Diffusion thermique dans des milieux hétérogènes. Problèmes de conditionnement et de préconditionnement.

On s'intéresse dans ce texte à la répartition de température dans une pièce mécanique plongée dans un milieu dont la température est connue en tout point.
On rappelle tout d'abord la loi de Fourier qui décrit les échanges de chaleur dans un matériau en fonction de la répartition de température. Cette loi nous dit qu'en tout point x du système physique considéré (donc en dimension 3 comme il se doit), si on se donne une surface plane élémentaire (c'est-à-dire très petite) notée δS et orientée par un vecteur unitaire normal $\vec{n}$ (voir figure 1), alors la quantité de chaleur qui traverse la surface δS dans le sens indiqué par $\vec{n}$ **par unité de temps** est donnée par

$$\delta Q = -k(x)(\nabla T(x), \vec{n}),$$

où $\nabla T(x)$ désigne le gradient de la température T au point x et $(\cdot, \cdot)$ désigne le produit scalaire usuel de $\mathbb{R}^3$. Le réel strictement positif $k(x)$ dépend du point considéré et est appelé conductivité thermique au point x du milieu étudié. Le signe $-$ dans cette expression indique que la chaleur passe des zones chaudes aux zones froides et non l'inverse.

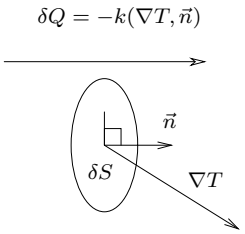


Figure 1: Surface élémentaire δS et loi de Fourier

Sauf précision contraire, on supposera en toute généralité que le système considéré est hétérogène au sens où sa conductivité thermique $k(x)$ est vraiment une fonction non constante de x.

Cette situation peut se produire si la pièce mécanique étudiée présente des usures ou des défauts, mais aussi de façon plus flagrante si elle est composée de plusieurs parties constituées de matériaux de natures vraiment différentes. Les exemples de telles situations sont nombreux : on peut par exemple penser à une pièce de moteur de voiture composée de deux alliages différents, ou encore à un système d'isolation thermique (de type double-vitrage) ...

Ce texte pésente certaines difficultés numériques que l'on rencontre lors de la discrétisation de l'équation vérifiée par la répartition de température dans le milieu étudié.

Le point-clé de l'étude qui va suivre est la notion de conditionnement.

DÉFINITION 1

Le conditionnement d'une matrice inversible A (ou plus précisément d'un système linéaire dont la matrice est A) est défini par :

$$Cond(A) = \|A\|.\|A^{-1}\|.$$

La norme choisie est ici la norme matricielle associée à la norme euclidienne sur $\mathbb{R}^n$.

Dans le cas d'une matrice symétrique définie positive, le conditionnement ainsi défini n'est autre que le rapport entre la plus grande et la plus petite des valeurs propres de la matrice.
On peut montrer que le conditionnement d'un système linéaire mesure la sensibilité du système aux variations dans les données (c'est-à-dire dans les coefficients du système, ou dans le second membre) et donc par conséquent la sensibilité du système aux erreurs d'arrondis qui apparaissent lorsqu'on utilise des méthodes numériques pour le résoudre. On verra dans la suite que le conditionnement mesure également l'efficacité de certaines méthodes de résolution du système linéaire.

1 Diffusion thermique dans le cas monodimensionnel

1.1 Obtention du modèle

Pour fixer les idées, on va supposer pour l'instant que le problème est monodimensionel. Plus précisément on suppose qu'on a affaire à une barre rectiligne (figure 2) modélisée par l'intervalle $[0, 1]$ de $\mathbb{R}$ dont toutes les caractéristiques (notament la conductivité et la température) ne dépendent que de l'abscisse x. On note S la surface de la section de la barre.
On suppose qu'elle est soumise, à l'abscisse x, à un apport de chaleur extérieur (rayonnement, radiation, moyen de refroidissement, réactions chimiques, etc ...) par unité de volume et par unité de temps noté $f(x)$. Déterminons l'équation vérifiée par la distribution de température $T(x)$ dans cette barre une fois qu'elle a atteint l'équilibre thermique.
Considérons la partie de la barre située entre les abscisses $x - \delta x$ et $x + \delta x$ (avec δx petit).
Celle-ci étant en équilibre thermique, on doit donc écrire que la chaleur totale échangée par cette section (par unité de temps) avec le reste de la barre et avec l'extérieur est nulle. La quantité de

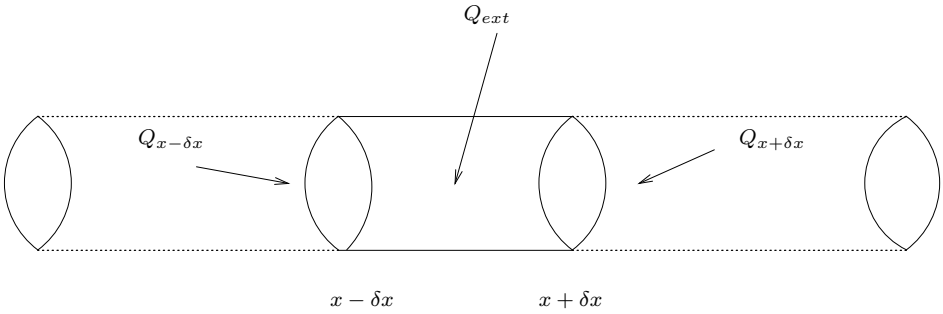


Figure 2: Section $[x-\delta x, x+\delta x]$ de la barre étudiée

chaleur **acquise** par unité de temps par la partie $[x-\delta x, x+\delta x]$ de la barre à travers son extrémité gauche nous est donnée par la loi de Fourier par :

$$Q_{x-\delta x} = -k(x-\delta x)T'(x-\delta x)S,$$

où $k(x-\delta x)$ est la conductivité thermique de la barre à l'abscisse $x-\delta x$. De la même façon, la quantité de chaleur **acquise** par unité de temps à travers son extrémité droite est donnée par

$$Q_{x+\delta x} = k(x+\delta x)T'(x+\delta x)S.$$

Par ailleurs, la chaleur acquise par unité de temps, par le morceau de barre étudié à partir de l'extérieur est donnée par

$$Q_{ext} = S\int_{x-\delta x}^{x+\delta x} f(u)\,du.$$

L'équilibre thermique s'écrit alors $Q_{x-\delta x} + Q_{x+\delta x} + Q_{ext} = 0$, ce qui donne l'équation

$$k(x+\delta x)T'(x+\delta x) - k(x-\delta x)T'(x-\delta x) + \int_{x-\delta x}^{x+\delta x} f(u)\,du = 0.$$

En divisant ceci par $2\delta x$ et faisant tendre δx vers 0 on obtient **formellement** l'équation différentielle

$$-\frac{d}{dx}\big(k(x)T'(x)\big) = f(x). \tag{1}$$

La température de la barre est par exemple fixée à ces deux extrémités, on ajoute donc à cette équation les conditions aux limites

$$T(0) = \alpha,\ \ T(1) = \beta.$$

L'équation (1) est appelée **équation de Laplace non homogène**.

1.2 Résolution numérique du problème 1D

Pour résoudre numériquement cette équation, on divise l'intervalle $[0,1]$ en $n+1$ intervalles de même longueur $h = \frac{1}{n+1}$ et on note $x_i = ih$, pour $0 \leq i \leq n+1$ les points de cette subdivision. On se propose de résoudre l'équation de Laplace non homogène (1) grâce au schéma classique

$$\begin{cases} -\frac{1}{h}\left(k\left(\frac{x_{i+1}+x_i}{2}\right)\frac{T_{i+1}-T_i}{h} - k\left(\frac{x_i+x_{i-1}}{2}\right)\frac{T_i-T_{i-1}}{h}\right) = f(x_i), \ \ \forall 1 \leq i \leq n, \\ T_0 = \alpha, \\ T_{n+1} = \beta. \end{cases}$$

Quand la conductivité thermique $x \mapsto k(x)$ est constante (égale à K), on montre analytiquement que le système linéaire ci-dessus a un conditionnement indépendant de la constante K et asymptotiquement proportionnel à n^2.
Par contre, dans le cas d'un matériau vraiment hétérogène (supposons par exemple que $k(x) = K_1$ pour $x \in [0, \frac{1}{2}]$, et $k(x) = K_2$ pour $x \in]\frac{1}{2}, 1]$), alors le conditionnement du système dépend fortement du rapport $\frac{K_1}{K_2}$. Dès que ce rapport devient vraiment grand, le conditionnement du système explose (à n fixé).

Pour résoudre le système précédent que l'on écrit sous une forme matricielle générale

$$At = f, \tag{2}$$

on peut envisager une méthode de type gradient à pas optimal ou gradient conjugué. On sait que la méthode du gradient à pas optimal a une erreur qui se comporte en

$$\left(\frac{Cond(A)-1}{Cond(A)+1}\right)^{2k},$$

où k est le nombre d'itérations. La méthode du gradient conjugué, elle, a une erreur en

$$\left(\frac{\sqrt{Cond(A)}-1}{\sqrt{Cond(A)}+1}\right)^{2k}.$$

Ainsi, pour des matrices à grand conditionnement, la méthode du gradient conjugué se comporte mieux que celle du gradient à pas optimal. On sait par ailleurs que la méthode du gradient conjugué converge, en théorie, en au plus n itérations.
Malheureusement, cette dernière propriété théorique est mise en défaut numériquement dès qu'on traite des matrices ayant un trop grand conditionnement, comme cela peut-être le cas dans l'exemple introduit ci-dessus. Les deux méthodes peuvent alors s'avérer assez inefficaces.

1.3 Principe général des méthodes de préconditionnement

Pour passer outre ce problème, nous allons mettre en place une méthode de préconditionnement. De manière générale, l'idée est de remplacer la résolution du système (2) par la résolution d'un système éuivalent

$$C^{-1}At = C^{-1}f,$$

où on choisit la matrice C (matrice de préconditionnement) qui soit "facilement inversible" et qui soit suffisament "proche" de la matrice A pour que la matrice $C^{-1}A$ soit proche de l'identité et donc que son conditionnement soit "proche" de 1. Bien entendu, toutes ses notions sont assez intuitives et seront partiellement précisées dans la suite.
Dans le cas où la matrice A est symétrique définie positive, il est plus astucieux de choisir la matrice C symétrique définie positive également et de procéder à un préconditionnement un peu différent en remplaçant le problème initial par le problème préconditionné suivant :

$$C^{-\frac{1}{2}}AC^{-\frac{1}{2}}y = C^{-\frac{1}{2}}f, \quad t = C^{\frac{1}{2}}y, \tag{3}$$

et en essayant de faire en sorte que $Cond(C^{-\frac{1}{2}}AC^{-\frac{1}{2}}) << Cond(A)$.
On peut alors appliquer les méthodes du gradient à pas optimal et du gradient conjugué à ce nouveau problème (3). Ces algorithmes peuvent alors s'écrire, **dans les variables initiales** (c'est-à-dire avec l'inconnue t et le résidu $r = f - At$ du problème de départ (2)) :

1. Algorithme du gradient à pas optimal préconditionné :

$$\text{Initialisation : } \quad r_0 = f - At_0,$$

$$\text{Itérations : } \begin{cases} \alpha_k = \dfrac{(C^{-1}r_k, r_k)}{(AC^{-1}r_k, C^{-1}r_k)}, \\ t_{k+1} = t_k + \alpha_k C^{-1}r_k, \\ r_{k+1} = r_k - \alpha_k AC^{-1}r_k. \end{cases}$$

2. Algorithme du gradient conjugué préconditionné :

$$\text{Initialisation : } \begin{cases} r_0 = f - At_0 \\ p_0 = C^{-1}r_0. \end{cases}$$

$$\text{Itérations : } \begin{cases} \alpha_k = \dfrac{(C^{-1}r_k, r_k)}{(Ap_k, p_k)}, \\ t_{k+1} = t_k + \alpha_k p_k, \\ r_{k+1} = r_k - \alpha_k Ap_k, \\ \beta_{k+1} = \dfrac{(C^{-1}r_{k+1}, r_{k+1})}{(C^{-1}r_k, r_k)}, \\ p_{k+1} = C^{-1}r_{k+1} + \beta_{k+1}p_k. \end{cases}$$

1.4 Choix des matrices de préconditionnement

Il reste maintenant à choisir les matrices de préconditionnement. Les critères de choix sont assez clairs :

- Il nous faut pouvoir résoudre des systèmes linéaires $Cy = d$ de façon extrêmement efficace.
- Il faut que la matrice C "contienne" suffisament d'informations de la matrice A pour que la matrice $C^{-\frac{1}{2}}AC^{-\frac{1}{2}}$ ait un conditionnement faible. Il faut donc que la matrice C contienne les parties les plus significatives de la matrice A qui sont la cause du mauvais conditionnement de la matrice A.

Ensuite, le choix de C est très empirique. Dans le cas du problème présenté ci-dessus, on peut penser dans un premier temps à prendre pour C, la matrice diagonale extraite de A. Cette méthode de préconditionnement fonctionne assez bien dans l'exemple proposé ici. Ceci est essentiellement dû à la forme particulièrement simple de la matrice obtenue sur le problème 1D. D'ailleurs, on peut remarquer que pour traiter ce problème 1D, il existe des méthodes plus performantes pour résoudre le système linéaire proposé. L'utilisation de méthodes de gradient dans ce cadre est purement théorique et sert ici à illustrer des problèmes plus généraux.

2 Cas de la dimension 2

2.1 Obtention du modèle

Reprenons l'exemple de l'équation de Laplace non-homogène décrivant la répartition de température dans un matériau hétérogène. En dimension 2, si on reprend la méthode de modélisation proposée ci-dessus, l'équation différentielle (1) devient une équation aux dérivées partielles qui s'écrit

$$-\text{div}\left(k(x)\nabla T(x)\right) = f(x), \tag{4}$$

avec des conditions aux limites $T = T_b$ sur le bord du domaine considéré. On rappelle que la divergence d'un champ de vecteurs de coordonnées $(u_1, u_2)^t$ dans la base canonique de $\mathbb{R}^2$, est donnée par

$$\text{div}\,(u) = \frac{\partial u_1}{\partial x_1} + \frac{\partial u_2}{\partial x_2}.$$

L'équation (4) s'obtient, en écrivant que le bilan des échanges de chaleur par unité de temps entre n'importe quel volume Ω dans le matériau étudié et le milieu qui l'entoure est nul. On utilise alors la loi de Fourier puis la formule de Stokes (ou théorème de la divergence) qui dit que pour tout champ de vecteur u on a

$$\int_\Omega \text{div}\,(u)\,dx = \int_{\partial\Omega} (u, \vec{n})\,dS,$$

où $\vec{n}$ est la normale unitaire sortante au domaine Ω.

2.2 Discrétisation bidimensionelle

On suppose, pour simplifier, que le domaine considéré est le carré $[0,1] \times [0,1]$. On considère un maillage régulier de ce carré constitué des points $x_{i,j}$ de coordonnées (ih, jh) pour $0 \leq i, j \leq n+1$ et pour $h = \frac{1}{n+1}$ (voir la figure 3). En un point $x_{i,j}$, avec $1 \leq i, j \leq n$, l'équation (4) est discrétisée, par une méthode de différences finies tout à fait analogue au cas monodimensionnel de la façon suivante :

$$\begin{aligned} &-\frac{1}{h}\left(k\left(\frac{x_{i+1,j}+x_{i,j}}{2}\right)\frac{T_{i+1,j}-T_{i,j}}{h} - k\left(\frac{x_{i,j}+x_{i-1,j}}{2}\right)\frac{T_{i,j}-T_{i-1,j}}{h}\right) \\ &-\frac{1}{h}\left(k\left(\frac{x_{i,j+1}+x_{i,j}}{2}\right)\frac{T_{i,j+1}-T_{i,j}}{h} - k\left(\frac{x_{i,j}+x_{i,j-1}}{2}\right)\frac{T_{i,j}-T_{i,j-1}}{h}\right) \\ &= f(x_{i,j}). \end{aligned} \tag{5}$$

Les valeurs au bord de $T_{i,j}$ (i.e. pour $i = 0$ et $i = n+1$ ainsi que pour $j = 0$ et $j = n+1$) sont prescrites par les conditions aux limites que l'on choisit.

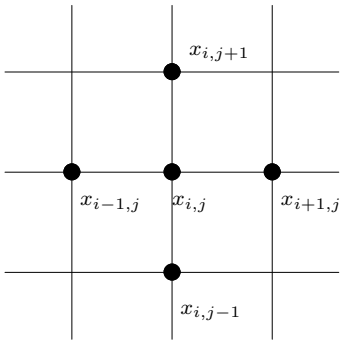
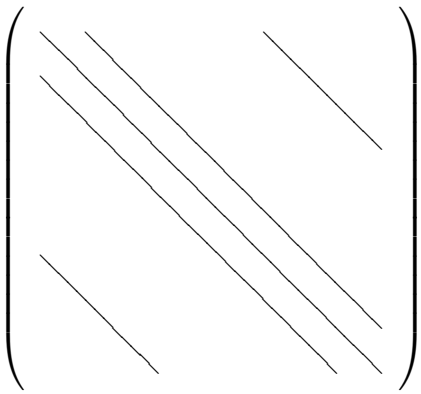


Figure 3: Description du maillage

Si on range les inconnues $(T_{i,j})_{1 \leq i,j \leq n}$ dans un vecteur colonne de la façon naturelle suivante

$$t = \begin{pmatrix} T_{1,1} \\ T_{2,1} \\ \vdots \\ T_{n,1} \\ T_{1,2} \\ \vdots \\ T_{n,2} \\ \vdots \\ \vdots \\ T_{n,n} \end{pmatrix},$$

on constate que la matrice du système linéaire (5) a une structure avec 5 diagonales non nulles :



Ainsi, les méthodes spécifiques au cas particulier du problème 1D ne s'appliquent plus ici et on voit bien qu'un préconditionnement diagonal ne va pas être d'une efficacité optimale car elle ne tient pas bien compte des 2 diagonales "extérieures" de la matrice.
Pour obtenir un préconditionnement de meilleure qualité, on s'inspire des méthodes itératives classiques. Plus précisément, si on écrit la matrice A sous la forme

$$A = L + D + L^t,$$

où L est la partie strictement sous-diagonale de A et D sa partie diagonale, la méthode de Gauss-Seidel consiste à écrire l'équation $At = f$ sous la forme

$$(L + D)t = -L^t t + f,$$

puis

$$t = -(L + D)^- L^t t + (L + D)^{-1} f, \tag{6}$$

et donc à faire une procédure de point fixe donnée par

$$t_{k+1} = -(L + D)^{-1} L^t t_k + (L + D)^{-1} f.$$

On constate que l'équation de base de la méthode de Gauss-Seidel (6) peut s'écrire sous la forme

$$(L + D)^{-1} At = (L + D)^{-1} f.$$

Ceci nous incite donc à utiliser la matrice $L + D$ comme préconditionnement pour la matrice A. Ainsi, dans les algorithmes préconditionnés exposés plus haut, "appliquer C^{-1}" revient à "faire une itération de la méthode de Gauss-Seidel".
Malheureusement la matrice $L + D$ n'est pas symétrique, ce qui comme on l'a vu n'est pas très recommandé dans le cas de la matrice A donnée par (5). Ceci nous donne l'idée de se rapporter plutôt à la version "symétrisée" de la méhode de Gauss-Seidel. Uune itération de cette méthode consiste à effectuer une itération de Gauss-Seidel puis une itération de la transposée de la méthode de Gauss-Seidel. Les itérations de la méthode s'écrivent donc

$$\begin{cases} t_{k+\frac{1}{2}} = -(L + D)^{-1} L^t t_k + (L + D)^{-1} f, \\ t_{k+1} = -(L^t + D)^{-1} L t_{k+\frac{1}{2}} + (L^t + D)^{-1} f, \end{cases}$$

ce qui fournit l'expression complète d'une itération

$$t_{k+1} = -(L^t + D)^{-1} L\big(-(L + D)^{-1} L^t t_k + (L + D)^{-1} f\big) + (L^t + D)^{-1} f,$$

ou encore

$$t_{k+1} = (L^t + D)^{-1} L(L + D)^{-1} L^t t_k + (L^t + D)^{-1} D(L + D)^{-1} f.$$

Ainsi, la méthode "symétrisée" peut-être vue comme une méthode de point fixe sur l'équation

$$t = (L^t + D)^{-1} L(L + D)^{-1} L^t t + (L^t + D)^{-1} D(L + D)^{-1} f. \tag{7}$$

Finalement, le système $At = f$, traité par la méthode de Gauss-Seidel symétrisée s'écrit sous la forme

$$(L^t + D)^{-1} D(L + D)^{-1} At = (L^t + D)^{-1} D(L + D)^{-1} f,$$

ce qui nous fournit une idée de matrice de préconditionnement possible

$$C = (L + D)D^{-1}(L^t + D).$$

De la même façon, en s'inspirant de la méthode de relaxation (SSOR), on constate que l'on peut considérer toute une famille de matrices de préconditionnement

$$C_\omega = (\omega L + D)D^{-1}(\omega L^t + D),$$

le paramètre ω donne un poids à la partie hors diagonale de la matrice A. Ce paramètre doit être ajusté dans chaque situation pour adapter la méthode à la matrice étudiée.
Quand on utilise la méthode du gradient à pas optimal préconditionné ou la méthode du gradient conjugué préconditionné sur des problèmes du type (5) où la conductivité $k(x)$ est une fonction qui présente des fortes discontinuités (ou des forts gradients), on constate aisément que les matrices C_ω (pour ω bien choisi) fournissent un meilleur résultat (en terme de vitesse de convergence de la méthode) que le simple préconditionnement diagonal.

Quelques développements possibles ...

- Illustrer numériquement les problèmes de conditionnement sur les exemples issus du texte. On pourra faire varier les contrastes de conductivité $\frac{K_1}{K_2}$.
- Quelles méthodes de résolution plus adaptées au problème 1D peut-on envisager en lieu et place des méthodes de gradient proposées dans le texte ? Pourquoi ces méthodes ne sont plus compétitives dans le cas 2D.
- Vérifier que le conditionnement de $C^{-1}A$ (ou de $C^{-\frac{1}{2}}AC^{-\frac{1}{2}}$) est effectivement meilleur que celui de A.
- Justifier les vitesses de convergence des méthodes de gradient à pas optimal et de gradient conjugué.

- Justifier les algorithmes préconditionnés proposés dans le texte. Quel est l'intérêt de l'écriture proposée pour ces algorithmes ? Quel est le coût supplémentaire engendré par la technique de préconditionnement ?
- Comparer les divers choix possibles de matrices de préconditionnement en terme de nombres d'itérations des méthodes, coût de calcul, etc ...