

## Décomposition de domaine et méthodes itératives pour les problèmes aux limites

---

**Résumé :** On analyse une méthode itérative pour la résolution numérique de problèmes aux limites linéaires, qui consiste à décomposer le domaine d'étude en deux sous-domaines de résolution.

**Mots clefs :** Algèbre linéaire. Méthodes itératives. Différences finies.

---

- *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. Le problème de Dirichlet

On s'intéresse à la résolution numérique du problème aux limites

$$(1) \quad -u''(x) = f(x) \quad \text{pour } x \in ]0, 1[, \quad u(0) = \alpha, \quad u(1) = \beta$$

où  $f : ]0, 1[ \rightarrow \mathbb{R}$  et  $\alpha, \beta \in \mathbb{R}$  sont donnés. Les méthodes qu'on va étudier reposent sur une décomposition du domaine  $[0, 1] = [0, x_*] \cup ]x_*, 1]$  avec la motivation de confier la résolution des sous-systèmes à des ordinateurs différents afin d'accélérer le temps de calcul. En fait, on va voir qu'il peut être pertinent de travailler avec une décomposition qui, éventuellement, se recouvre partiellement  $[0, 1] = [0, x_*^r] \cup ]x_*, 1]$ , avec  $x_*^r > x_*$ . Le problème est alors résolu de manière itérative, construite en faisant communiquer les domaines  $[0, x_*^r]$  et  $]x_*, 1]$ .

**Algorithme 1.** Connaissant  $u^{(n)}$  :

- Résoudre  $-(u^{(n+1/2)})''(x) = f(x)$  sur  $]0, x_*^r[$  avec  $u^{(n+1/2)}(0) = \alpha$ ,  $u^{(n+1/2)}(x_*^r) = u^{(n)}(x_*^r)$ ,
- Puis résoudre  $-(u^{(n+1)})''(x) = f(x)$  sur  $]x_*, 1[$  avec  $u^{(n+1)}(x_*) = u^{(n+1/2)}(x_*)$ ,  $u^{(n+1)}(1) = \beta$  et poser  $u^{(n+1)}(x) = u^{(n+1/2)}(x)$  pour  $x \in ]0, x_*^r[$ .

Cette approche n'est pas très satisfaisante du point de vue de l'objectif annoncé. On lui préfère donc la méthode suivante qui va être étudiée dans le reste du texte.

**Algorithme 2.** Connaissant  $u^{(n)}$  :

- Résoudre  $-(u_g)''(x) = f(x)$  sur  $]0, x_*^r[$  avec  $u_g(0) = \alpha$ ,  $u_g(x_*^r) = u^{(n)}(x_*^r)$ ,
- Résoudre  $-(u_d)''(x) = f(x)$  sur  $]x_*, 1[$  avec  $u_d(x_*) = u^{(n)}(x_*)$ ,  $u_d(1) = \beta$ ,
- Poser  $u^{(n+1)}(x) = u_g(x)\mathbf{1}_{[0, x_*^r]}(x) + u_d(x)\mathbf{1}_{]x_*, 1]}(x)$ . (On n'utilise donc pas l'information contenue dans  $u_g(x)$  pour  $x_* \leq x \leq x_*^r$ .)

Bien entendu ces questions sont surtout motivées par des problèmes multi-dimensionnels pour lesquels on ne connaît pas la solution. Par exemple ces approches ont permis de démontrer l'existence de solutions, et de les calculer numériquement, pour le problème  $-\Delta u = f$  avec conditions de Dirichlet, posé sur la réunion d'un disque et d'un carré.

## 2. Exemples numériques

Pour motiver l'analyse, on réalise des simulations sur le problème (1) avec  $\alpha = 1$ ,  $\beta = 2$  et  $f(x) = 10 \sin(3x)$ . On prend  $J = 500$  points de discrétisation. Le point  $x_*$  est fixé au milieu du domaine. Le premier itéré est pris nul. L'algorithme s'arrête quand l'erreur relative  $\frac{\|u^{(n+1)} - u^{(n)}\|}{\|u^{(n)}\|}$  est inférieure à  $10^{-4}$ . La figure 1 compare une solution de référence, obtenue en résolvant directement le système (2) pour un maillage fin, et quelques itérations de l'algorithme 2. La figure 2 représente l'évolution de l'erreur relative entre deux itérés consécutifs pour 10 ou 17 points de recouvrements. Cette expérience semble indiquer que la convergence est plus rapide quand le recouvrement est plus large. Dans la suite on va discuter des éléments d'analyse qui étayent ces observations et prouvent la convergence de l'algorithme 2.

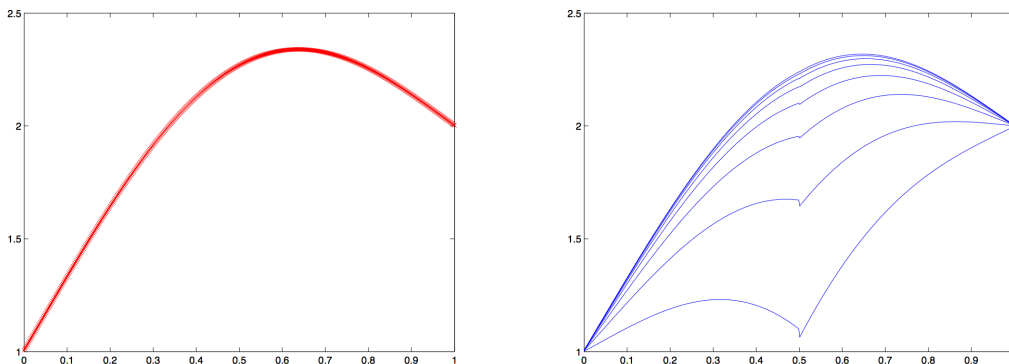


FIGURE 1. Solution de référence de (1) (à gauche) et itérations obtenues par l'algorithme 2 (à droite).

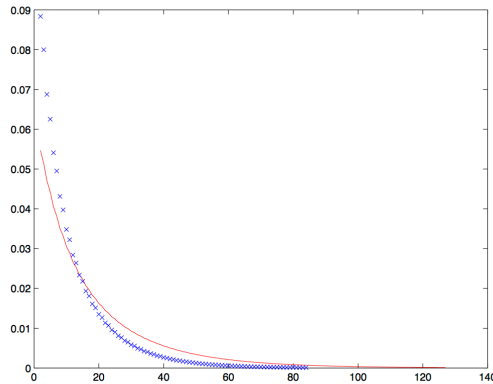


FIGURE 2. Erreur relative en fonction du nombre d'itérations pour 10 (trait plein) et 17 (croix) points de recouvrement.

### 3. Le point de vue discret

La résolution numérique du problème (1) — tout comme celle des problèmes impliqués dans les algorithmes 1 et 2 — conduit à considérer le système linéaire

$$(2) \quad AU = b, \quad A = \frac{1}{h^2} \begin{pmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & & & \\ 0 & & \ddots & & \\ \vdots & & & 2 & -1 \\ 0 & \dots & 0 & -1 & 2 \end{pmatrix}, \quad b = \begin{pmatrix} f(h) \\ f(2h) \\ \vdots \\ f((J-1)h) \\ f(Jh) \end{pmatrix} + \begin{pmatrix} \alpha/h^2 \\ 0 \\ \vdots \\ 0 \\ \beta/h^2 \end{pmatrix},$$

pour certains  $J \in \mathbb{N}$  et  $0 < h \ll 1$ . L'énoncé suivant met en évidence des propriétés remarquables de la matrice  $A$ .

**Lemme 1.** Soit  $T \in \mathcal{M}_J(\mathbb{R})$  une matrice irréductible<sup>1</sup> telle que pour tout  $i \in \{1, \dots, J\}$ , on a  $t_{ii} > 0$ ,  $t_{ij} \leq 0$  si  $i \neq j$  et  $\sum_{j=1}^J t_{ij} \geq 0$ . On suppose de plus qu'il existe  $k \in \{1, \dots, J\}$  tel que  $\sum_{j=1}^J t_{kj} > 0$ . Alors  $T$  est inversible. De plus  $T$  est une  $\mathcal{M}$ -matrice : si les coordonnées de  $b \in \mathbb{R}^J$  sont positives, alors les coordonnées de  $x$ , solution de  $Tx = b$ , le sont aussi (et les coefficients de  $T^{-1}$  sont positifs).

**Preuve.** On raisonne par l'absurde en supposant que 0 est valeur propre de  $T$ . Si  $Tx = 0$ , on pose  $I = \{i \in \{1, \dots, J\}, |x_i| = \|x\|_\infty\}$ . On montre alors que pour tout  $i \in I$ ,  $\sum_{j=1}^J t_{ij} = 0$ , puis que ou bien  $t_{ij} = 0$ , ou bien  $j \in I$ . Les hypothèses sur  $T$  permettent de conclure à une contradiction. Ces arguments s'adaptent pour justifier que  $T$  est une  $\mathcal{M}$ -matrice.  $\square$

1. c'est-à-dire que, en disant que  $(k, l)$  est un arc de  $T$  si  $t_{kl} \neq 0$ , on peut trouver pour tous  $i, j \in \{1, \dots, J\}$  une suite d'arcs reliant  $i$  à  $j$ .

On va analyser l'algorithme 2 du point de vue discret. On dispose donc d'une partition  $x_0 = 0 < x_1 = h < x_2 = 2h < \dots < x_{J+1} = (J+1)h = 1$  où on identifie  $x_* = J_* h \leq x_*^r = J_*^r h$ . L'algorithme 2 peut s'écrire en termes de systèmes linéaires  $A_g^r U_g = f_g^r - B_g^r U^{(n)}$ ,  $A_d U_d = f_d - B_d U^{(n)}$  qui ont la même forme que (2), où les matrices  $A_g^r$ ,  $A_d$  ont la même structure que  $A$  mais sont de tailles  $J_*^r \times J_*^r$  et  $(J - J_*) \times (J - J_*)$  respectivement,  $f_g^r$  contient les  $J_*^r$  premières coordonnées du vecteur  $f$  et  $f_d$  ses  $J - J_*$  dernières coordonnées, etc. Puis on pose  $U^{(n+1)} = (U_{g,1}, \dots, U_{g,J_*}, U_{d,1}, \dots, U_{d,J-J_*})$ . En fait, ces systèmes correspondent à des décompositions par blocs de la matrice  $A$  (où les blocs  $\star$  ne sont pas impliqués dans les calculs) :

$$(3) \quad A = \begin{pmatrix} \star & \star \\ B_d & A_d \end{pmatrix} = \begin{pmatrix} A_g^r & B_g^r \\ \star & \star \end{pmatrix}.$$

On introduit  $R_g^r$ ,  $\tilde{R}_g^r$  et  $R_d$  les matrices telles que pour  $x = (x_1, \dots, x_J) \in \mathbb{R}^J$

$$(4) \quad R_g^r x = \begin{pmatrix} x_1 \\ \vdots \\ x_{J_*^r} \end{pmatrix} \in \mathbb{R}^{J_*^r}, \quad \tilde{R}_g^r x = \begin{pmatrix} x_1 \\ \vdots \\ x_{J_*} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbb{R}^{J_*^r}, \quad R_d x = \begin{pmatrix} x_{J_*+1} \\ \vdots \\ x_J \end{pmatrix} \in \mathbb{R}^{J-J_*}.$$

Si  $J_*^r = J_*$ , on pose simplement  $R_g^r = R_g = \tilde{R}_g^r$ . L'algorithme peut alors s'écrire sous la forme

$$(5) \quad U^{(n+1)} = (\tilde{R}_g^r)^T (A_g^r)^{-1} (R_g^r b - \tilde{B}_g^r U^{(n)}) + R_d^T (A_d)^{-1} (R_d b - \tilde{B}_d U^{(n)})$$

qui fait intervenir les matrices blocs  $\tilde{B}_g^r = (0_{J_*^r \times J_*^r} \quad B_g^r)$  et  $\tilde{B}_d = (B_d \quad 0_{J_* \times J_*})$ . Or on a  $R_g^r A = (A_g^r \quad B_g^r)$  et  $R_d A = (B_d \quad A_d)$ , ce qui donne<sup>2</sup>

$$(6) \quad U^{(n+1)} = Lb + (\mathbb{I} - LA)U^{(n)}, \quad L = (\tilde{R}_g^r)^T (A_g^r)^{-1} R_g^r + R_d^T (A_d)^{-1} R_d.$$

#### 4. Analyse de la convergence de la méthode

Ainsi, l'analyse de l'algorithme 2 se ramène à étudier la convergence de la méthode itérative (6), qu'on relie aux propriétés du rayon spectral de  $\mathbb{I} - LA$ . Commençons par le cas sans recouvrement  $J_* = J_*^r$ . L'algorithme correspond alors à la décomposition par blocs de la matrice  $A$

$$(7) \quad A = M - N, \quad M = \begin{pmatrix} A_g & 0 \\ 0 & A_d \end{pmatrix}, \quad N = - \begin{pmatrix} 0 & B_g \\ B_d & 0 \end{pmatrix} \geq 0,$$

où, à partir de maintenant, on note  $T \geq 0$  (resp.  $x \geq 0$ ) lorsque les coefficients de la matrice  $T \in \mathcal{M}_J(\mathbb{R})$  (resp. le vecteur  $x \in \mathbb{R}^J$ ) sont positifs, et  $T \geq S$  lorsque  $T - S \geq 0$ . On note que  $M$  est une  $\mathcal{M}$ -matrice et on a  $M^{-1}N \geq 0$  ainsi que  $A^{-1}N \geq 0$ . En écrivant  $A^{-1}N = (\mathbb{I} - M^{-1}N)^{-1} - \mathbb{I}$ , on observe que  $\tau$  est valeur propre de  $A^{-1}N$  si et seulement si  $\frac{\tau}{1+\tau}$  est valeur propre de  $M^{-1}N$ ,

2. en désignant par  $\mathbb{I}$  la matrice identité

avec les mêmes vecteurs propres associés. Or  $\rho(A^{-1}N)$  est valeur propre de  $A^{-1}N$  associée à un vecteur propre  $x \geq 0$ . On en déduit que

$$(8) \quad \rho(M^{-1}N) = \frac{\rho(A^{-1}N)}{1 + \rho(A^{-1}N)} < 1.$$

Pour aborder le cas  $J'_* > J_*$ , on définit  $E_g = (\tilde{R}_g^r)^\top R_g^r = R_g^\top R_g$  et  $E_d = R_d^\top R_d$ . On retient que  $E_g + E_d = \mathbb{1}$ , l'identité sur  $\mathbb{R}^J$ .

**Lemme 2.** *La matrice  $L$  est inversible :  $L = M^{-1} = E_g M_g^{-1} + E_d M_d^{-1}$  où les matrices  $M_j$  sont composées des blocs  $A_g^r$  et  $A_d$ , respectivement, complétés par les éléments diagonaux de  $A$ .*

**Preuve.** On remarque que  $(\tilde{R}_g^r)^\top = E_g (R_g^r)^\top$  donc

$$(9) \quad (\tilde{R}_g^r)^\top (A_g^r)^{-1} R_g^r = E_g (R_g^r)^\top (A_g^r)^{-1} R_g^r = E_g \begin{pmatrix} (A_g^r)^{-1} & 0 \\ 0 & 0 \end{pmatrix} = E_g \begin{pmatrix} (A_g^r)^{-1} & 0 \\ 0 & (\Delta_g^r)^{-1} \end{pmatrix}$$

qu'on écrit  $E_g M_g^{-1}$ , où  $\Delta_g^r = \text{diag}(A_{J-J'_*+1, J-J'_*+1}, \dots, A_{J, J})$ . Un calcul similaire peut être mené avec  $A_d$ . On déduit de ces expressions que  $L$  est injective.  $\square$

L'algorithme (6) s'interprète encore comme une méthode itérative associée à la décomposition  $A = M - N$  et la convergence dépend des propriétés spectrales de  $\mathbb{1} - M^{-1}A = M^{-1}N$ .

**Lemme 3.** *Soit  $A$  une  $\mathcal{M}$ -matrice qu'on décompose sous la forme  $A = M - N$  avec  $M$  une  $\mathcal{M}$ -matrice et  $M^{-1}N$  à coefficients positifs. Alors le rayon spectral de  $\mathbb{1} - M^{-1}A$  vérifie  $\rho(\mathbb{1} - M^{-1}A) < 1$ .*

**Preuve.** On pose  $H = M^{-1}N \geq 0$ . Comme  $M^{-1} = (\mathbb{1} - H)A^{-1}$  on a

$$(10) \quad 0 \leq (\mathbb{1} + H + \dots + H^m)M^{-1} = (\mathbb{1} - H^{m+1})A^{-1} \leq A^{-1}.$$

Or  $M^{-1}$  contient au moins un élément strictement positif sur chaque ligne. Il s'ensuit que  $\sum_{k=0}^{\infty} H^k$  converge, ce qui implique  $\rho(H) < 1$ .  $\square$

Pour appliquer le Lemme 3, le point crucial consiste à montrer que les matrices  $M_j$  sont des  $\mathcal{M}$ -matrices. Ceci est une conséquence du fait que  $A$  est une  $\mathcal{M}$ -matrice et que

$$(11) \quad A \leq M_j \leq D = \text{diag}(A_{1,1}, \dots, A_{J,J}).$$

En effet, on aura alors  $M^{-1} \geq 0$  et on déduit de  $(M_j)^{-1}A \leq \mathbb{1}$  que

$$(12) \quad M^{-1}N = \mathbb{1} - M^{-1}A = \mathbb{1} - (E_g M_g^{-1}A + E_d M_d^{-1}A) \geq \mathbb{1} - (E_g + E_d) = 0.$$

Ici, le Lemme 3 s'applique directement à la décomposition  $A = D - (D - A)$  :  $\rho(\mathbb{1} - D^{-1}A) < 1$ . La condition (11) implique que  $0 \leq \mathbb{1} - D^{-1}M_j \leq \mathbb{1} - D^{-1}A$ . On en déduit que  $\rho(\mathbb{1} - D^{-1}M_j) < 1$ , donc  $\sum_{k=0}^{\infty} (\mathbb{1} - D^{-1}M_j)^k$  converge et sa somme, qui n'est autre que  $M_j^{-1}D$ , est une matrice à coefficients positifs, d'où finalement,  $M_j^{-1} \geq 0$ . On peut aussi se rendre compte que le schéma obtenu en remplaçant  $(\tilde{R}_g^r)^\top$  par  $(R_g^r)^\top$  dans (6) ne converge pas. En effet, il existe des vecteurs  $e \in \mathbb{R}^J$  tels que  $e = (R_g^r)^\top R_g^r e = R_d^\top R_d e$ . Comme  $A_i = R_i A R_i^\top$ , il s'ensuit que  $(-1)$  est valeur propre de  $M^{-1}N$ .

## 5. Rôle du recouvrement

On peut justifier qu'il y a toujours intérêt à travailler avec un recouvrement. En effet, soit  $T_0 = M_0^{-1}N_0$  la matrice obtenue quand  $J_*^r = J_*$  et  $T = \mathbb{I} - M^{-1}A = M^{-1}N$  correspondant à un cas  $J_*^r > J_*$ . On note  $C_0 = N_0M_0^{-1} \geq 0$ . L'analyse repose sur les deux observations suivantes :

- d'une part, on a  $M^{-1}NA^{-1} = A^{-1} - M^{-1} = A^{-1}NM^{-1}$ ,
- d'autre part, il existe  $\phi \geq 0$ ,  $u \geq 0$  non nuls tels que  $C_0u = \rho(C_0)u$  et  $T^T\phi = \rho(T)\phi$ .

Alors, la relation

$$(13) \quad u \cdot (TA^{-1})^T\phi = \rho(T)A^{-1}u \cdot \phi \leq A^{-1}C_0u \cdot \phi = \rho(C_0)A^{-1}u \cdot \phi$$

permet de conclure que  $\rho(C_0) \geq \rho(T)$  puis  $\rho(T) \leq \rho(T_0) < 1$ .

On peut ensuite chercher à comparer numériquement (l'analyse étant encore essentiellement un problème ouvert) l'efficacité des méthodes associées à deux recouvrements différents. On s'aperçoit que, pour une erreur relative fixée, le nombre d'itérations décroît quand on augmente le recouvrement. D'un point de vue pratique, il faut aussi tenir compte du fait qu'élargir le recouvrement augmente aussi la taille des systèmes linéaires à résoudre, ce qui impacte les temps de calculs, notamment si on envisage une extension à des situations multi-dimensionnelles. Aussi l'efficacité de la méthode est affaire de compromis.

## 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.*
  - Expliquer en quoi l'algorithme 2 est plus satisfaisant au regard de l'objectif annoncé.
  - Illustrer les performances de la méthode et les énoncés du texte en s'inspirant des exemples donnés à la section 2. Comparer numériquement les algorithmes 1 et 2.
  - Détailler le lien entre le problème (1) et le système (2). Discuter la preuve du Lemme 1 et expliquer son intérêt en lien avec la résolution de (1). On pourra éventuellement remplacer l'équation différentielle de (1) par  $\lambda u - u'' = f$  avec  $\lambda > 0$ , certains arguments pouvant être alors plus simples. Il est aussi loisible de proposer des démonstrations différentes des mêmes faits.
  - Détailler la décomposition par blocs (3) et l'expression de l'algorithme sous la forme (5) et (6). On pourra s'appuyer sur des dessins.
  - Justifier la convergence de l'algorithme en expliquant le rôle du rayon spectral et en détaillant notamment le Lemme 3.
  - Analyser, théoriquement et/ou numériquement, le rôle du recouvrement.
  - Expliquer l'extension à une situation bi-dimensionnelle évoquée à la fin de la section 1.