Équation de la chaleur : traitement numérique

1 Position du problème

Nous nous intéressons ici au traitement numérique de l'équation de la chaleur :

$$\pi^2 \frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = 0$$

Ce problème, avec en outre une condition initiale (la donnée de $u(0,x)$ pour tout x), est à l'origine de nombreux problèmes souvent difficiles, et pourrait nous faire couler beaucoup d'encre.

Pour être plus concret dans ce qui suit, nous allons nous restreindre à un exemple précis, dans lequel on connaît bien la solution exacte. Ainsi, nous serons à même de comparer les résultats numériques donnés par les différentes méthodes présentées à la solution exacte, et donc d'aprécier leur efficacité et leur précision.

Nous considérons ce problème sur un domaine borné en espace ($0 \leq x \leq 1$), et pour des temps positifs ($t \geq 0$). Nous allons également nous placer dans des conditions initiales relativement simples :

$$\forall x \in [\,0\,;1\,], \quad u(0,x) = \sin(\pi x)$$

On impose également les conditions de bord :

$$\forall t \geq 0, \quad u(t,0) = u(t,1) = 0$$

On vérifie alors aisément que la fonction $(t,x) \longmapsto e^{-t}\sin(\pi x)$ est solution de notre problème. Nous pourrons donc comparer les solutions approchées que nous allons calculer à cette fonction de référence.

N.B. S'il est aisé de vérifier par dérivation que notre fonction vérifie bien l'équation de la chaleur, et qu'elle satisfait par ailleurs aux conditions initiales et aux conditions de bord imposées, l'unicité d'une telle solution est un problème autrement plus complexe.

Par ailleurs, pour des données initiales moins régulières, on construit sans trop de problème une solution, mais là encore l'unicité dépend d'un certain nombre de paramètres, notamment de régularité de la condition initiale.

2 Résolution numérique de l'équation

Bien que nous connaissions ici la solution exacte du problème, ou plutôt justement parceque nous la connaissons, il est instructif de voir comment l'on peut la calculer numériquement, ou plus exactement l'approcher numériquement.

Aussi puissant que soit votre ordinateur, un programme de calcul numérique ne pourra jamais calculer qu'une quantité finie de valeurs de la fonction solution du problème. Il nous faut donc comme toujours discrétiser le problème en espace ET en temps.

On se donne donc un pas spatial h, c'est-à-dire que l'on cherche à calculer les valeurs de notre solution u aux points d'abscisse $x = 0, h, 2h, \ldots, 1$. On doit aussi se restreindre à un ensemble discret d'instants auxquels on prétend calculer ces valeurs, on fixe donc un pas temporel k et l'on cherchera à calculer les valeurs de u aux instants $t = k, 2k, 3k, \ldots$.

Dans toute la suite, nous noterons u_n^i la valeur calculée (approchée) de la solution u à l'instant nk ($n > 0$) et au point d'abscisse ih ($0 \leq i \leq \frac{1}{h}$). Discrétiser l'équation, c'est alors transformer ses termes en des expressions faisant intervenir uniquement nos valeurs u_n^i.

2.1 Un première méthode

La dérivée temporelle $\frac{\partial u}{\partial t}$ au point (nk, ih) peut être approchée par le rapport $\frac{u\left((n+1)k, ih\right) - u\left(nk, ih\right)}{k}$. En effet, un développement limité à l'ordre 1 nous donne :

$$\frac{u\left((n+1)k, ih\right) - u\left(nk, ih\right)}{k} = \frac{\partial u}{\partial t}\left(nk, ih\right) + \mathrm{o}\left(1\right)$$

Remplaçant les valeurs de la fonction $u\left((n+1)k, ih\right)$ et $u\left(nk, ih\right)$ par leurs valeurs approchées u_{n+1}^i et u_n^i, on approchera donc $\frac{\partial u}{\partial t}\left(nk, ih\right)$ par le rapport $\frac{u_{n+1}^i - u_n^i}{k}$.

De même, un développement limité à l'ordre 2 par rapport à la variable x au point (nk, ih) nous donne :

$$u\left(nk, (i+1)h\right) = u\left(nk, ih\right) + h\left(\frac{\partial u}{\partial x}\right)\left(nk, ih\right) + \frac{h^2}{2}\frac{\partial^2 u}{\partial x^2}\left(nk, ih\right) + h^2\mathrm{o}\left(1\right)$$

et $u\left(nk, (i-1)h\right) = u\left(nk, ih\right) - h\left(\frac{\partial u}{\partial x}\right)\left(nk, ih\right) + \frac{h^2}{2}\frac{\partial^2 u}{\partial x^2}\left(nk, ih\right) + h^2\mathrm{o}\left(1\right)$

Par sommation, on trouve donc :

$$\frac{u\left(nk, (i+1)h\right) + u\left(nk, (i-1)h\right) - 2u\left(nk, ih\right)}{h^2} = \frac{\partial^2 u}{\partial x^2}\left(nk, ih\right) + \mathrm{o}\left(1\right)$$

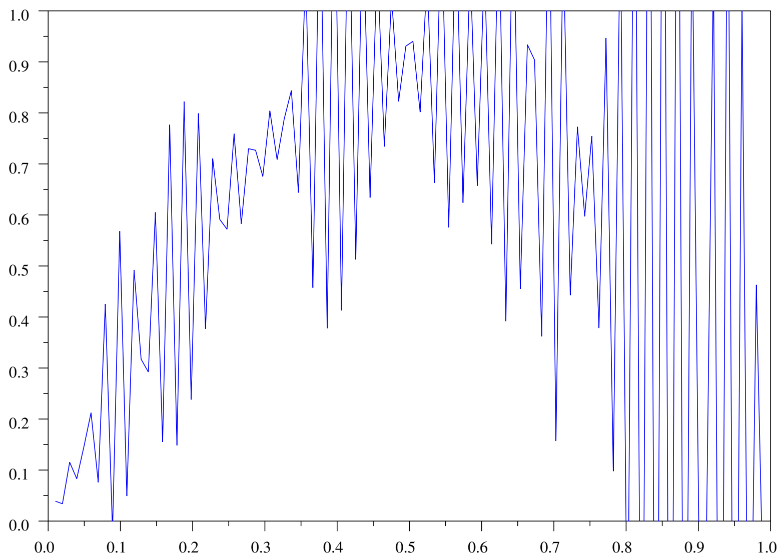
D'où l'idée, encore une fois, de remplacer le terme $\frac{\partial^2 u}{\partial x^2}$ de l'équation par le rapport $\frac{u_n^{i+1} + u_n^{i-1} - 2u_n^i}{h^2}$.

On obtient, pour tous les couples d'indices (n, i), l'équation approchée :

$$\boxed{\pi^2\frac{u_{n+1}^i - u_n^i}{k} - \frac{u_n^{i+1} + u_n^{i-1} - 2u_n^i}{h^2} = 0} \qquad \textbf{(1)}$$

Grâce à ce schéma, on peut calculer de proche en proche les valeurs de $(u_n^i)_{0\leq i\leq \frac{1}{h}}$ sans trop de difficultés : connaissant toutes ces valeurs à l'instant nk, chacune de ces équations (pour une valeur de i donnée) nous permet de calculer u_{n+1}^i. Seul se pose un petit problème pour $i=0$ et $i=\dfrac{1}{h}$, valeurs pour lesquelles u_n^{i-1} et u_n^{i+1} respectivement ne sont pas définies. Mais comme nous avons imposé $u(t,0)=u(t,1)=0$ pour tout t, il suffit à chaque étape de calcul de poser $u_{n+1}^0=u_{n+1}^{\frac{1}{h}}=0$ et de résoudre les équations pour i variant de 1 à $\dfrac{1}{h}-1$ uniquement...

C'est ce procédé de calcul qu'utilise le programme chaleurexpl. Voyons ce qu'il donne, pour $h=\dfrac{1}{100}$ et $k=\dfrac{1}{100}$ par exemple. Après 10 étapes de calculs ($t=0,1$ donc), ça commence à devenir franchement anarchique :



(voir également l'animation java)

Conclusion : même si les premiers pas de calcul sont relativement satisfaisant, on se rend compte qu'au bout de quelques temps, ça ne marche plus du tout ! Et ceci quel que soient les pas h et k choisi. En raffinant ceux-ci, on peut au mieux retarder le moment où la solution diverge complètement.

2.2 Essayons autre chose

Si l'approximation du terme $\dfrac{\partial^2 u}{\partial x^2}$ semble assez satisfaisante car après tout très symétrique, il n'en va pas de même pour le terme $\dfrac{\partial u}{\partial t}$: pourquoi l'approcher par le rapport $\dfrac{u_{n+1}^i - u_n^i}{k}$ plutôt que par $\dfrac{u_n^i - u_{n-1}^i}{k}$? Testons donc cette solution. Elle nous conduit aux équations :

$$\pi^2 \frac{u_n^i - u_{n-1}^i}{k} - \frac{u_n^{i+1} + u_n^{i-1} - 2u_n^i}{h^2} = 0$$

c'est-à-dire, en remplaçant n par $n+1$ (*i.e.* en changeant de point) :

$$\boxed{\pi^2 \frac{u_{n+1}^i - u_n^i}{k} - \frac{u_{n+1}^{i+1} + u_{n+1}^{i-1} - 2u_{n+1}^i}{h^2} = 0} \qquad \textbf{(2)}$$

Cette fois, le calcul des valeurs $(u_{n+1}^i)_{0\leq i\leq \frac{1}{h}}$ à partir des $(u_n^i)_{0\leq i\leq \frac{1}{h}}$ est un peu plus compliqué : chaque équation fait intervenir trois inconnues. Nous sommes donc face à un système à résoudre, ou ce qui revient au même, face à une matrice à inverser. Heureusement, de nombreux algorithmes font ça de manière assez efficace, et l'on peut donc programmer cette méthode sans trop de difficultés.

C'est le procédé qu'utilise le programme chaleurimpl. Pour que la comparaison avec le précédent ait un sens, prenons les mêmes valeurs pour les pas h et k (toujours $\dfrac{1}{100}$). Au temps $t = 0,1$, soit après 10 étapes de calcul, l'erreur maximale par rapport à la solution exacte est de seulement $4,57.10^{-4}$, elle est d'environ 2.10^{-3} après 100 étapes de calcul. Rien à voir donc avec le comportement anarchique de la solution trouvée par la première méthode ! (voir également l'animation java)

Conclusion : Bien que l'équation semble à première vue très semblable à la précédente, encore qu'un peu plus difficile à résoudre, cette méthode est autrement plus efficace et précise. On s'aperçoit que l'écart à la solution exacte est plus que raisonnable, et ce pendant très longtemps. Bien sûr, les erreurs d'approximation dans les calculs s'accumulant (le terme en $o(h^2)$ négligé lors de l'approximation plus des erruers numériques dues au fait que l'ordinateur ne prends en compte qu'un nombre fini de décimales), on va peu à peu s'écarter (en relatif) de la solution exacte de l'équation de la chaleur, mais cette méthode permet d'obtenir des résultats tout à fait exploitables.

N.B. En réalité, on reste assez proche. En effet, la solution numérique, tout comme la solution exacte, tend exponentiellement vers 0. Même pour des temps très grands, elles ne se séparent donc que très peu...

2.3 Pourquoi cette différence ?

La question est tout à fait légitime : dans un cas comme dans l'autre, on approxime la dérivée temporelle par une expression du type $\dfrac{u_{n+1}^i - u_n^i}{k}$ et la dérivée seconde en espace par quelquechose du type $\dfrac{u_n^{i+1} + u_n^{i-1} - 2u_n^i}{h^2}$. Les deux équations obtenues sont après tout très semblables, à quelques indices près... Transformons légèrement les équations obtenues.

<u>Convergence de la première méthode :</u> L'équation **(1)** se réécrit en

$$\begin{aligned} u_{n+1}^i &= u_n^i + \frac{k}{\pi^2 h^2}\left(u_n^{i+1} + u_n^{i-1} - 2u_n^i\right) \\ &= \left(1 - \frac{2k}{\pi^2 h^2}\right) u_n^i + \frac{k}{\pi^2 h^2}\left(u_n^{i+1} + u_n^{i-1}\right) \end{aligned}$$

Cette méthode est dite explicite, car elle fournit des formules explicites pour le calculs des valeurs u_{n+1}^i à partir des valeurs de u_n^i $(0 \leq i \leq \dfrac{1}{h})$.

Lorsque les pas k et h vérifient la condition $\dfrac{2k}{\pi^2 h^2} < 1$, aucun problème. La valeur u_{n+1}^i est alors une combinaison convexe (*i.e.* un barycentre à coefficients positifs) des valeurs u_n^{i-1}, u_n^i et u_n^{i+1}. Comme nous partons d'une fonction positive, la solution calculée reste constamment positive, on peut alors passer aux valeurs absolues et l'on obtient :

$$\left|u_{n+1}^i\right| \leq \max\left(\left|u_n^{i-1}\right|, \left|u_n^i\right|, \left|u_n^{i+1}\right|\right)$$

et donc
$$\max_i\left(\left|u_{n+1}^i\right|\right) \leq \max_i\left(\left|u_n^i\right|\right)$$

c'est-à-dire que notre solution calculée décroit, tout comme la solution exacte. Par ailleurs, on a nécessairement une courbe "lisse", u_{n+1}^i étant une combinaison convexe des valeurs u_n^{i-1}, u_n^i et u_n^{i+1}.

Mais cette condition sur h et k est pour le moins difficile à réaliser. En effet, il est naturel de prendre un pas spatial (h) petit, pour avoir de bonnes approximations de la dérivée seconde selon x. Ici, nous avons pris $h = \dfrac{1}{100}$. Il faudrait donc prendre un pas temporel k vérifiant :

$$k < \frac{10^{-4}\pi^2}{2} \simeq 5.10^{-4}$$

Pour un tel pas, le calcul de la solution approchée avance pour le moins lentement... Et si k ne vérifie pas cette condition, on a des barycentre à coefficients négatifs et tous nos calculs peuvent diverger assez rapidement.

<u>Convergence de la seconde méthode :</u> Passons à la seconde méthode, dite implicite. L'équation **(2)** se réécrit en :

$$\left(1 + \frac{2k}{\pi^2 h^2}\right) u_{n+1}^i - \frac{k}{\pi^2 h^2}\left(u_{n+1}^{i+1} + u_{n+1}^{i-1}\right) = u_n^i$$

On obtient cette fois les coefficients u_{n+1}^i en résolvant un système de la forme :

$$\left(\mathrm{I}_n + \frac{k}{\pi^2 h^2}\mathrm{A}\right)\mathrm{U}_{n+1} = \mathrm{U}_n$$

où I_n est la matrice identité, U_m le vecteur de coordonnées $(u_m^i)_{0 \leq i \leq \frac{1}{h}}$; et la matrice A est la matrice tridiagonale ayant des 2 sur la diagonale et des -1 sur la sur-diagonale et la sous-diagonale. (À l'exception du premier et du dernier coefficient de chacune de ces diagonales, car $u_n^0 = u_n^{\frac{1}{h}} = u_{n+1}^0 = u_{n+1}^{\frac{1}{h}} = 0$, donc ces coefficients sont des 0).

La matrice $\mathrm{I}_n + \dfrac{k}{\pi^2 h^2}\mathrm{A}$ est une matrice à diagonale dominante, on montre qu'elle est inversible et que son inverse est à coefficients positifs (voir en annexe). C'est-à-dire que cette fois, quels que soient les valeurs de h et k, les nombres u_{n+1}^i sont des combinaisons convexes des $(u_n^j)_{0 \leq j \leq \frac{1}{h}}$.

On a donc cette fois aussi, et sans conditions sur h et k, une propriété du type :

$$\boxed{\max_i\left(\left|u_{n+1}^i\right|\right) \leq \max_i\left(\left|u_n^i\right|\right)}$$

c'est-à-dire une solution qui décroît, tout comme la solution exacte.

Un autre point de vue : Une façon plus simple, mais un peu moins complète, d'étudier la différence entre ces deux méthodes est d'introduire la quantité $a(n) = \sum_{i=0}^{1/h} (-1)^i u_n^i$. Dans le cas de la première méthode, en sommant terme à terme toutes les équations **(1)** (multipliée respectivement par $(-1)^i$), on trouve :

$$\frac{a(n+1) - a(n)}{k} = \frac{-4a(n)}{\pi^2 h^2}$$

et donc
$$a(n+1) = a(n)\left(1 - \frac{4k}{\pi^2 h^2}\right)$$

On retrouve la condition $k < \dfrac{\pi^2 h^2}{2}$ pour que ce facteur multiplicatif soit inférieur à 1 en valeur absolue. Dans le cas contraire, on a une suite $(a(n))_{n\in\mathbb{N}}$ qui diverge, et donc la solution approchée ne peut pas rester très longtemps proche de la solution exacte, qui tends elle vers 0. Notons d'ailleurs que de grandes valeurs de la suite a indiquent de fortes oscillations de la solution calculée.

Dans le cas de la seconde méthode, en effectuant la même somme, on trouve :

$$\frac{a(n+1) - a(n)}{k} = \frac{-4a(n+1)}{\pi^2 h^2}$$

c'est-à-dire
$$a(n+1) = \frac{a(n)}{1 + \dfrac{4k}{\pi^2 h^2}}$$

Là, aucun problème. Quelle que soient les valeurs de h et k, la quantité $a(n)$ décroît, ce qui est compatible avec le fait que la solution tends vers 0, et que notre courbe reste à peu près "lisse", sans grandes oscillations.

Conclusion : Attention, ceci n'est pas une preuve du fait que la méthode implicite marche forcément, mais seulement du fait qu'elle est meilleure que la méthode explicite...Ces deux méthodes sont illustrées par des animations graphiques disponibles sur ce site (tout comme les deux programmes chaleurexpl et chaleurimpl mentionnés, en versions matlab et scilab), à l'addresse :

http ://www.dma.ens.fr/culturemath/maths/html/chaleur/chaleur.html

Annexe : matrices à diagonale dominante

Écrivons tout d'abord notre matrice $\mathrm{B} = \mathrm{I}_n + \dfrac{k}{\pi^2 h^2}\mathrm{A}$. Pour simplifier les notations, nous allons poser ici $c = \dfrac{k}{\pi^2 h^2}$ (c est donc un réel positif). Notre matrice s'écrit :

$$\begin{pmatrix} 1 & 0 & \dots & \dots & \dots & 0 \\ 0 & 1+2c & -c & & & \vdots \\ \vdots & -c & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & -c & \vdots \\ \vdots & & & -c & 1+2c & 0 \\ 0 & \dots & \dots & \dots & 0 & 1 \end{pmatrix}$$

Nous allons montrer que cette matrice B vérifie la propriété suivante : si X est un vecteur tel que BX a toutes ses coordonnées positives, alors il en est de même pour X.

Soit en effet X un tel vecteur, de coordonnées $x_1, \dots, x_n$, et i_0 l'indice vérifiant $x_{i_0} = \min_{1 \leq i \leq n} x_i$. (En cas d'égalité, on prend le plus petit i_0 vérifiant cette propriété). Supposons $x_{i_0} < 0$.

Les cas $i_0 = 1$ et $i_0 = n$ sont immédiatement exclus. En effet, la première (resp dernière) coordonnée de BX est x_1 (resp x_n), positif par hypothèse.

Supposons $3 \leq i_0 \leq n-2$. Alors l'hypothèse de positivité des coefficients de BX nous donne, pour le i_0-ième :

$$-cx_{i_0-1} + (1+2c)\,x_{i_0} - cx_{i_0+1} \geq 0$$

Or, par définition de i_0, on a $2cx_{i_0} - c\,(x_{i_0-1} + x_{i_0+1}) \leq 0$. D'autre part, on a supposé $x_{i_0} < 0$. On obtient par sommation :

$$-cx_{i_0-1} + (1+2c)\,x_{i_0} - cx_{i_0+1} < 0$$

ce qui est absurde. Les cas $i_0 = 2$ et $i_0 = n-1$ se traitent bien sûr de la même façon. Le coefficient x_{i_0}, minimal, est donc positif, et donc X est à coefficients positifs.

De cette propriété de la matrice B, on déduit que B est inversible et que B^{-1} est à coefficients positifs.

En effet, B est injective : si X_1 et X_2 sont deux vecteurs de même image par B, alors on a $\mathrm{B}(\mathrm{X}_1 - \mathrm{X}_2) = 0$. Donc $\mathrm{B}(\mathrm{X}_1 - \mathrm{X}_2)$ est un vecteur de coordonnées positives, donc $\mathrm{X}_1 - \mathrm{X}_2$ également.

Symétriquement, on obtient que les coordonnées du vecteur $\mathrm{X}_2 - \mathrm{X}_1$ sont aussi positives, pour les mêmes raisons. Et donc, nécessairement, $\mathrm{X}_1 = \mathrm{X}_2$, c'est-à-dire que la matrice B est injective.

Elle est donc aussi bijective, donc inversible. Son inverse B^{-1} transforme tout vecteur à coordonnées positives en un vecteur à coordonnées positives. C'est le cas en particulier des vecteurs de la base canonique, donc les vecteurs colonne de B^{-1} sont à coordonnées positives, c'est-à-dire que B^{-1} est à coefficients positifs.