Approximation polynomiale au sens des moindres carrés
(notes de cours)

Jean-Paul Chehab
Université de Picardie Jules Vernes
LAMFA CNRS 6140

version du 6 mai 2009

Contents

1 Formulation du problème

Soit V un espace vectoriel et $K \subset V$, un sous espace-vectoriel de K. On s'intéresse au problème suivant : soit $f \in V$

$$\text{Trouver } u \in K \text{ qui } \underline{\text{minimise}} \ \|u - f\|$$

c'est à dire tel que

$$\|u - f\| \leq \|v - f\| \ \forall v \in K.$$

Ici $\|.\|$ désigne une norme sur V.

La difficulté de la résolution de ce problème est étroitement liée au choix de la norme ; on peut en effet munir un même espace vectoriel de normes différentes, par exemple, pour $V = \mathbb{R}^N$, nous pouvons considérer les normes

$$\|u\|_2 = \sqrt{\sum_{i=1}^{N} u_i^2}, \ \|u\|_\infty = \max_{i=1,\cdots N} |u_i|, \ \|u\|_1 = \sum_{i=1}^{N} |u_i|.$$

Nous allons nous restreindre ici au cas des normes de type $\|.\|_2$ et nous parlerons donc d'approximation au sens des moindres carrés.

On suppose que V est muni du *produit scalaire* $< ., >$, c'est à dire que $< ., >$ vérifie le propriétés

Definition 1 *Soit V un espace vectoriel réel (pour simplifier), on rappelle que l'application*

$$< ., . >: V \times V \mapsto \mathbb{R}$$

est un produit scalaire si

i. $(u, v) \in V \times V \mapsto \phi(u, v) =< u, v >$ *est une forme bilinéaire sur* $V \times V$, *i.e.*

$$< \alpha u + \beta v, w >= \alpha < u, w > +\beta < v, w > \ \ et \ \ < u, \alpha y + \beta z >= \alpha < u, y > +\beta < u, z >$$

ii. ϕ *est* *<u>symétrique</u>, i.e.,* $\phi(u, v) = \phi(v, u) \ \forall (u, v) \in V \times V$

iii. ϕ *est* *<u>définie positive</u> :* $< u, u >\geq 0 \forall u \in V, \ < u, u >= 0 \iff u = 0$

Lorsque $V = \mathbb{R}^N$, le produit scalaire commun est le <u>produit scalaire euclidien</u> : $< u, v >= \displaystyle\sum_{i=1}^{N} u_i v_i$.

Nous rappelons enfin le résultat suivant :

Proposition 2 *Soit V un espace vectoriel muni du produit scalaire $< ., >$. alors l'application*

$$\begin{aligned} &\|.\| : V \to \mathbb{R}^+ \\ &v \mapsto \|v\| \end{aligned}$$

est une norme sur V.

Preuve. Il est facile de vérifier que

- $\|u\| \geq 0 \ \forall u \in V$ et $\|u\| = 0 \Rightarrow u = 0$.
- $\|\lambda u\| = \sqrt{< \lambda u, \lambda u >}\sqrt{\lambda^2 < u, u >} = |\lambda| \|u\|$.
- L'inégalité triangulaire est une conséquence de l'inégalité de Cauchy-Schwarz : $| < u, v > | \leq \|u\|.\|v\|$ que vérifie tout produit scalaire.

■

1.1 Le problème

Ici K est un espace vectoriel de dimension N. Soit $< .,>$ un produit saclaire sur $V \times V$, on note $\|.\| = \sqrt{< .,. >}$ la norme induite. On considère le problème

$$(\mathcal{M}) \quad \begin{array}{l} \text{Trouver } u \in K \text{ tel que} \\ \\ \|f - u\| \leq \|v - f\|, \ \forall v \in K. \end{array}$$

1.2 Le résultat principal

Commençons par caractériser la solution $(\mathcal{M})$. Nous avons la

Théorème 3 *Le problème $(\mathcal{M})$ admet une unique solution $u \in K$ qui est caractérisée par la relation*

$$< u - f, v >= 0, \forall v \in K.$$

Preuve. Démontrons tout d'abord que $u \in K$ est solution de $(\mathcal{M})$ si et seulement si u vérifie $< u - f, v >= 0, \forall v \in K$. Supposons tout d'abord que $(\mathcal{M})$ admette une solution u. Nous avons alors

$$\|f - u\|^2 \leq \|f - v\|^2, \forall (u, v) \in K.$$

Nous pouvons écrire $v = u + \lambda w$ avec $\lambda \in \mathbb{R}$ et $v \in K$ quelconques. Nous avons

$$\|f-v\|^2 = \|(f-u)-\lambda w\|^2 =< (f-u)-\lambda w, (f-u)-\lambda w >=< f-u, f-u > -2\lambda < f-u, w > +\lambda^2 < w, w > .$$

On déduit de l'inégalité précédente que

$$\|f - u\|^2 \leq \|f - u\|^2 - 2\lambda < f - u, w > +\lambda^2 \|w\|^2$$

donc que

$$-2\lambda < f - u, w > +\lambda^2 \|w\|^2 \geq 0, \ \forall w \in K, \forall \lambda \in \mathbb{R}.$$

Remplaçons λ par λ_k le terme général d'une suite de réels strictement positifs, tendant vers 0. Nous avons

$$-2 < f - u, w > +\lambda_k \|w\|^2$$

En passant à la limite, nous trouvons

$$< f - u, w >\leq 0 \ \forall w \in K.$$

On procède de même en remplaçant λ par $\lambda_k < 0$ avec $\lim_{k \to infty} \lambda_k = 0$ et on trouve

$$< f - u, w >\geq 0 \ \forall w \in K.$$

En conclusion, si u est solution de $(\mathcal{M})$, alors u vérifie

$$< f - u, w >= 0, \forall w \in K. \tag{1}$$

Il est facile de vérifier que si u satisfait cette dernière condition alors u est solution de $(\mathcal{M})$. Nous avons donc l'équivalence. Cherchons à présent à montrer que cette solution u est unique. Pour cela rappelons-nous que l'espace vectoriel K est de dimension N : on peut donc considérer e_i, $i = 1, \cdots N$ une base de K. Nous remarquons que

$$< f - u, w >= 0 \iff < f - u, e_i >= 0, i = 1, \cdots, N \tag{2}$$

En effet, si u satisfait (2) alors u satisfait (2) en prenant $w = e_i$ sucessivement pour $i = 1, \cdots, N$. Réciproquement si u est solution de (2), alors on a

$$\sum_{i=1}^{N} \alpha_i < f - u, e_i >= 0 =< f - u, \sum_{i=1}^{N} \alpha_i e_i >, \forall \alpha_i \in \mathbb{R}^N$$

C'est à dire que la relation a lieu pour toute combinaison linéaire des e_i. Comme les e_i forment une base de K, nous obtenons (1).

Maintenant, rechercher $u \in K$ revient à rechercher ses coefficients dans la base e_i. On cherche donc u sous la forme

$$u = \sum_{j=1}^{N} a_j e_j.$$

En remplaçant u par cette expression dans (2), nous trouvons

$$< f - \sum_{j=1}^{N} a_j e_j, e_i >= 0, \ i = 1, \cdots, N$$

soit

$$\sum_{j=1}^{N} < e_j, e_i > a_j =< f, e_i >, \ i = 1, \cdots, N,$$

ou bien, comme $< ., >$ est symétrique

$$\sum_{j=1}^{N} < e_i, e_j > a_j =< f, e_i >, \ i = 1, \cdots, N,$$

on reconnaît à gauche la i-ème composante du produit matrice-vecteur AU avec $A_{i,j} =< e_i, e_j >$ et $U = (a_1, \cdots, a_N)$. Ainsi le problème de départ se reformule de manère équivalente en résoudre le système linéaire

$$AU = F$$

avec $F = (< f, e_1 >, \cdots, < f, e_N >)$. Pour conclure la démonstration, il faut et il suffit de montrer que la matrice A est inversible. Soit donc $Z = \sum_{i=1}^{N} Z_i e_i$ un vecteur de K représenté par $(Z_1, \cdots, Z_N)$ tel que $AZ = 0$. Nous allons montrer que nécessairement $Z = 0$. Nous avons

$$AZ = 0 \Rightarrow (AZ, Z) = 0 \iff \sum_{i=1}^{N} (\sum_{j=1}^{N} < e_i, e_j > Z_j) Z_i) = 0$$

où $(., .)$ est le produit scalaire euclidien. Ainsi

$$< \sum_{i=1}^{N} Z_i e_i, \sum_{j=1}^{N} z_j e_j >= \|Z\|^2 = 0$$

et donc $Z = 0$. ∎

2 Approximation continue

Soit f une fonction continue sur l'intervalle $[0,1]$. On considère $V = \{u, \int_0^1 u^2 dx < +\infty\}$. On montre que V est un espace vectoriel, que $f \in V$ et que l'on peut munir V du produit scalaire

$$< f, g >= \int_0^1 fg dx$$

Soit $\subset V$, on considère le problème

$$\begin{array}{l} \text{Trouver } u \in K \text{ tel que} \\ \int_0^1 (u-f)^2 dx \leq \int_0^1 (v-f)^2 dx, \ \forall v \in K \end{array}$$

Soit $K = ¶_N$, l'espace vectoriel des polynômes de degré inférieur ou égal à N et soit $e_i = x^{i-1}, i = 1, \cdots N+1$ la base canonique de P_N. Le théorème précédent nous assure qu'il existe une unique solution $u = \sum_{i=0}^{N} a_i x^i$ au problème de minimisation et que les a_i sont solution de $AU = F$ avec

- $F_i = \int_0^1 f x^{i-1} dx, \ i = 1, \cdots N+1$
- $A_{i,j} = \int_0^1 x^{i-1} x^{j-1} dx = \dfrac{1}{i+j+1}, \ \cdots i, j = 1, \cdots N.$

La matrice du problème est la matrice de Hilbert. C'est une matrice trs mal conditionnée, c'est à dire que le rapport entre la plus grande et la plus petite valeur propre est trs grand, il est de l'ordre de $e^{\frac{7n}{2}}$. La conséquence est qu'une petite variation des données, par exemple le second membre, peut induire uen variation importante du résultat. Nous comparons ci-après le calcul du polynôme approchant e^x sur $[0,1]$ au sens des moindres carrés lorsque les termes F_i sont calculés exactement et lorsque leur valeur numérique est obtenu en ne prenant que les 3 premiers chiffres significatifs. Nous remarquons qu ele polynôme obtenu n'a plus rien à voir avec celui sans perturbation.

On peut remplacer le produit scalaire par

$$< f, g >= \int_a^b fgw dx$$

où w est une fonction intégrable, à valeurs positives, appelée fonction poids. Par exemple, si $[a,b] = [-1,1]$ et si $w(x) = \dfrac{1}{\sqrt{1-x^2}}$, nous pouvons chercher u non plus dans la base canonique mais dans la base des polynômes de Tchebytecheff. Grâce à leur orthogonalité, dans cette base, la matrice devient diagonale, le système est donc très facile à résoudre.

3 Approximation discrète

3.1 Cadre général

On se donne $(x_j)_{j=1}^{N+1}$, $N+1$ réels deux à deux distincts et f une application continue sur un intervale I contenant tous les x_j. On considère le problème

$$\text{Minimiser } E(f) = \sum_{j=1}^{N+1} (f(x_j) - p(x_j))^2 \text{ pour p, polynôme de } P_M \tag{3}$$

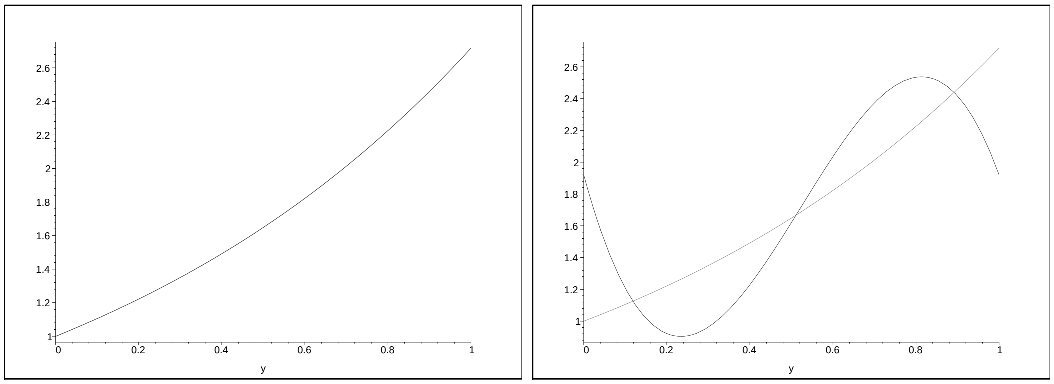

Figure 1: Approximation polynômiale au sens des moindres carrés de $f(x) = e^x$ dans $[0, 1]$, la matrice de Gram est la matrice de Hilbert. Comparaison entre solution exacte et approchée. A gauche résolution exacte, à droite résolution avec second membre tronqué

Cette expression est le carré d'un produit scalaire dans $\mathbb{R}^{N+1}$, ce problème rentre donc dans le cadre décrit plus haut.

En cherchant p dans la base canonique,

$$p = \sum_{k=0}^{M} a_k x^k$$

Théorème 4 *Le problème de minimisation (3) admet une unique solution* $p = \sum_{k=0}^{M} a_k x^k$ *caractérisée par les relations*

$$\sum_{k=0}^{M} \left(\sum_{j=1}^{N+1} x_j^k x_j^\ell \right) a_k = \sum_{j=1}^{N+1} f(x_j) x_j^\ell, \ \ \ell = 0, \cdots, M.$$

Preuve. L'existence et l'unicité découlent de la présence d'un produit scalaire. La caractérisation de la solution dans la base canonique est aussi une conséquence du théorème. Nous avons

$$\sum_{j=1}^{N+1} (f(x_j) - \sum_{k=0}^{M} a_k x_j^k) x_j^\ell = 0, \ell = 0, \cdots M.$$

■

Cette dernière formule peut être interprétée comme une approximation de celle obtenue dans le cas continue. Supposons en effet que les x_i soient régulièrement espacés par h. Nous avons en effet, d'une part,

$$\sum_{j=1}^{N+1} (f(x_j) - \sum_{k=0}^{M} a_k x_j^k) x_j^\ell h = 0, \ell = 0, \cdots M.$$

ou encore

$$\sum_{k=0}^{M}\left(\sum_{j=1}^{N+1} x_j^k x_j^\ell h\right) a_k = \sum_{j=1}^{N+1} f(x_j) x_j^\ell h.$$

D'autre part,

$$\int_{x_1}^{x_{N+1}} f(x)dx \simeq \sum_{j=1}^{N+1} f(x_j)x_j^\ell h \text{ et } \int_{x_1}^{x_{N+1}} x^k x dx \simeq \sum_{j=1}^{N+1} x_j^k x_j^\ell h$$

Enfin, il est important de remarquer qu'en général $M \neq N$. Lorsque $M = N$, $E(f) = 0$ quand p est le polynôme d'interpolation de f aux points x_j et c'est donc la solution du problème de minimisation. Montrons que le système obtenu est

$$V^T V U = V^T F$$

où V est la matrice de Vandermonde. Les coefficients de V sont

$$V_{i,j} = x_i^{j-1}, i, j = 1, \cdots, N+1$$

Les coefficients de $V^T V$ sont

$$(V^T V)_{i,j} = \sum_{k=1}^{N+1} x_k^{i-1} x_k^{j-1}.$$

Le système obtenu est donc bien $V^T V U = V^T F$ et V étant inversible, il est équivalent à $VU = F$ dont la solution est le vecteur de composantes les coefficients du polynôme d'interpolation dans la base canonique.
Exemple $f(x) = e^x$, $I = [0, 1]$, $x_1 = 0, x_2 = \frac{1}{2}, x_3 = 1$, $M = 1$.

References

[1] A. Quarteroni, R. Sacco, F. Saleri, *Méthodes numériques pour le calcul scientifique*, Springer Paris, 2000.

[2] M. Schatzman, *Analyse numérique pour la licence*, InterEdition, 1993.