Agrégation marocaine de Mathématiques, session 2025

Épreuve de modélisation et Calcul Scientifique

Méthodes d'approximation numérique des solutions d'un problème de minimisation

Résumé : On s'intéresse dans ce texte à différentes méthodes d'approximation numérique des solutions d'un problème de minimisation sous contraintes modélisant un phénomène de conduction thermique dans une barre métallique .

Mots clefs : Optimisation. Algèbre linéaire. Méthodes itératives.

► *Il est rappelé que le jury n'exige pas une compréhension exhaustive du texte. La présentation, bien que totalement libre, doit être organisée et le jury apprécie qu'un plan soit annoncé en préliminaire. L'exposé doit être construit en évitant la paraphrase et mettant en lumière les connaissances, à partir des éléments du texte. Il doit contenir des illustrations informatiques réalisées sur ordinateur, ou, à défaut, des propositions de telles illustrations. Des pistes de réflexion, indicatives et largement indépendantes les unes des autres, vous sont proposées en fin de texte.*

1. Système linéaire et problème de minimisation

De très nombreux problèmes issus des sciences et techniques conduisent en définitive à résoudre un système linéaire de la forme

$$Ax = b, \tag{1}$$

où b est un vecteur donné de $\mathbb{R}^N$ et A une matrice symétrique définie positive (SDP) de taille $N \times N$. La question qui va nous intéresser peut se résumer ainsi : les coefficients de la matrice A ou le second membre b proviennent de mesures et sont entachés d'erreurs, mais on peut effectuer des observations de sorte qu'on va en fait chercher une solution x qui réalise un compromis entre la réalisation de (1) et l'accord avec les observations.

Le point de vue adopté repose sur une interprétation de (1) comme un problème de minimisation.

Théorème 1. *Le vecteur $x \in \mathbb{R}^N$ est solution de* (1) *si et seulement si il réalise le minimum sur $\mathbb{R}^N$ de*

$$\mathscr{J}(y) = \frac{1}{2}\langle Ay, y\rangle - \langle b, y\rangle, \tag{2}$$

où $\langle \cdot, \cdot \rangle$ désigne le produit scalaire usuel sur $\mathbb{R}^N$.

Le problème (1) peut être résolu numériquement par de nombreuses méthodes, directes ou itératives. Le théorème 1 suggère qu'on peut obtenir la solution x comme limite d'un algorithme de gradient visant à minimiser la fonctionnelle $\mathscr{J}$. C'est cette démarche que nous allons suivre et adapter.

2. Un exemple d'application

On s'intéresse à la répartition de température dans une barre métallique de longueur $L>0$ et de section constante Σ, dont on désigne l'aire par $\sigma>0$, soumise à une source de chaleur S. On suppose que toutes les grandeurs physiques sont homogènes dans les sections de sorte que le problème se réduit à une description mono-dimensionnelle, où toutes ces grandeurs sont des fonctions de la seule position $0\le x\le L$. Les propriétés thermiques de la barre sont décrites par la conductivité, qu'on note $a(x)>0$. On suppose qu'il existe deux constantes $a^\star, a_\star>0$ telles que $0<a_\star\le a(x)\le a^\star$ pour tout $x\in[0,L]$. La loi de Fourier définit alors le flux de chaleur [1] à travers la surface d'abscisse x comme $J(x)=-a(x)\frac{\mathrm{d}}{\mathrm{d}x}u(x)$, $x\mapsto u(x)$ désignant la température. On s'intéresse uniquement au régime permanent, c'est-à-dire quand toutes les grandeurs peuvent être considérées indépendantes du temps. Le bilan d'énergie dans le volume $\Sigma\times[x,x+\delta x]$ s'énonce

$$\sigma\big(J(x+\delta x)-J(x)\big)=\int_x^{x+\delta x} S(y)\,\sigma\,\mathrm{d}y. \tag{3}$$

On en déduit que la température $x\mapsto u(x)$ obéit à l'équation différentielle

$$-\frac{\mathrm{d}}{\mathrm{d}x}\Big(a(x)\frac{\mathrm{d}}{\mathrm{d}x}u(x)\Big)=S(x),\qquad \text{pour } 0<x<L, \tag{4}$$

alors qu'on suppose les bords maintenus à une température fixe

$$u(0)=0=u(L) \tag{5}$$

(pour un choix d'unités convenables).

On peut chercher la solution de ce problème sous la forme d'une série trigonométrique $u(x)=\sum_{j=1}^{\infty}u_j\sin(j\pi x/L)$. Une méthode de résolution *approchée* de (4)–(5) consiste, pour un entier $N\in\mathbb{N}$ fixé, à définir $u^{(N)}(x)=\sum_{j=1}^{N}u_j^{(N)}\sin(j\pi x/L)$ où $\mathbb{U}^{(N)}=(u_1^{(N)},\dots,u_N^{(N)})\in\mathbb{R}^N$ est obtenu comme unique solution du système linéaire

$$\begin{cases} \mathbb{A}^{(N)}\mathbb{U}^{(N)}=\mathbb{S}^{(N)},\\ \mathbb{A}^{(N)}_{jk}=\dfrac{2}{L}\dfrac{\pi^2}{L^2}\,jk\displaystyle\int_0^L a(x)\cos\Big(\frac{\pi jx}{L}\Big)\cos\Big(\frac{\pi kx}{L}\Big)\,\mathrm{d}x, & j,k\in\{1,\dots,N\},\\ \mathbb{S}^{(N)}_{j}=\dfrac{2}{L}\displaystyle\int_0^L \sin\Big(\frac{\pi jx}{L}\Big)\,S(x)\,\mathrm{d}x, & j\in\{1,\dots,N\}. \end{cases} \tag{6}$$

En fait, $u^{(N)}$ apparaît ainsi comme la projection de u sur $\mathrm{Vect}\big\{\sin(\pi nx/L),\ n\in\{1,\dots,N\}\big\}$ pour le produit scalaire défini par

$$p(u,v)=\int_0^L a(x)\frac{\mathrm{d}}{\mathrm{d}x}u(x)\frac{\mathrm{d}}{\mathrm{d}x}v(x)\,\mathrm{d}x. \tag{7}$$

On note que $(u,v)\mapsto p(u,v)$ est bien défini sur l'ensemble $\mathscr{C}=\{u\in C^1([0,L]),\ u(0)=0=u(L)\}$ et que la forme quadratique $u\mapsto p(u,u)$ est définie positive sur $\mathscr{C}$: on dispose ainsi

1. c'est-à-dire la quantité d'énergie traversant la surface Σ par unité de temps.

d'un espace préhilbertien et on cherche à projeter sur un ensemble de dimension finie. Ces remarques sont à la base des arguments qui permettent de justifier le

Théorème 2. *Pour tout $S \in L^2([0,L])$, quand $N \to \infty$, $u^{(N)}$ converge vers u dans $L^2([0,L])$, où u est solution de* (4)–(5) .

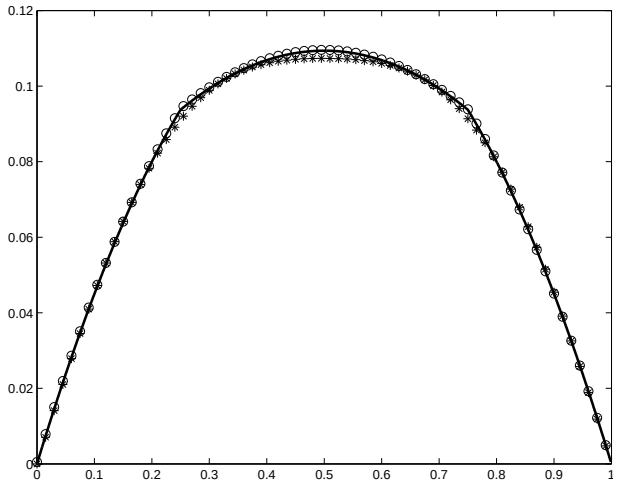

FIGURE 1. Données : $a(x) = 1 + \mathbf{1}_{1/4<x<3/4}$, $S(x) = 1$, $L = 1$. Solution exacte $u(x) = (x/2 - x^2/2)(\mathbf{1}_{0\le x\le 1/4} + \mathbf{1}_{3/4\le x\le 1}) + (3/64 + x/4 - x^2/4)\mathbf{1}_{1/4<x<3/4}$ (trait plein). Solution approchée $u^{(4)}$ (symboles $*$) et solution par différences finies (symboles $\circ$) avec $\Delta x = 0.01$.

En pratique, la connaissance de la conductivité $x \mapsto a(x)$ est affectée d'erreurs. En conséquence, la solution u obtenue ne reproduit pas fidèlement l'échauffement de la barre. L'idée consiste à améliorer la prédiction de la température en tenant compte de mesures. On désigne par $m \ll N$ le nombre de mesures disponibles. On note par $\xi_1, \ldots, \xi_m$ les points de mesure, éléments de $]0,L[$ et $\mathbb{V}^{\rm obs} = (v_1^{\rm obs}, \ldots, v_m^{\rm obs}) \in \mathbb{R}^m$ désigne les températures observées à ces points de mesure. On introduit l'application Ω qui, à un vecteur $\mathbb{U} = (u_1, \ldots, u_N) \in \mathbb{R}^N$, associe le vecteur de $\mathbb{R}^m$ dont les coordonnées sont $\sum_{j=1}^N u_j \sin(\pi j \xi_\mu / L)$, pour tout $\mu \in \{1, \ldots, m\}$. On notera encore Ω la matrice canoniquement associée à cette application. Alors, au lieu de résoudre le système linéaire (6), on minimise la fonctionnelle

$$\mathbb{U} \in \mathbb{R}^N \longmapsto \mathscr{J}^{(N)}(\mathbb{U}) = \frac{1}{2}\langle \mathbb{A}^{(N)}\mathbb{U}, \mathbb{U}\rangle - \langle \mathbb{S}^{(N)}, \mathbb{U}\rangle, \tag{8}$$

sous la contrainte

$$\Omega\mathbb{U} = \mathbb{V}^{\rm obs}. \tag{9}$$

Ce problème amène à introduire une inconnue supplémentaire $\lambda \in \mathbb{R}^m$ et à résoudre le système

$$\begin{pmatrix} \mathbb{A}^{(N)} & \Omega^T \\ \Omega & 0 \end{pmatrix} \begin{pmatrix} \mathbb{U} \\ \lambda \end{pmatrix} = \begin{pmatrix} \mathbb{S}^{(N)} \\ \mathbb{V}^{\rm obs} \end{pmatrix}. \tag{10}$$

Proposition 1. *On suppose que* Ω *est de rang maximal. Alors, le problème* (10) *admet une unique solution* $(\mathbb{U},\lambda)$ *où* $\mathbb{U}$ *est solution du problème de minimisation de* $\mathscr{J}^{(N)}$ *sous la contrainte* $\Omega\mathbb{U} = \mathbb{V}^{\rm obs}$.

Démonstration. Le point clef consiste à écrire le problème sous la forme : $\mathbb{U} = \big[\mathbb{A}^{(N)}\big]^{-1}\big(\mathbb{S}^{(N)} - \Omega^T\lambda\big)$ puis, avec la contrainte, $\Omega\big[\mathbb{A}^{(N)}\big]^{-1}\Omega^T\lambda = \Omega\big[\mathbb{A}^{(N)}\big]^{-1}\mathbb{S}^{(N)} - \mathbb{V}^{\rm obs}$. □

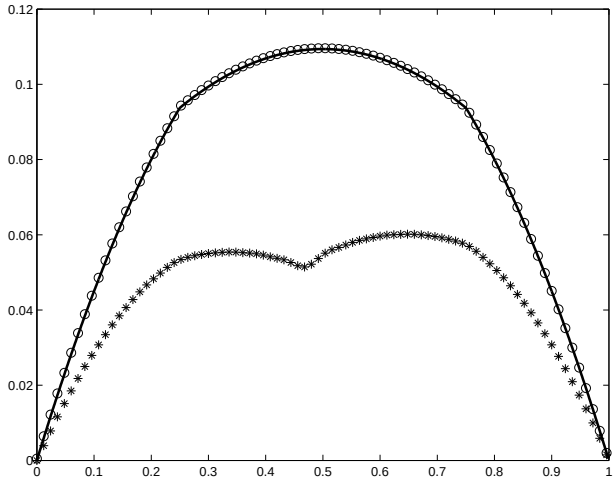

FIGURE 2. Même problème que pour la figure 1. Points de mesure : $(\xi_1, v_1^{\rm obs}) = (0.4711, 0.0515)$ et $(\xi_2, v_2^{\rm obs}) = (0.5005, 0.0547)$, calculs avec $N = 40$. Solution exacte (trait plein), solution approchée $u^{(40)}$ (symboles $*$) et solution par différences finies (symboles $\circ$).

3. Diverses méthodes numériques

On va maintenant exploiter des formulations différentes du problème qui vont guider la mise en œuvre de méthodes numériques dont on comparera les performances.

3.1. *Un schéma itératif de minimisation*

On pose $\mathscr{A} = \Omega\big[\mathbb{A}^{(N)}\big]^{-1}\Omega^T$. En combinant la preuve de la proposition 1 et le théorème 1, λ apparaît comme solution du problème consistant à minimiser

$$\zeta \mapsto \frac{1}{2}\langle \mathscr{A}\zeta,\zeta\rangle - \langle \Omega\big[\mathbb{A}^{(N)}\big]^{-1}\mathbb{S}^{(N)} - \mathbb{V}^{\rm obs},\zeta\rangle, \tag{11}$$

où on conserve la notation $\langle\cdot,\cdot\rangle$ pour désigner le produit scalaire usuel sur $\mathbb{R}^m$. Cette remarque motive la mise en œuvre d'algorithmes de type « méthode de gradient » où λ est obtenu comme limite d'une suite vérifiant

$$\lambda_{n+1} = \lambda_n - \varrho\big(\mathscr{A}\lambda_n + \mathbb{V}^{\rm obs} - \Omega\big[\mathbb{A}^{(N)}\big]^{-1}\mathbb{S}^{(N)}\big), \tag{12}$$

avec $\varrho > 0$ un paramètre à choisir « assez petit » pour assurer la convergence de la méthode. Ayant déterminé λ, on obtient simplement $\mathbb{U}$ en résolvant un système linéaire. La difficulté technique pour exploiter cette méthode est qu'elle nécessite d'avoir accès à la matrice inverse $\left[\mathbb{A}^{(N)}\right]^{-1}$.

3.2. *Une méthode de recherche de point–selle*

En pratique, on utilise plutôt l'algorithme itératif suivant qui définit à la fois des valeurs de $\mathbb{U}$ et λ approchées. À cette fin, on interprète (10) comme un problème de point–selle.

Proposition 2. *Le couple* $(\mathbb{U},\lambda)$ *est solution de* (10) *si et seulement si* $(\mathbb{U},\lambda)$ *est point–selle de*

$$\mathscr{L}(\mathbb{W},\zeta) = \mathscr{J}(\mathbb{W}) + \langle \zeta, \Omega\mathbb{W} - \mathbb{V}^{\text{obs}} \rangle, \tag{13}$$

c'est-à-dire que pour tout $(\mathbb{W},\zeta)$*, on a* $\mathscr{L}(\mathbb{U},\zeta) \le \mathscr{L}(\mathbb{U},\lambda) \le \mathscr{L}(\mathbb{W},\lambda)$.

En d'autres termes, on a $\mathscr{L}(\mathbb{U},\lambda) = \max_\lambda \min_{\mathbb{W}} \mathscr{L}(\mathbb{W},\lambda) = \min_{\mathbb{W}} \max_\lambda \mathscr{L}(\mathbb{W},\lambda)$. Étant donnés $\mathbb{U}_0$ et λ_0, on construit les suites définies par les relations

$$\mathbb{U}_{n+1} = \mathbb{U}_n - \varrho\left[\mathbb{A}^{(N)}\mathbb{U}_n - \mathbb{S}^{(N)} + \Omega^T\lambda_n\right], \qquad \lambda_{n+1} = \lambda_n + \varrho(\Omega\mathbb{U}_{n+1} - \mathbb{V}^{\text{obs}}). \tag{14}$$

Théorème 3. *On suppose que* Ω *est de rang maximal. Il existe* $\varrho_\star > 0$ *tel que si* $0 < \varrho < \varrho_\star$*, la suite* $\left(\mathbb{U}_n,\lambda_n\right)_{n\in\mathbb{N}}$ *converge vers* $(\mathbb{U},\lambda)$ *solution de* (10).

Preuve. On note $e_n = \mathbb{U}_n - \mathbb{U}$ et $r_n = \lambda_n - \lambda$ qui vérifient donc $e_{n+1} = (I - \varrho\mathbb{A}^{(N)})e_n - \varrho\Omega^T r_n$ et $r_{n+1} = r_n + \varrho\Omega e_{n+1}$. On remarque alors que $\|e_{n+1}\|^2 = \langle (I-\varrho\mathbb{A}^{(N)})e_n, e_{n+1}\rangle - \varrho\langle r_n, \Omega e_{n+1}\rangle$. On utilise cette relation pour calculer

$$\|r_{n+1}\|^2 = \|r_n\|^2 + \varrho^2\|\Omega e_{n+1}\|^2 - 2\|e_{n+1}\|^2 + 2\langle e_n, (I-\varrho\mathbb{A}^{(N)})e_{n+1}\rangle. \tag{15}$$

On en déduit que

$$\|r_{n+1}\|^2 + \|e_{n+1}\|^2 \le \|r_n\|^2 + \|e_n\|^2 - \left(1 - \varrho^2\|\Omega\|^2 - \|I-\varrho\mathbb{A}^{(N)}\|^2\right)\|e_{n+1}\|^2. \tag{16}$$

Pour ϱ assez petit, le dernier terme contribue négativement de sorte que la suite $\left(\|r_n\|^2 + \|e_n\|^2\right)_{n\in\mathbb{N}}$ est décroissante. Comme elle est minorée par 0, elle converge et en revenant à (16) on en déduit que $\|e_n\|$ tend vers 0 quand $n \to \infty$. Alors, la relation $e_{n+1} = (I-\varrho\mathbb{A}^{(N)})e_n - \varrho\Omega^T r_n$ entraîne que $\Omega^T r_n$ converge aussi vers 0, et, finalement, puisque Ω est surjective, que r_n converge vers 0. □

3.3. *Pénalisation*

Une autre approche, qui permet de s'affranchir de l'introduction du vecteur λ, traite les observations par pénalisation : pour $0 < \varepsilon \ll 1$ donné, on cherche $\mathbb{U}^\varepsilon$ qui minimise sur $\mathbb{R}^N$

$$\mathscr{J}^{(N)}(\mathbb{U}) + \frac{1}{2\varepsilon}\|\Omega\mathbb{U} - \mathbb{V}^{\text{obs}}\|^2. \tag{17}$$

Autrement dit, $\mathbb{U}^\varepsilon$ est solution du système linéaire

$$\mathbb{A}^{(N)}\mathbb{U}^\varepsilon + \frac{1}{\varepsilon}\Omega^T\Omega\mathbb{U}^\varepsilon = \mathbb{S}^{(N)} + \frac{1}{\varepsilon}\Omega^T\mathbb{V}^{\text{obs}}. \tag{18}$$

La pertinence de cette approche est validée par l'énoncé suivant.

Théorème 4. *Quand* $\varepsilon \to 0$, $\mathbb{U}^\varepsilon$ *converge vers* $\mathbb{U}$ *dans* $\mathbb{R}^N$ *et il existe* $\lambda \in \mathbb{R}^m$ *tel que* $(\mathbb{U}, \lambda)$ *est solution de* (10).

L'avantage de cette formulation est qu'elle permet de travailler avec un système dont la taille est réduite par rapport à (10) ; toutefois la matrice $\mathbb{A}^{(N)} + \frac{1}{\varepsilon}\Omega^T\Omega$ est mal conditionnée quand $0 < \varepsilon \ll 1$.

Suggestions et pistes de réflexion

► *Les pistes de réflexion suivantes ne sont qu'indicatives et il n'est pas obligatoire de les suivre. Vous pouvez choisir d'étudier, ou non, certains des points proposés, de façon plus ou moins approfondie, mais aussi toute autre question à votre initiative. Vos investigations comporteront une partie traitée sur ordinateur et, si possible, des représentations graphiques de vos résultats. À défaut, si vos illustrations informatiques n'ont pas abouti, il est conseillé d'expliquer ce que vous auriez souhaité mettre en œuvre.*

— Commenter la méthode d'approximation de l'équation (4) proposée dans le texte. Démontrer le théorème 2.
— Le problème (6) implique le calcul de diverses intégrales. Expliquer comment évaluer numériquement ces quantités et comparer différentes méthodes.
— Proposer d'autres schémas numériques pour résoudre (4) et comparer les résultats. On pourra notamment considérer des données S continues ou au contraire présentant des discontinuités. On pourra mener la même discussion avec le coefficient a et étudier le comportement de la méthode quand N varie.
— Justifier le théorème 1 et la proposition 1.
— Démontrer le théorème 3 et mettre en œuvre la méthode numérique correspondante.
— Démontrer le théorème 4 et mettre en œuvre la méthode numérique correspondante. Étudier numériquement le comportement de la méthode quand $\varepsilon \to 0$.
— Étudier numériquement l'influence du nombre m de mesures disponibles, de la répartition $\xi_1, \ldots, \xi_m$ et de l'amplitude relative des mesures (quantités que l'on peut fixer de manière aléatoire). À défaut de réaliser ces expérimentations, décrire précisément un ensemble de tests permettant d'évaluer les limites éventuelles du modèle.