

CHAPITRE 12

Introduction aux méthodes numériques.

Dans ce chapitre, nous présenterons des méthodes nous permettant d'obtenir des solutions explicites d'équations aux dérivées partielles. Nous allons illustrer ceci pour le problème de la chaleur tel que formulé au chapitre 7. Dans ce cas, nous pouvons calculer exactement la solution étant donnée la température initiale connue. Nous serons ainsi en mesure de comparer la solution explicite obtenue des méthodes numériques avec la solution exacte et d'analyser l'erreur due à l'approximation liée aux méthodes numériques.

Rappelons premièrement le théorème du développement de Taylor pour des fonctions à valeurs réelles d'une seule variable réelle. Soient $n \in \mathbf{N}$, $f(x)$ une fonction dérivable $(n + 1)$ fois dans un intervalle ouvert I contenant $x = x_0$ et $h \in \mathbf{R}$ tel que $x_0 + h \in I$, alors il existe $\alpha \in \mathbf{R}$ compris entre x_0 et $x_0 + h$ tel que

$$f(x_0 + h) = f(x_0) + \sum_{i=1}^n \frac{f^{(i)}(x_0)}{i!} h^i + R_n \quad \text{où} \quad R_n = \frac{f^{(n+1)}(\alpha)}{(n+1)!} h^{(n+1)}.$$

Par la suite, nous supposons toujours que $f(x)$ est suffisamment dérivable.

Ceci nous permet d'approximer $f(x_0 + h)$ lorsque $h \approx 0$. En effet, nous obtenons que

$$f(x_0 + h) \approx f(x_0) + \sum_{i=1}^n \frac{f^{(i)}(x_0)}{i!} h^i \quad (\text{éq. [1]})$$

et le terme d'erreur dans l'utilisation de cette approximation est

$$R_n = \frac{f^{(n+1)}(\alpha)}{(n+1)!} h^{(n+1)}.$$

De plus si $f^{(n+1)}(x)$ est continue dans un voisinage de x_0 , alors $f^{(n+1)}(\alpha) \approx f^{(n+1)}(x_0)$ et nous avons

$$R_n \approx \frac{f^{(n+1)}(x_0)}{(n+1)!} h^{(n+1)} \quad \Rightarrow \quad |R_n| \leq C h^{n+1} \quad \text{où } C \text{ est une constante.}$$

L'équation [1] nous permet de formuler plusieurs approximations de la dérivée première $f'(x_0)$. Ainsi en prenant $h > 0$, disons $h = \Delta x > 0$,

$$f(x_0 + \Delta x) \approx f(x_0) + f'(x_0) (\Delta x) \quad \Rightarrow \quad f'(x_0) \approx \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}$$

et nous dirons alors que

$$\frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}$$

est l'approximation de $f'(x_0)$ par différence finie progressive. Nous pouvons évaluer le terme d'erreur E_p dans l'utilisation de cette approximation. E_p est obtenu par

$$f(x_0 + \Delta x) = f(x_0) + f'(x_0) (\Delta x) + \frac{f^{(2)}(\alpha)}{2} (\Delta x)^2 \quad \Rightarrow \quad f'(x_0) = \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x} - \frac{f^{(2)}(\alpha)}{2} (\Delta x)$$

où α est compris entre x_0 et $x_0 + \Delta x$. Ainsi

$$E_p = -\frac{f^{(2)}(\alpha)}{2} (\Delta x) \quad \text{et} \quad |E_p| = \frac{|f^{(2)}(\alpha)|}{2} (\Delta x).$$

Par contre, si nous prenons $h < 0$, disons $h = -\Delta x$ avec $\Delta x > 0$, alors

$$f(x_0 - \Delta x) \approx f(x_0) - f'(x_0)(\Delta x) \quad \Rightarrow \quad f'(x_0) \approx \frac{f(x_0) - f(x_0 - \Delta x)}{\Delta x}$$

et nous dirons que

$$\frac{f(x_0) - f(x_0 - \Delta x)}{\Delta x}$$

est l'approximation de $f'(x_0)$ par différence finie régressive. Nous pouvons aussi évaluer le terme d'erreur E_r dans l'utilisation de cette approximation. E_r est obtenu par

$$f(x_0 - \Delta x) = f(x_0) - f'(x_0)(\Delta x) + \frac{f^{(2)}(\beta)}{2}(\Delta x)^2 \quad \Rightarrow \quad f'(x_0) = \frac{f(x_0) - f(x_0 - \Delta x)}{\Delta x} + \frac{f^{(2)}(\beta)}{2}(\Delta x)$$

où β est compris entre $x_0 - \Delta x$ et x_0 . Ainsi

$$E_r = \frac{f^{(2)}(\beta)}{2}(\Delta x) \quad \text{et} \quad |E_r| = \frac{|f^{(2)}(\beta)|}{2}(\Delta x).$$

Nous pouvons aussi faire la moyenne arithmétique de ces deux approximations de $f'(x_0)$ pour obtenir l'approximation de $f'(x_0)$ par différence finie centrée. Plus précisément, nous avons

$$\begin{aligned} f(x_0 + \Delta x) &\approx f(x_0) + f'(x_0)(\Delta x) + \frac{f^{(2)}(x_0)}{2}(\Delta x)^2 \quad \text{et} \\ f(x_0 - \Delta x) &\approx f(x_0) - f'(x_0)(\Delta x) + \frac{f^{(2)}(x_0)}{2}(\Delta x)^2. \end{aligned}$$

En soustrayant la deuxième expression de la première, nous obtenons

$$f(x_0 + \Delta x) - f(x_0 - \Delta x) \approx 2f'(x_0)(\Delta x) \quad \Rightarrow \quad f'(x_0) \approx \frac{f(x_0 + \Delta x) - f(x_0 - \Delta x)}{2(\Delta x)}$$

et nous disons que

$$\frac{f(x_0 + \Delta x) - f(x_0 - \Delta x)}{2(\Delta x)}$$

est l'approximation de $f'(x_0)$ par différence finie centrée. Nous pouvons aussi évaluer le terme d'erreur E_c dans l'utilisation de cette approximation. Nous avons que

$$\begin{aligned} f(x_0 + \Delta x) &= f(x_0) + f'(x_0)(\Delta x) + \frac{f^{(2)}(x_0)}{2}(\Delta x)^2 + \frac{f^{(3)}(\alpha')}{6}(\Delta x)^3 \quad \text{et} \\ f(x_0 - \Delta x) &= f(x_0) - f'(x_0)(\Delta x) + \frac{f^{(2)}(x_0)}{2}(\Delta x)^2 - \frac{f^{(3)}(\beta')}{6}(\Delta x)^3 \end{aligned}$$

où α' est compris entre x_0 et $x_0 + \Delta x$, alors que β' est compris entre $x_0 - \Delta x$ et x_0 . En soustrayant la deuxième équation de la première, nous obtenons

$$f(x_0 + \Delta x) - f(x_0 - \Delta x) = 2f'(x_0)(\Delta x) + \left[\frac{f^{(3)}(\alpha') + f^{(3)}(\beta')}{6} \right] (\Delta x)^3$$

et conséquemment

$$f'(x_0) = \frac{f(x_0 + \Delta x) - f(x_0 - \Delta x)}{2(\Delta x)} - \left[\frac{f^{(3)}(\alpha') + f^{(3)}(\beta')}{12} \right] (\Delta x)^2$$

Donc

$$E_c = - \left[\frac{f^{(3)}(\alpha') + f^{(3)}(\beta')}{12} \right] (\Delta x)^2 \quad \text{et} \quad |E_c| = \frac{|f^{(3)}(\alpha') + f^{(3)}(\beta')|}{12} (\Delta x)^2.$$

Si $f^{(3)}(x)$ est continue sur l'intervalle $[x_0 - \Delta x, x_0 + \Delta x]$, alors il existe $\gamma' \in [x_0 - \Delta x, x_0 + \Delta x]$ tel que

$$f^{(3)}(\gamma') = \frac{f^{(3)}(\alpha') + f^{(3)}(\beta')}{2}.$$

En effet, il suffit d'utiliser la connexité de l'intervalle $[x_0 - \Delta x, x_0 + \Delta x]$, du fait que l'image d'un ensemble connexe par une fonction continue est connexe et que

$$\frac{f^{(3)}(\alpha') + f^{(3)}(\beta')}{2}$$

est le milieu du segment d'extrémités: $f^{(3)}(\alpha')$ et $f^{(3)}(\beta')$. Conséquemment

$$E_c = - \frac{f^{(3)}(\gamma')}{6} (\Delta x)^2 \quad \text{et} \quad |E_c| = \frac{|f^{(3)}(\gamma')|}{6} (\Delta x)^2.$$

Ces trois approximations de $f'(x_0)$ sont consistantes, i.e. si $\Delta x \rightarrow 0$, alors chacune de ces approximations par différence finie approche $f'(x_0)$.

Nous pouvons illustrer ceci dans un exemple. Considérons la fonction $f(x) = \sqrt{1+x}$ autour de $x_0 = 0$. Comme $f'(x) = 1/2\sqrt{1+x}$, alors $f'(0) = 1/2$. Si nous utilisons chacune des approximations ci-dessus pour $\Delta x = 0.1$, nous obtenons pour l'approximation par différence finie progressive

$$f'(0) \approx \frac{f(\Delta x) - f(0)}{\Delta x} = \frac{\sqrt{1.1} - 1}{0.1} = 0.4881$$

et l'erreur relative est 2.38%; pour l'approximation par différence finie régressive

$$f'(0) \approx \frac{f(0) - f(-\Delta x)}{\Delta x} = \frac{1 - \sqrt{0.9}}{0.1} = 0.513167$$

et l'erreur relative est 2.6334% et finalement pour l'approximation par différence finie centrée

$$f'(0) \approx \frac{f(\Delta x) - f(-\Delta x)}{2(\Delta x)} = \frac{\sqrt{1.1} - \sqrt{0.9}}{2(0.1)} = 0.500628$$

et l'erreur relative est 0.1256%.

Ceci n'est pas trop surprenant parce que le terme d'erreur pour les approximations par différence finie progressive et régressive est de l'ordre de Δx , alors que celui de l'approximation par différence finie centrée est de l'ordre de $(\Delta x)^2$. Plus précisément, comme

$$f^{(2)}(x) = - \frac{1}{4\sqrt{(1+x)^3}} \quad \text{et} \quad f^{(3)}(x) = \frac{3}{8\sqrt{(1+x)^5}},$$

nous avons pour l'approximation par différence finie progressive

$$|E_p| = \left| \frac{-1}{4\sqrt{(1+\alpha)^3}} \right| \frac{\Delta x}{2} \quad \text{avec} \quad 0 \leq \alpha \leq \Delta x \quad \Rightarrow \quad |E_p| \leq \frac{\Delta x}{8}.$$

Dans ce cas, l'erreur relative est

$$\left| \frac{E_p}{f'(0)} \right| = 2|E_p| \quad \Rightarrow \quad \left| \frac{E_p}{f'(0)} \right| \leq \frac{\Delta x}{4}.$$

Cette dernière expression est inférieure à 2.5% lorsque $\Delta x = 0.1$. Pour l'approximation par différence finie régressive,

$$|E_r| = \left| \frac{-1}{4\sqrt{(1+\beta)^3}} \right| \frac{\Delta x}{2} \quad \text{avec } -\Delta x \leq \beta \leq 0 \quad \Rightarrow \quad |E_r| \leq \frac{\Delta x}{8\sqrt{(1-\Delta x)^3}}.$$

Cette dernière expression est égale à 0.014640174353 lorsque $\Delta x = 0.1$. Dans ce cas, l'erreur relative est

$$\left| \frac{E_r}{f'(0)} \right| = 2|E_r| \quad \Rightarrow \quad \left| \frac{E_r}{f'(0)} \right| \leq \frac{\Delta x}{4\sqrt{(1-\Delta x)^3}}.$$

Cette dernière expression est inférieure à 2.93% lorsque $\Delta x = 0.1$. Finalement pour l'approximation par différence finie centrée,

$$|E_c| = \left| \frac{3}{8\sqrt{(1+\gamma')^5}} \right| \frac{(\Delta x)^2}{6} \quad \text{avec } -\Delta x \leq \gamma' \leq \Delta x \quad \Rightarrow \quad |E_c| \leq \frac{(\Delta x)^2}{16\sqrt{(1-\Delta x)^5}}.$$

Cette dernière expression est égale à 0.00081334302 lorsque $\Delta x = 0.1$. Dans ce cas, l'erreur relative est

$$\left| \frac{E_c}{f'(0)} \right| = 2|E_c| \quad \Rightarrow \quad \left| \frac{E_c}{f'(0)} \right| \leq \frac{(\Delta x)^2}{8\sqrt{(1-\Delta x)^5}}.$$

Cette dernière expression est inférieure à 0.163% lorsque $\Delta x = 0.1$.

À partir des bornes supérieures obtenues pour le terme d'erreur, nous pourrions déterminer Δx pour obtenir une approximation de $f'(0)$ avec le degré de précision désiré.

Il est aussi possible d'approximer la dérivée seconde $f''(x_0)$. En effet, nous avons

$$\begin{aligned} f(x_0 + \Delta x) &= f(x_0) + f'(x_0)(\Delta x) + \frac{f^{(2)}(x_0)}{2}(\Delta x)^2 + \frac{f^{(3)}(x_0)}{6}(\Delta x)^3 + \frac{f^{(4)}(\alpha'')}{24}(\Delta x)^4 \quad \text{et} \\ f(x_0 - \Delta x) &= f(x_0) - f'(x_0)(\Delta x) + \frac{f^{(2)}(x_0)}{2}(\Delta x)^2 - \frac{f^{(3)}(x_0)}{6}(\Delta x)^3 + \frac{f^{(4)}(\beta'')}{24}(\Delta x)^4 \end{aligned}$$

où $\alpha'' \in [x_0, x_0 + \Delta x]$ et $\beta'' \in [x_0 - \Delta x, x_0]$. En additionnant ces deux équations et en laissant tomber les dérivées quatrièmes, nous obtenons

$$f(x_0 + \Delta x) + f(x_0 - \Delta x) \approx 2f(x_0) + f^{(2)}(x_0)(\Delta x)^2 \quad \Rightarrow \quad f^{(2)}(x_0) \approx \frac{f(x_0 + \Delta x) - 2f(x_0) + f(x_0 - \Delta x)}{(\Delta x)^2}.$$

Nous dirons que

$$\frac{f(x_0 + \Delta x) - 2f(x_0) + f(x_0 - \Delta x)}{(\Delta x)^2}$$

est l'approximation par différence finie centrée de la dérivée seconde $f''(x_0)$. Il est aussi possible de déterminer le terme d'erreur $E_c^{(2)}$ dans l'utilisation de cette approximation. Nous avons en additionnant les deux équations ci-dessus

$$f(x_0 + \Delta x) + f(x_0 - \Delta x) = 2f(x_0) + f^{(2)}(x_0)(\Delta x)^2 + \left[\frac{f^{(4)}(\alpha'') + f^{(4)}(\beta'')}{24} \right] (\Delta x)^4$$

et conséquemment

$$f^{(2)}(x_0) = \frac{f(x_0 + \Delta x) - 2f(x_0) + f(x_0 - \Delta x)}{(\Delta x)^2} - \left[\frac{f^{(4)}(\alpha'') + f^{(4)}(\beta'')}{24} \right] (\Delta x)^2.$$

Ainsi le terme d'erreur est

$$E_c^{(2)} = - \left[\frac{f^{(4)}(\alpha'') + f^{(4)}(\beta'')}{24} \right] (\Delta x)^2 \quad \text{et} \quad |E_c^{(2)}| = \frac{|f^{(4)}(\alpha'') + f^{(4)}(\beta'')|}{24} (\Delta x)^2.$$

Si $f^{(4)}(x)$ est continue sur l'intervalle $[x_0 - \Delta x, x_0 + \Delta x]$, alors il existe $\lambda'' \in [x_0 - \Delta x, x_0 + \Delta x]$ tel que

$$f^{(4)}(\lambda'') = \frac{f^{(4)}(\alpha'') + f^{(4)}(\beta'')}{2}$$

et le terme d'erreur est

$$E_c^{(2)} = - \left[\frac{f^{(4)}(\lambda'')}{12} \right] (\Delta x)^2 \quad \text{et} \quad |E_c^{(2)}| = \frac{|f^{(4)}(\lambda'')|}{12} (\Delta x)^2.$$

Jusqu'à présent, nous avons décrit des approximations des dérivées première et seconde pour une fonction d'une seule variable. Mais bien entendu, nous pouvons aussi utiliser ces approximations pour les dérivées partielles de fonctions de plusieurs variables. Ceci est possible à cause de la définition de ces dérivées dans laquelle toutes les variables sont constantes sauf celle relativement à laquelle nous dérivons. Par exemple, si $u = u(x, y)$, alors

$$\frac{\partial u}{\partial x}(x_0, y_0) \approx \frac{u(x_0 + \Delta x, y_0) - u(x_0 - \Delta x, y_0)}{2(\Delta x)} \quad \text{et} \quad \frac{\partial u}{\partial y}(x_0, y_0) \approx \frac{u(x_0, y_0 + \Delta y) - u(x_0, y_0 - \Delta y)}{2(\Delta y)}$$

où $\Delta x > 0$ et $\Delta y > 0$, lorsque nous utilisons les approximations par différence finie centrée.

Mais nous aurions tout aussi bien pu utiliser les autres approximations. Ce choix dépendra du problème à étudier. Dans tous les cas, nous pouvons déterminer le terme d'erreur. Nous allons maintenant illustrer comment ces approximations peuvent être utilisées pour le problème de la chaleur étudié au chapitre 7, i.e.

$$\left\{ \begin{array}{l} \frac{\partial u}{\partial t} = c^2 \frac{\partial^2 u}{\partial x^2} \quad \text{où } u = u(x, t), \quad 0 \leq x \leq \ell, \quad t \geq 0 \quad \text{avec} \\ u(0, t) = 0 \quad \text{pour tout } t \geq 0 \\ u(\ell, t) = 0 \quad \text{pour tout } t \geq 0 \quad \text{et} \\ u(x, 0) = f(x) \quad \text{pour tout } x \in [0, \ell] \end{array} \right.$$

Fixons $\Delta x > 0$ et $\Delta t > 0$. Alors par ce que nous avons vu précédemment

$$\frac{\partial u}{\partial t}(x, t) = \frac{u(x, t + \Delta t) - u(x, t)}{\Delta t} - \frac{\Delta t}{2} \frac{\partial^2 u}{\partial t^2}(x, \alpha) \quad \text{avec } t \leq \alpha \leq t + \Delta t$$

si nous voulons utiliser l'approximation par différence finie progressive pour

$$\frac{\partial u}{\partial t}(x, t)$$

et

$$\frac{\partial^2 u}{\partial x^2}(x, t) = \frac{u(x + \Delta x, t) - 2u(x, t) + u(x - \Delta x, t)}{(\Delta x)^2} - \frac{(\Delta x)^2}{12} \frac{\partial^4 u}{\partial x^4}(\beta, t) \quad \text{avec } x - \Delta x \leq \beta \leq x + \Delta x$$

si nous voulons utiliser l'approximation par différence finie centrée pour

$$\frac{\partial^2 u}{\partial x^2}(x, t).$$

Donc l'équation de la chaleur devient

$$\frac{u(x, t + \Delta t) - u(x, t)}{\Delta t} = c^2 \left(\frac{u(x + \Delta x, t) - 2u(x, t) + u(x - \Delta x, t)}{(\Delta x)^2} \right) + E$$

où le terme d'erreur de discrétisation E est

$$E = \frac{(\Delta t)}{2} \left[\frac{\partial^2 u}{\partial t^2}(x, \alpha) \right] - \frac{c^2(\Delta x)^2}{12} \left[\frac{\partial^4 u}{\partial x^4}(\beta, t) \right].$$

Nous avons supposé ci-dessus que les dérivées partielles sont continues. Ainsi nous avons comme approximation

$$\frac{u(x, t + \Delta t) - u(x, t)}{\Delta t} \approx c^2 \left(\frac{u(x + \Delta x, t) - 2u(x, t) + u(x - \Delta x, t)}{(\Delta x)^2} \right).$$

Nous cherchons à déterminer une approximation $\tilde{u}(x, t)$ de la solution exacte $u(x, t)$. Bien entendu, nous voulons que cette approximation $\tilde{u}(x, t)$ soit le plus près de la solution exacte $u(x, t)$. Il est alors naturel d'exiger que cette approximation $\tilde{u}(x, t)$ satisfasse l'équation

$$\frac{\tilde{u}(x, t + \Delta t) - \tilde{u}(x, t)}{\Delta t} = c^2 \left(\frac{\tilde{u}(x + \Delta x, t) - 2\tilde{u}(x, t) + \tilde{u}(x - \Delta x, t)}{(\Delta x)^2} \right). \quad (\text{éq. [2]})$$

Plutôt que de déterminer $\tilde{u}(x, t)$ pour tout $x \in [0, \ell]$ et $t \geq 0$, nous allons résoudre l'équation [2] pour les points d'un maillage de notre domaine.

Subdivisons l'intervalle $[0, \ell]$ en N sous-intervalles égaux de longueur $\Delta x = \ell/N$. Les extrémités de ces sous-intervalles sont $x_0 = 0, x_1 = (\Delta x), x_2 = 2(\Delta x), \dots, x_i = i(\Delta x), \dots, x_N = N(\Delta x) = \ell$. De façon similaire, nous pouvons subdiviser l'intervalle $[0, \infty[$ en sous-intervalles égaux de longueur Δt . Dans ce cas, le nombre de sous-intervalles est infini et les extrémités de ces sous-intervalles sont $t_0 = 0, t_1 = (\Delta t), t_2 = 2(\Delta t), \dots, t_j = j(\Delta t), \dots$. Ainsi nous obtenons un maillage de notre domaine dont les points sont (x_i, t_j) pour $0 \leq i \leq N$ et $j \geq 0$.

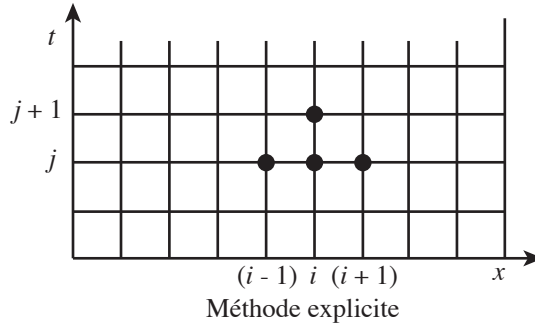
Posons $\tilde{u}(x_i, t_j) = U_i^{(j)}$ où $0 \leq i \leq N$ et $j \geq 0$. Dans ce cas, l'équation [2] devient pour les points (x_i, t_j) du maillage

$$\frac{U_i^{(j+1)} - U_i^{(j)}}{\Delta t} = c^2 \left(\frac{U_{i+1}^{(j)} - 2U_i^{(j)} + U_{i-1}^{(j)}}{(\Delta x)^2} \right) \quad \text{pour } 1 \leq i \leq (N-1), j \geq 0$$

Cette dernière équation nous permet de calculer $U_i^{(j+1)}$ en fonction de $U_k^{(j)}$ pour $k = (i-1), i, (i+1)$. En effet,

$$U_i^{(j+1)} = U_i^{(j)} + \left[\frac{c^2(\Delta t)}{(\Delta x)^2} \right] (U_{i+1}^{(j)} - 2U_i^{(j)} + U_{i-1}^{(j)}). \quad (\text{éq. [3]})$$

Nous avons illustré ci-dessous les points du maillage qui interviennent dans cette équation [3].



De plus il nous est possible de préciser les valeurs de $U_i^{(j)}$ lorsque $j = 0$ ou encore lorsque $i = 0$ et $i = N$ pour tenir compte des conditions au bord et de la condition initiale. Parce que $u(0, t) = 0$ pour tout $t \geq 0$, alors nous posons comme condition $U_0^{(j)} = 0$ pour tout $j \geq 0$. Parce que $u(\ell, t) = 0$ pour tout $t \geq 0$, alors nous posons comme condition $U_N^{(j)} = 0$ pour tout $j \geq 0$. Finalement $u(x, 0) = f(x)$ pour tout $x \in [0, \ell]$, alors nous posons comme condition $U_i^{(0)} = f(x_i)$ pour tout $i = 0, 1, 2, \dots, N$.

Si nous tenons compte de ces valeurs pour les points au bord du maillage et de l'équation [3], nous pouvons complètement déterminer les valeurs $U_i^{(j)}$ et cette solution est unique. Nous avons ainsi une méthode explicite. Ce qualificatif "explicite" sera mieux compris lorsque nous décrirons plus tard une méthode implicite, par exemple celle de Crank-Nicolson.

Nous allons illustrer ceci dans l'exemple suivant. Considérons le problème:

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \quad \text{où } u = u(x, t), \quad 0 \leq x \leq 1, \quad t \geq 0 \quad \text{avec}$$

$$u(0, t) = 0 \quad \text{et} \quad u(1, t) = 0 \quad \text{pour tout } t \geq 0 \quad \text{et}$$

$$u(x, 0) = f(x) = \begin{cases} 0, & \text{si } 0 \leq x \leq 0.25; \\ (x - 0.25), & \text{si } 0.25 \leq x \leq 0.5; \\ (0.75 - x), & \text{si } 0.5 \leq x \leq 0.75; \\ 0, & \text{si } 0.75 \leq x \leq 1. \end{cases} \quad \text{pour tout } x \in [0, \ell].$$

Ici $\ell = 1$ et $c = 1$. Il est possible de déterminer la solution exacte de ce problème avec les résultats obtenus au chapitre 7. Nous obtenons ainsi

$$u(x, t) = \sum_{n=1}^{\infty} \frac{2}{n^2 \pi^2} \left[-\sin\left(\frac{n\pi}{4}\right) + 2\sin\left(\frac{n\pi}{2}\right) - \sin\left(\frac{3n\pi}{4}\right) \right] \sin(n\pi x) e^{-n^2 \pi^2 t}.$$

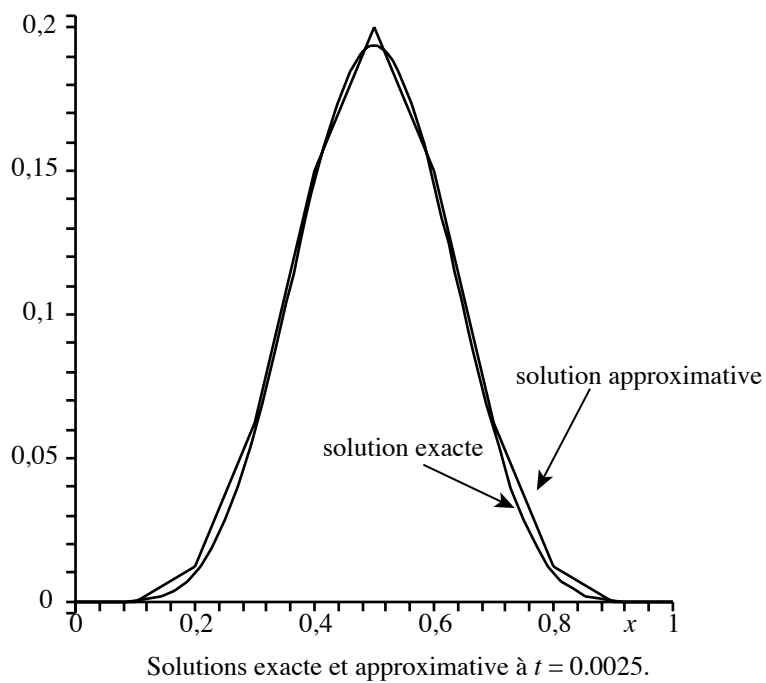
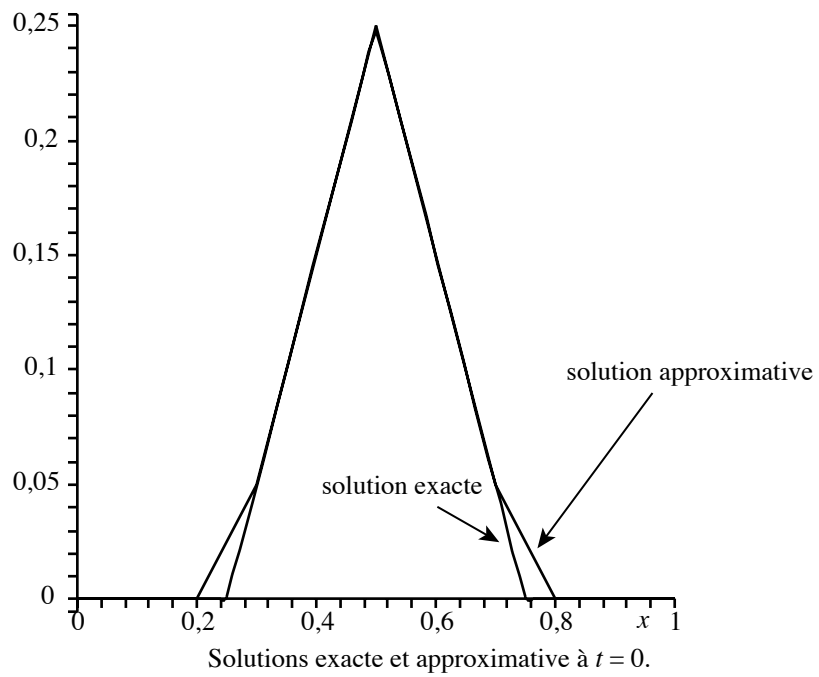
Pour le calcul des valeurs $U_i^{(j)}$, il nous faut fixer Δx et Δt . Prenons $\Delta x = 0.1$ et $\Delta t = 0.0025$, alors l'équation [3] devient dans ce cas particulier

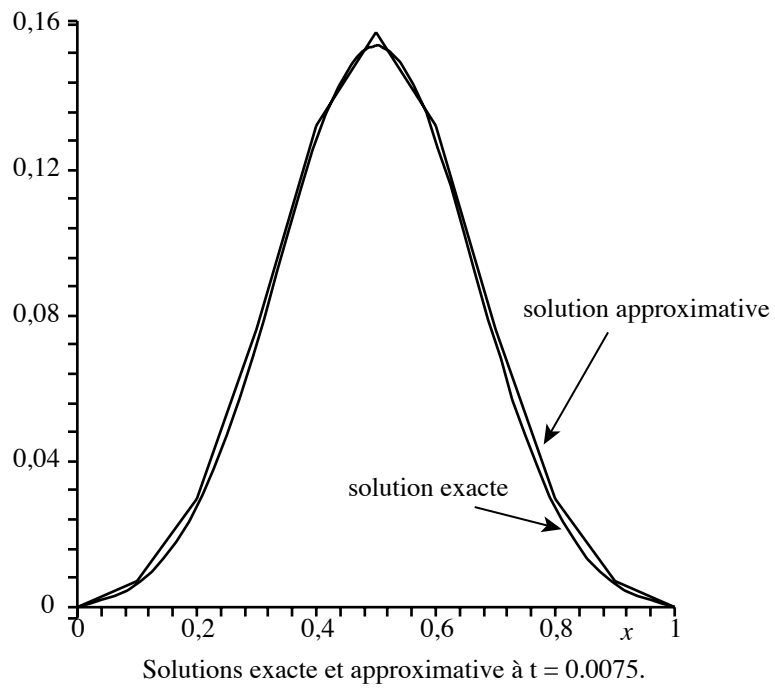
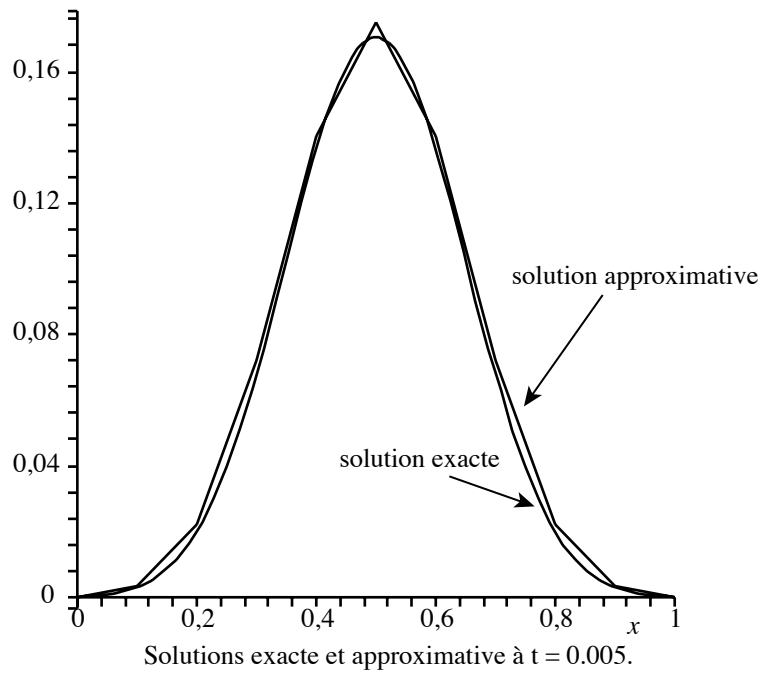
$$U_i^{(j+1)} = U_i^{(j)} + \left[\frac{1^2 (0.0025)}{(0.1)^2} \right] (U_{i+1}^{(j)} - 2U_i^{(j)} + U_{i-1}^{(j)}) = 0.25U_{i+1}^{(j)} + 0.5U_i^{(j)} + 0.25U_{i-1}^{(j)}.$$

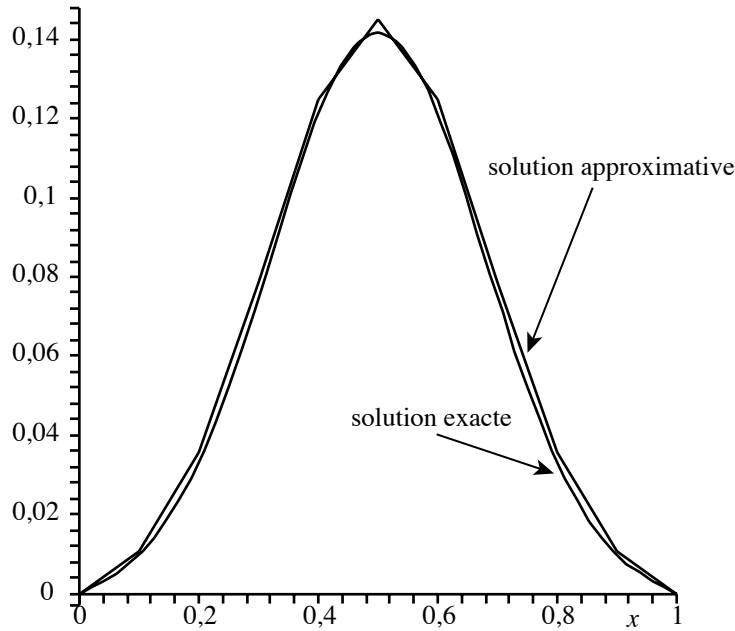
À cette équation, il nous faut aussi ajouter les conditions $U_0^{(j)} = U_{10}^{(j)} = 0$ pour $j \geq 0$ et $U_i^{(0)} = f(i/10)$ pour $i = 0, 1, 2, \dots, 10$. Nous obtenons ainsi le tableau suivant des valeurs des $U_i^{(j)}$.

i	$U_i^{(0)}$	$U_i^{(1)}$	$U_i^{(2)}$	$U_i^{(3)}$	$U_i^{(4)}$
0	0	0	0	0	0
1	0	0	0.003125	0.00703125	0.0109375
2	0	0.0125	0.021875	0.0296875	0.003574219
3	0.05	0.0625	0.071875	0.0765625	0.07871094
4	0.15	0.15	0.140625	0.13203125	0.12460938
5	0.25	0.2	0.175	0.1578125	0.14492188
6	0.15	0.15	0.140625	0.13203125	0.12460938
7	0.05	0.0625	0.071875	0.0765625	0.07871094
8	0	0.0125	0.021875	0.0296875	0.003574219
9	0	0	0.003125	0.00703125	0.0109375
10	0	0	0	0	0

En traçant les graphes de la solution exacte pour les valeurs de t suivantes: $t = 0$, $t = \Delta t = 0.0025$, $t = 2(\Delta t) = 0.005$, $t = 3(\Delta t) = 0.0075$ et $t = 4(\Delta t) = 0.01$ et en comparant avec les valeurs des $U_i^{(j)}$ ci-dessus, nous vérifions que ces dernières sont de bonnes approximations.







Solutions exacte et approximative à $t = 0.01$.

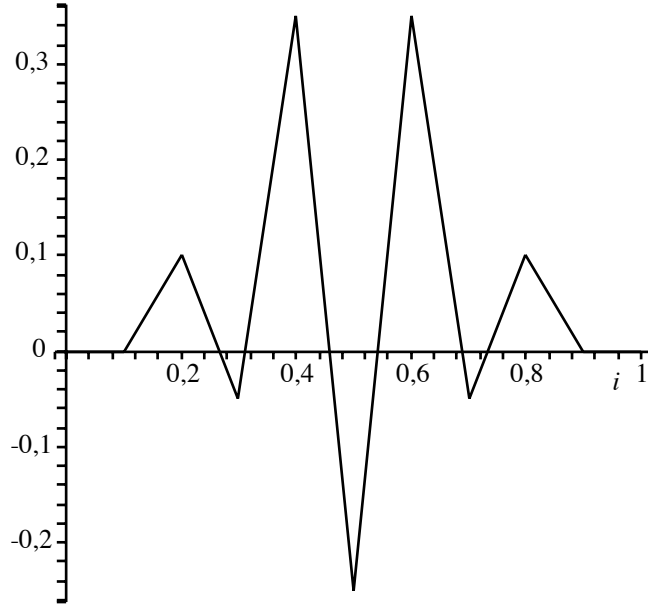
Nous pourrions penser que si Δx et Δt sont presque nuls, l'approximation devrait être bonne. Cependant ceci n'est pas suffisant. Il faut aussi tenir compte d'un problème d'instabilité dans ces situations. Nous allons maintenant illustrer ce phénomène. Considérons toujours le même problème. Mais si, pour le calcul des valeurs $U_i^{(j)}$, nous avons fixé $\Delta x = 0.1$ et $\Delta t = 0.01$, alors l'équation [3] devient dans ce cas particulier

$$U_i^{(j+1)} = U_i^{(j)} + \left[\frac{1^2 (0.01)}{(0.1)^2} \right] (U_{i+1}^{(j)} - 2U_i^{(j)} + U_{i-1}^{(j)}) = U_{i+1}^{(j)} - U_i^{(j)} + U_{i-1}^{(j)}.$$

Nous obtenons ainsi le tableau suivant des valeurs des $U_i^{(j)}$.

i	$U_i^{(0)}$	$U_i^{(1)}$	$U_i^{(2)}$	$U_i^{(3)}$	$U_i^{(4)}$
0	0	0	0	0	0
1	0	0	0.05	0	0.1
2	0	0.05	0.05	0.1	-0.15
3	0.05	0.1	0.1	-0.05	0.5
4	0.15	0.15	0	0.35	-0.65
5	0.25	0.05	0.25	-0.25	0.95
6	0.15	0.15	0	0.35	-0.65
7	0.05	0.1	0.1	-0.05	0.5
8	0	0.05	0.05	0.1	-0.15
9	0	0	0.05	0	0.1
10	0	0	0	0	0

Ces valeurs ne peuvent correspondre à notre problème. Nous devrions voir une décroissance exponentielle et ce n'est pas du tout ce qui se passe pour ces valeurs. Pour illustrer ceci, nous avons tracé le graphe de $U_i^{(3)}$ ci-dessous.



Graphique de $U_i^{(3)}$

Il y a là un phénomène d'instabilité, qui est associé à la valeur prise par $\sigma = c^2(\Delta t)/(\Delta x)^2$. Nous allons maintenant expliquer les raisons de cette instabilité et sa relation avec σ .

Pour bien analyser la solution approximative $U_i^{(j)}$, il faut revenir à l'équation[3] et d'obtenir la solution générale du problème intermédiaire (\clubsuit) suivant:

$$(\clubsuit) \quad \begin{cases} (U_i^{(j+1)} - U_i^{(j)}) = U_i^{(j)} + \sigma(U_{i+1}^{(j)} - 2U_i^{(j)} + U_{i-1}^{(j)}) \\ \text{avec les conditions } U_0^{(j)} = 0 \text{ et } U_N^{(j)} = 0 \text{ pour tout } j \geq 0 \end{cases}$$

Nous avons laissé de côté la condition initiale $U_i^{(0)} = f(i\Delta x)$. Nous ajouterons celle-ci plus tard à notre analyse.

L'équation de différences finies ci-dessus est linéaire, i.e. les combinaisons linéaires de deux solutions de l'équation sont aussi des solutions. Pour résoudre le problème (\clubsuit) ci-dessus, il est possible d'utiliser une variante de la méthode de séparation de variables pour obtenir des solutions et ensuite de considérer les combinaisons linéaires de ces solutions. Nous allons maintenant développer ceci.

Considérons des solutions non triviales du problème (\clubsuit) ci-dessus de la forme $U_i^{(j)} = F(i)G(j)$, où $F : \{0, 1, 2, \dots, N\} \rightarrow \mathbf{R}$ et $G : \mathbf{N} \rightarrow \mathbf{R}$. Après avoir substitué ceci dans l'équation, nous obtenons

$$F(i) (G(j+1) - G(j)) = \sigma(F(i+1) - 2F(i) + F(i-1)) G(j) \quad \text{pour tout } 1 \leq i \leq (N-1), j \geq 0.$$

Si nous divisons les deux côtés de cette dernière équation par $\sigma F(i)G(j)$, nous obtenons

$$\frac{G(j+1) - G(j)}{\sigma G(j)} = \frac{F(i+1) - 2F(i) + F(i-1)}{F(i)} \quad \text{pour tout } 1 \leq i \leq (N-1), j \geq 0.$$

Le terme de gauche est une fonction de j , alors que celui de droite est une fonction de i . Pour que cette dernière équation soit possible, il faut que chacun de ces termes soit constant. Donc

$$\frac{G(j+1) - G(j)}{\sigma G(j)} = \frac{F(i+1) - 2F(i) + F(i-1)}{F(i)} = \lambda \quad \text{pour tout } 1 \leq i \leq (N-1), j \geq 0$$

pour un λ et nous avons ainsi deux équations à différences finies

$$\left\{ \begin{array}{l} G(j+1) - (1 + \sigma\lambda)G(j) = 0 \quad \text{pour tout } j \geq 0 \\ F(i+1) - (2 + \lambda)F(i) + F(i-1) = 0 \quad \text{pour tout } 1 \leq i \leq (N-1) \end{array} \right.$$

Si nous considérons les conditions au bord $U_0^{(j)} = 0$ et $U_N^{(j)} = 0$ pour tout $j \geq 0$, alors nous pouvons déduire de ces conditions que $F(0) = 0$ et $F(N) = 0$. Nous avons donc à résoudre le système suivant:

$$\left\{ \begin{array}{l} G(j+1) - (1 + \sigma\lambda)G(j) = 0 \quad \text{pour tout } j \geq 0 \\ F(i+1) - (2 + \lambda)F(i) + F(i-1) = 0 \quad \text{pour tout } 1 \leq i \leq (N-1) \\ \text{avec } F(0) = 0 \text{ et } F(N) = 0. \end{array} \right.$$

Nous devons maintenant rappeler que la solution générale d'une relation de récurrence d'ordre 2 de la forme $H(i+2) + aH(i+1) + bH(i) = 0$ pour tout $i \geq 0$, où $a, b \in \mathbf{R}$, est obtenue en considérant les racines du polynôme $x^2 + ax + b$. Il y a trois cas à considérer:

1) Si $x^2 + ax + b$ a deux racines réelles distinctes: r_1, r_2 , alors la solution générale de $H(i+2) + aH(i+1) + bH(i) = 0$ est $H(i) = Ar_1^i + Br_2^i$, où A et B sont des constantes.

2) Si $x^2 + ax + b$ a une racine réelle double: r , alors la solution générale de $H(i+2) + aH(i+1) + bH(i) = 0$ est $H(i) = Ar^i + Bir^i$, où A et B sont des constantes.

3) Si $x^2 + ax + b$ a deux racines complexes non réelles: r_1, r_2 , alors celles-ci sont conjuguées et de la forme $r_1 = \rho \exp(\sqrt{-1}\theta)$, $r_2 = \rho \exp(-\sqrt{-1}\theta)$, où $\rho \in \mathbf{R}, \rho > 0$ et $\theta \in \mathbf{R}$ et, dans ce cas, la solution générale de $H(i+2) + aH(i+1) + bH(i) = 0$ est $H(i) = A\rho^i \cos(i\theta) + B\rho^i \sin(i\theta)$, où A et B sont des constantes.

Si maintenant nous revenons au système à différences finies ci-dessus, il faut alors considérer le polynôme $x^2 - (2 + \lambda)x + 1$ pour résoudre l'équation $F(i+1) - (2 + \lambda)F(i) + F(i-1) = 0$. Le polynôme $x^2 - (2 + \lambda)x + 1$ a comme racines:

$$r_1 = \frac{(2 + \lambda) + \sqrt{\lambda^2 + 4\lambda}}{2} \quad \text{et} \quad r_2 = \frac{(2 + \lambda) - \sqrt{\lambda^2 + 4\lambda}}{2}.$$

Si $\lambda < -4$ ou $\lambda > 0$, alors r_1 et r_2 sont deux racines réelles distinctes et la solution générale de $F(i+1) - (2 + \lambda)F(i) + F(i-1) = 0$ est $F(i) = Ar_1^i + Br_2^i$. Mais si nous tenons compte maintenant des conditions au bord, nous obtenons

$$\left\{ \begin{array}{l} F(0) = A + B = 0 \\ F(N) = Ar_1^N + Br_2^N = 0 \end{array} \right\} \Rightarrow A = B = 0.$$

En effet, nous avons que $r_1 \neq r_2$ et $r_1 \neq -r_2$. Dans ce dernier cas, c'est parce que $r_1 + r_2 = (2 + \lambda)$ et $\lambda \neq -2$. Maintenant le déterminant

$$\begin{vmatrix} 1 & 1 \\ r_1^N & r_2^N \end{vmatrix} = r_2^N - r_1^N = r_1^N \left[\left(\frac{r_2}{r_1} \right)^N - 1 \right] \neq 0$$

car $r_1 \neq 0$ et $(r_2/r_1)^N = 1 \Rightarrow r_2 = \pm r_1$. Mais ceci est absurde. Nous devons ainsi exclure le cas où $\lambda < -4$ ou $\lambda > 0$.

Si $\lambda = -4$ ou $\lambda = 0$, alors $r_1 = r_2 = r$ est une racine réelle double et la solution générale de $F(i+1) - (2 + \lambda)F(i) + F(i-1) = 0$ est $F(i) = Ar^i + Bir^i$. Mais si nous tenons compte maintenant des conditions au bord, nous obtenons

$$\left\{ \begin{array}{l} F(0) = A = 0 \\ F(N) = Ar^N + BNr^N = 0 \end{array} \right\} \Rightarrow A = 0 \text{ et } Br^N = 0 \Rightarrow A = B = 0.$$

En effet, $r = -1$ ou 1 et $r \neq 0$.

Il nous reste à considérer le cas où $-4 < \lambda < 0$. Alors r_1 et r_2 sont deux racines complexes distinctes non réelles conjuguées. Dans ce cas, ces racines sont de la forme $r_1 = \rho \exp(\sqrt{-1} \theta)$ et $r_2 = \rho \exp(-\sqrt{-1} \theta)$. Nous pouvons être plus précis pour ρ . En effet,

$$\rho = \sqrt{\frac{(2 + \lambda)^2 + (-\lambda^2 - 4\lambda)}{4}} = 1$$

Parce que $\rho = 1$, nous obtenons que la solution générale de $F(i + 1) - (2 + \lambda)F(i) + F(i - 1) = 0$ est $F(i) = A \cos(i\theta) + B \sin(i\theta)$. Mais si nous tenons compte des conditions au bord, nous obtenons

$$\left\{ \begin{array}{l} F(0) = A = 0 \\ F(N) = A \cos(N\theta) + B \sin(N\theta) = 0 \end{array} \right\} \Rightarrow \left\{ \begin{array}{l} A = 0; \\ B \sin(N\theta) = 0. \end{array} \right\}$$

Comme nous cherchons à déterminer une solution non triviale et que $A = 0$, nous pouvons supposer que $B \neq 0$. Ainsi nous obtenons que $\sin(N\theta) = 0$. Conséquemment $\theta = (n\pi/N)$ où $n \in \mathbf{N}, n \geq 1$ et nous obtenons que $F(i) = B \sin(n\pi i/N)$. De plus nous pouvons calculer $\lambda = \lambda_n$ correspondant à cet angle θ . En effet,

$$\frac{(2 + \lambda) + \sqrt{\lambda^2 + 4\lambda}}{2} = \exp\left(\frac{n\pi\sqrt{-1}}{N}\right) \Rightarrow \frac{(2 + \lambda)}{2} = \cos\left(\frac{n\pi}{N}\right) \Rightarrow \lambda_n = 2 \left[\cos\left(\frac{n\pi}{N}\right) - 1 \right].$$

Comme $-4 < \lambda = \lambda_n < 0$, nous pouvons alors ajouter la condition: $n \neq 0 \pmod{N}$

Avant de considérer l'équation $G(j + 1) - (1 + \sigma\lambda_n)G(j) = 0$, rappelons que la solution générale d'une relation de récurrence d'ordre 1 de la forme $H(i + 1) = aH(i) + b$ est

$$H(i) = \begin{cases} A' a^i + b \left(\frac{a^i - 1}{a - 1} \right), & \text{si } a \neq 1; \\ A' + bi, & \text{si } a = 1. \end{cases}$$

Nous pouvons maintenant utiliser cette dernière remarque. Notons premièrement que $(1 + \sigma\lambda_n) \neq 1$. Sinon $\sigma\lambda_n = 0$ et $\lambda_n < 0 \Rightarrow \sigma = 0$. Ceci est absurde, car $\Delta t, \Delta x$ et c sont > 0 . Nous pouvons donc écrire que $G(j) = A' (1 + \sigma\lambda_n)^j$.

Notons $(1 + \sigma\lambda_n)$ par κ_n . Il est possible de calculer κ_n :

$$\kappa_n = (1 + \sigma\lambda_n) = \left(1 - 2\sigma \left(1 - \cos\left(\frac{n\pi}{N}\right) \right) \right)$$

Pour chaque $n \in \mathbf{N}, n \geq 1$ et $n \neq 0 \pmod{N}$, nous obtenons donc des solutions du problème (\clubsuit):

$$U_i^{(j)} = a_n \sin\left(\frac{n\pi i}{N}\right) (\kappa_n)^j \quad \text{où } a_n \text{ est une constante.}$$

Il est possible de restreindre les valeurs de n pour obtenir une base des solutions. Plus précisément, nous pouvons nous restreindre aux valeurs de n entre 1 et $N - 1$.

Finalement nous obtenons que la solution générale du problème (\clubsuit) est

$$U_i^{(j)} = \sum_{n=1}^{N-1} a_n \sin\left(\frac{n\pi i}{N}\right) (\kappa_n)^j$$

Si nous ajoutons la condition initiale $U_i^{(0)} = f(x_i)$, alors nous pouvons déterminer les coefficients a_n en considérant le système de $(N - 1)$ équations linéaires à $(N - 1)$ inconnues:

$$\sum_{n=1}^{N-1} a_n \sin\left(\frac{n\pi i}{N}\right) = f(x_i) \quad \text{pour tout } i = 1, 2, \dots, (N - 1).$$

Lemme 1 Le déterminant D de la matrice $n \times n$ suivante:

$$A = \begin{pmatrix} \sin(x) & \sin(2x) & \cdots & \sin(qx) & \cdots & \sin(nx) \\ \sin(2x) & \sin(4x) & \cdots & \sin(2qx) & \cdots & \sin(2nx) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \sin(px) & \sin(2px) & \cdots & \sin(pqx) & \cdots & \sin(pnx) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \sin(nx) & \sin(2nx) & \cdots & \sin(nqx) & \cdots & \sin(n^2x) \end{pmatrix}$$

est

$$D = \det(A) = 2^{n(n-1)/2} \left[\prod_{i=1}^n \sin(ix) \right] \left[\prod_{1 \leq j < k \leq n} (\cos(kx) - \cos(jx)) \right].$$

En particulier, ce déterminant ne s'annule pas si $x = \pi/N$ et $n = N - 1$.

Esquisse de preuve: Nous n'allons qu'esquisser cette preuve. Il faut premièrement noter qu'en utilisant la formule de de Moivre:

$$\left(\cos(\theta) + \sin(\theta)\sqrt{-1} \right)^k = \left(\cos(k\theta) + \sin(k\theta)\sqrt{-1} \right),$$

nous pouvons montrer que $\sin(k\theta) = \sin(\theta)P_k(\cos(\theta))$, où P_k est un polynôme de degré $(k - 1)$ dont le coefficient de la plus grande puissance, celle de degré $(k - 1)$, est 2^{k-1} . Par exemple,

$$\sin(2\theta) = \sin(\theta) \left[2 \cos(\theta) \right], \quad \sin(3\theta) = \sin(\theta) \left[4 \cos^2(\theta) - 1 \right] \quad \text{et} \quad \sin(4\theta) = \sin(\theta) \left[8 \cos^3(\theta) - 4 \cos(\theta) \right].$$

Si nous considérons la p^{e} ligne, nous obtenons que

$$\left(\sin(px), \sin(2px), \dots, \sin(pqx), \dots, \sin(pnx) \right) = \sin(px) \left(1, P_2(\cos(px)), \dots, P_q(\cos(px)), \dots, P_n(\cos(px)) \right).$$

En factorisant $\sin(px)$ de la p^{e} ligne et en faisant des opérations colonnes, nous obtenons que le déterminant recherché D est $\prod_{1 \leq i \leq n} \sin(ix)$ fois le déterminant de la matrice

$$\begin{pmatrix} 1 & 2 \cos(x) & \cdots & 2^{q-1} \cos^{q-1}(x) & \cdots & 2^{n-1} \cos^{n-1}(x) \\ 1 & 2 \cos(2x) & \cdots & 2^{q-1} \cos^{q-1}(2x) & \cdots & 2^{n-1} \cos^{n-1}(2x) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 2 \cos(px) & \cdots & 2^{q-1} \cos^{q-1}(px) & \cdots & 2^{n-1} \cos^{n-1}(px) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 2 \cos(nx) & \cdots & 2^{q-1} \cos^{q-1}(nx) & \cdots & 2^{n-1} \cos^{n-1}(nx) \end{pmatrix}$$

Mais il est facile d'obtenir le déterminant de cette dernière matrice. Il s'agit de

$$2^{1+2+\cdots+(n-1)} = 2^{n(n-1)/2} \times \text{déterminant de Vandermonde}$$

évalué aux valeurs $x_1 = \cos(x)$, $x_2 = \cos(2x)$, \dots , $x_n = \cos(nx)$. Ce déterminant est bien connu et nous obtenons finalement que

$$D = 2^{n(n-1)/2} \left[\prod_{i=1}^n \sin(ix) \right] \left[\prod_{1 \leq j < k \leq n} \left(\cos(kx) - \cos(jx) \right) \right].$$

Pour terminer la preuve du lemme, il suffit de noter que si $x = (\pi/N)$ et $n = N - 1$, chacun des termes $\sin\left(\frac{i\pi}{N}\right) \neq 0$ si $i = 1, 2, \dots, (N - 1)$ et $\left(\cos\left(\frac{k\pi}{N}\right) - \cos\left(\frac{j\pi}{N}\right)\right) \neq 0$ si $1 \leq j < k \leq (N - 1)$

□

Comme nous voulons que la solution décroisse exponentiellement, peu importe le choix de la condition initiale, une condition nécessaire pour ceci est que $|\kappa_n| < 1$ pour $n = 1, 2, \dots, (N - 1)$. Sinon nous aurions pour certaines fonctions $f(x)$ des solutions comportant des oscillations très grandes pour des j suffisamment grands.

Nous dirons donc que notre méthode numérique est **stable** si $|\kappa_n| < 1$ pour tout $n = 1, 2, \dots, (N - 1)$. Sinon elle est **instable**. Comme

$$\kappa_n = \left(1 - 2\sigma \left(1 - \cos\left(\frac{n\pi}{N}\right) \right) \right)$$

et $\sigma > 0$, nous avons toujours $\kappa_n < 1$ pour $n = 1, 2, \dots, (N - 1)$. Pour la stabilité de notre méthode, nous devons donc avoir que $\kappa_n > -1$. Ainsi

$$\kappa_n = \left(1 - 2\sigma \left(1 - \cos\left(\frac{n\pi}{N}\right) \right) \right) > -1 \quad \Rightarrow \quad \sigma < \frac{1}{\left(1 - \cos\left(\frac{n\pi}{N}\right) \right)} \quad \text{pour tout } n = 1, 2, \dots, (N - 1).$$

Le terme de droite de cette dernière inégalité décroît avec n . Il suffit alors que

$$\sigma < \frac{1}{\left(1 - \cos\left(\frac{(N-1)\pi}{N}\right) \right)}$$

Lorsque N est grand, nous avons que $\cos((N-1)\pi/N) \approx -1$. Nous avons en fait que

$$\frac{1}{2} < \frac{1}{\left(1 - \cos\left(\frac{(N-1)\pi}{N}\right) \right)} \quad \text{pour tout } N \in \mathbf{N}, N > 0.$$

Si nous prenons $\sigma \leq (1/2)$, alors notre méthode numérique sera stable. Dans nos exemples numériques précédents, nous avons premièrement

$$\sigma = \frac{1^2 (0.0025)}{(0.1)^2} = \frac{1}{4} \quad \text{si } \Delta x = 0.1 \text{ et } \Delta t = 0.0025$$

et nos calculs illustraient la stabilité; alors que

$$\sigma = \frac{1^2 (0.01)}{(0.1)^2} = 1 \quad \text{si } \Delta x = 0.1 \text{ et } \Delta t = 0.01$$

dans la deuxième situation et nos calculs illustraient bien l'instabilité.

Nous devons considérer aussi la question de la convergence de la solution numérique. Fixons $\sigma = c^2 (\Delta t)/(\Delta x)^2$ dans notre méthode numérique. L'équation [3] est alors

$$U_i^{(j+1)} = U_i^{(j)} + \sigma(U_{i+1}^{(j)} - 2U_i^{(j)} + U_{i-1}^{(j)})$$

et considérons des maillages de notre domaine tels que $\Delta x \rightarrow 0$ et $\Delta t \rightarrow 0$ (avec σ fixé), i.e. que les mailles deviennent de plus en plus fines, nous dirons alors que la méthode numérique converge si pour tout point (x, t) dans notre domaine et $(x_i, t_j) = (i(\Delta x), j(\Delta t)) \rightarrow (x, t)$, alors $U_i^{(j)} \rightarrow u(x, t)$. Ici nous supposons que $x \neq 0, \ell$ et $t > 0$.

Nous allons premièrement esquisser un argument heuristique justifiant la convergence. Supposons que le point (x, t) appartient à tous les maillages sur lesquels nous considérons la limite, i.e. $x = x_i = i(\Delta x)$ et $t = t_j = j(\Delta t)$. Nous pouvons toujours nous restreindre à ce cas à cause de la continuité de $u(x, t)$. Nous avons vu plus tôt que $U_i^{(j)}$ est une somme finie de fonctions de la forme

$$\sin\left(\frac{n\pi i}{N}\right) \kappa_n^j = \sin\left(\frac{n\pi i}{N}\right) \left[1 - 2\sigma \left(1 - \cos\left(\frac{n\pi}{N}\right)\right)\right]^j,$$

alors que $u(x, t)$ est une somme infinie de fonctions de la forme

$$\sin\left(\frac{n\pi x}{\ell}\right) \exp\left[-\left(\frac{cn\pi}{\ell}\right)^2 t\right].$$

Dans ce dernier cas, il faut noter que la convergence de la série est très rapide et que $u(x, t)$ est approximativement égal à une somme finie. Maintenant esquissons la preuve que

$$\sin\left(\frac{n\pi i}{N}\right) \kappa_n^j \rightarrow \sin\left(\frac{n\pi x}{\ell}\right) \exp\left[-\left(\frac{cn\pi}{\ell}\right)^2 t\right]$$

lorsque $\Delta x \rightarrow 0$ et $\Delta t \rightarrow 0$ en supposant que $(n/N) \ll 1$, i.e. (n/N) est petit relativement à 1. En effet,

$$N(\Delta x) = \ell \quad \text{et} \quad i(\Delta x) = x \quad \Rightarrow \quad \sin\left(\frac{n\pi i}{N}\right) = \sin\left(\frac{n\pi i(\Delta x)}{\ell}\right) = \sin\left(\frac{n\pi x}{\ell}\right).$$

De même, parce que

$$\frac{n}{N} \ll 1, \quad \sigma = \frac{c^2(\Delta t)}{(\Delta x)^2}, \quad N(\Delta x) = \ell \quad \text{et} \quad t = j(\Delta t) \quad \Rightarrow \quad \cos\left(\frac{n\pi}{N}\right) \approx 1 - \frac{1}{2} \left(\frac{n\pi}{N}\right)^2 \quad \text{et}$$

$$\begin{aligned} \kappa_n^j &= \left[1 - 2\sigma \left(1 - \cos\left(\frac{n\pi}{N}\right)\right)\right]^j \approx \left[1 - \sigma \left(\frac{n\pi}{N}\right)^2\right]^j = \left[1 - \frac{c^2(\Delta t)}{(\Delta x)^2} \left(\frac{n\pi(\Delta x)}{\ell}\right)^2\right]^j \\ &\approx \left[1 - \left(\frac{cn\pi}{\ell}\right)^2 (\Delta t)\right]^j = \left[1 - \left(\frac{cn\pi}{\ell}\right)^2 \left(\frac{t}{j}\right)\right]^j. \end{aligned}$$

Mais $(\Delta x \rightarrow 0 \text{ et } i(\Delta x) = x) \Rightarrow i \rightarrow \infty$, car $x \neq 0$. De même, $(\Delta t \rightarrow 0 \text{ et } j(\Delta t) = t) \Rightarrow j \rightarrow \infty$, car $t \neq 0$. Avant de considérer la limite, rappelons que

$$\lim_{j \rightarrow \infty} \left[1 + \frac{a}{j}\right]^j = e^a \quad \text{pour tout } a \in \mathbf{R}.$$

Finalement si nous considérons la limite lorsque $\Delta x \rightarrow 0$ et $\Delta t \rightarrow 0$, i.e. $i \rightarrow \infty$ et $j \rightarrow \infty$, alors

$$\sin\left(\frac{n\pi i}{N}\right) \kappa_n^j \approx \sin\left(\frac{n\pi i}{N}\right) \left[1 - \left(\frac{cn\pi}{\ell}\right)^2 \left(\frac{t}{j}\right)\right]^j \rightarrow \sin\left(\frac{n\pi x}{\ell}\right) \exp\left[-\left(\frac{cn\pi}{\ell}\right)^2 t\right].$$

Ainsi la somme finie

$$\sum_{n=1}^{N-1} a_n \sin\left(\frac{n\pi i}{N}\right) \kappa_n^j$$

est approximativement égale à la somme

$$\sum_{n=1}^{N-1} a_n \sin\left(\frac{n\pi x}{\ell}\right) \exp\left[-\left(\frac{cn\pi}{\ell}\right)^2 t\right]$$

qui est approximativement égale à $u(x, t)$.

Si nous voulons être plus précis, il faut tenir compte des termes d'erreur. C'est ce que nous allons maintenant faire. Nous allons vérifier que si $\sigma < 1/2$, alors $U_i^{(j)}$ converge vers $u(x_i, t_j)$. Dans ce qui suivra, nous allons supposer que les dérivées partielles

$$\frac{\partial^2 u}{\partial t^2} \quad \text{et} \quad \frac{\partial^4 u}{\partial x^4}$$

sont continues sur tout domaine de la forme $[0, 1] \times [0, t_F]$ où $t_F \in \mathbf{R}$, $t_F > 0$.

Étant donné $\Delta x > 0$ et $\Delta t > 0$, nous avons vu au début du chapitre que

$$u(x, t + \Delta t) = u(x, t) + \frac{\partial u}{\partial t}(x, t) (\Delta t) + \frac{1}{2} \frac{\partial^2 u}{\partial t^2}(x, \alpha) (\Delta t)^2 \quad (\text{éq. [4]})$$

et

$$u(x + \Delta x, t) + u(x - \Delta x, t) = 2u(x, t) + \frac{\partial^2 u}{\partial x^2}(x, t) (\Delta x)^2 + \frac{1}{12} \frac{\partial^4 u}{\partial x^4}(\beta, t) (\Delta x)^4 \quad (\text{éq. [5]})$$

où $t \leq \alpha \leq t + \Delta t$ et $x - \Delta x \leq \beta \leq x + \Delta x$. En multipliant l'équation [4] par $1/(\Delta t)$, l'équation [5] par $c^2/(\Delta x)^2$ et en soustrayant ces deux équations, nous obtenons pour une solution u de l'équation de la chaleur

$$\frac{1}{(\Delta t)} u(x, t + \Delta t) - \frac{c^2}{(\Delta x)^2} (u(x + \Delta x, t) + u(x - \Delta x, t))$$

est égal à

$$\left[\frac{1}{(\Delta t)} - \frac{2c^2}{(\Delta x)^2} \right] u(x, t) + \left[\frac{1}{2} \frac{\partial^2 u}{\partial t^2}(x, \alpha) (\Delta t) - \frac{c^2}{12} \frac{\partial^4 u}{\partial x^4}(\beta, t) (\Delta x)^2 \right].$$

Nous avons utilisé le fait que u est une solution de l'équation de la chaleur, i.e.

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

Rappelons que $\sigma = c^2(\Delta t)/(\Delta x)^2$. Alors

$$u(x, t + \Delta t) = [\sigma u(x + \Delta x, t) + (1 - 2\sigma) u(x, t) + \sigma u(x - \Delta x, t)] + \left[\frac{1}{2} \frac{\partial^2 u}{\partial t^2}(x, \alpha) (\Delta t)^2 - \frac{c^4}{12\sigma} \frac{\partial^4 u}{\partial x^4}(\beta, t) (\Delta t)^2 \right].$$

Noter que

$$\left[\frac{1}{2} \frac{\partial^2 u}{\partial t^2}(x, \alpha) (\Delta t)^2 - \frac{c^4}{12\sigma} \frac{\partial^4 u}{\partial x^4}(\beta, t) (\Delta t)^2 \right] = E(\Delta t),$$

où E est le terme d'erreur de discrétisation

Nous allons maintenant analyser le terme d'erreur $\epsilon_i^{(j)} = u(x_i, t_j) - U_i^{(j)}$ pour $1 \leq i \leq (N - 1)$ et $j \geq 0$. Ici $x_i = i(\Delta x)$ et $t_j = j(\Delta t)$. Ainsi avec l'équation ci-dessus et l'équation [3] définissant $U_i^{(j)}$, nous obtenons

$$\begin{aligned} \epsilon_i^{(j+1)} &= u(x_i, t_{j+1}) - U_i^{(j+1)} \\ &= \sigma u(x_{i+1}, t_j) + (1 - 2\sigma) u(x_i, t_j) + \sigma u(x_{i-1}, t_j) + \left[\frac{1}{2} \frac{\partial^2 u}{\partial t^2}(x_i, \alpha_j) (\Delta t)^2 - \frac{c^4}{12\sigma} \frac{\partial^4 u}{\partial x^4}(\beta_i, t_j) (\Delta t)^2 \right] \\ &\quad - [\sigma U_{i+1}^{(j)} + (1 - 2\sigma) U_i^{(j)} + \sigma U_{i-1}^{(j)}] \\ &= \sigma \epsilon_{i+1}^{(j)} + (1 - 2\sigma) \epsilon_i^{(j)} + \sigma \epsilon_{i-1}^{(j)} + \left[\frac{1}{2} \frac{\partial^2 u}{\partial t^2}(x_i, \alpha_j) (\Delta t)^2 - \frac{c^4}{12\sigma} \frac{\partial^4 u}{\partial x^4}(\beta_i, t_j) (\Delta t)^2 \right] \end{aligned}$$

où $x_i - (\Delta x) \leq \beta_i \leq x_i + (\Delta x)$ et $t_j \leq \alpha_j \leq t_j + (\Delta t)$.

Soit $M^{(j)} = \max\{|\epsilon_i^{(j)}| \mid i = 1, 2, \dots, (N-1)\}$, l'erreur maximum en valeur absolue pour un temps donné t_j . En utilisant l'inégalité du triangle et le fait que les coefficients σ et $(1-2\sigma)$ sont ≥ 0 parce que $0 < \sigma < (1/2)$, nous obtenons

$$\begin{aligned} |\epsilon_i^{(j+1)}| &\leq |\sigma \epsilon_{i+1}^{(j)}| + |(1-2\sigma) \epsilon_i^{(j)}| + |\sigma \epsilon_i^{(j)}| + \left| \frac{1}{2} \frac{\partial^2 u}{\partial t^2}(x_i, \alpha_j) (\Delta t)^2 - \frac{c^4}{12\sigma} \frac{\partial^4 u}{\partial x^4}(\beta_i, t_j) (\Delta t)^2 \right| \\ &\leq \sigma M^{(j)} + (1-2\sigma) M^{(j)} + \sigma M^{(j)} + \left| \frac{1}{2} \frac{\partial^2 u}{\partial t^2}(x_i, \alpha_j) (\Delta t)^2 - \frac{c^4}{12\sigma} \frac{\partial^4 u}{\partial x^4}(\beta_i, t_j) (\Delta t)^2 \right| \\ &\leq M^{(j)} + \left| \frac{1}{2} \frac{\partial^2 u}{\partial t^2}(x_i, \alpha_j) - \frac{c^4}{12\sigma} \frac{\partial^4 u}{\partial x^4}(\beta_i, t_j) \right| (\Delta t)^2. \end{aligned}$$

Comme les points (x_i, α_j) , (β_i, t_j) appartiennent à un domaine $[0, 1] \times [0, t_F]$ pour un certain $t_F > 0$ et que

$$\frac{1}{2} \frac{\partial^2 u}{\partial t^2} - \frac{c^4}{12\sigma} \frac{\partial^4 u}{\partial x^4}$$

est continue sur ce domaine compact (i.e. un ensemble borné et fermé), alors cette fonction est bornée, en particulier il existe un nombre réel positif R tel que

$$\left| \frac{1}{2} \frac{\partial^2 u}{\partial t^2}(x_i, \alpha_j) - \frac{c^4}{12\sigma} \frac{\partial^4 u}{\partial x^4}(\beta_i, t_j) \right| \leq R \quad \text{pour tout } i, j.$$

Ainsi $|\epsilon_i^{(j+1)}| \leq M^{(j)} + R(\Delta t)^2$ et conséquemment $M^{(j+1)} \leq M^{(j)} + R(\Delta t)^2$ pour tout $j \geq 0$. Nous avons que $M^{(0)} = 0$ par notre choix de $U_i^{(0)}$. Il est alors facile de déduire par récurrence que $M^{(j)} \leq j R(\Delta t)^2$. Comme précédemment, nous supposons que (x, t) est un des points de maillage pour lesquels nous calculons la limite, i.e. $t = j(\Delta t)$ et $M^{(j)} \leq R t(\Delta t)$. Donc si $(\Delta t) \rightarrow 0$, alors $j \rightarrow \infty$ et $M^{(j)} \rightarrow 0$. Ceci montre que l'algorithme converge si $\sigma < (1/2)$.

Pour clore ce chapitre, nous allons présenter une autre méthode numérique pour approximer la solution $u(x, t)$, celle de Crank-Nicolson.

Une approche à laquelle nous pourrions penser pour notre problème de la chaleur est d'utiliser l'approximation par différence finie centrée pour la dérivée partielle de u relativement à t . Celle-ci a été proposée par Richardson en 1910. En d'autres mots, nous aurions alors que

$$\frac{\partial u}{\partial t}(x, t) \approx \frac{u(x, t + (\Delta t)) - u(x, t - (\Delta t))}{2(\Delta t)}$$

et nous aurions comme équation à résoudre

$$\frac{U_i^{(j+1)} - U_i^{(j-1)}}{2(\Delta t)} = \frac{c^2}{(\Delta x)^2} [U_{i+1}^{(j)} - 2U_i^{(j)} + U_{i-1}^{(j)}].$$

Noter que dans ce cas, il nous faut ajouter les conditions: $U_0^{(j)} = U_0^{(N)} = 0$ pour tout $j \geq 0$, ainsi que des conditions initiales permettant de déterminer $U_i^{(0)}$ et $U_i^{(1)}$ pour $i = 1, 2, \dots, (N-1)$. Cependant cette méthode n'est maintenant jamais utilisée, parce qu'elle est instable peu importe les valeurs de c^2 , Δx et Δt .

John Crank et Phyllis Nicolson ont proposé en 1947 une méthode alternative. Il est possible de visualiser l'approximation par différence finie progressive pour la dérivée partielle de u relativement à t comme étant l'approximation par différence finie centrée autour de $t + (\Delta t/2)$. Pour la dérivée partielle d'ordre deux par rapport à x , nous voudrions l'évaluer à $t + (\Delta t/2)$, nous faisons la moyenne arithmétique des approximations de cette dérivée évaluée à (x, t) et $(x, t + \Delta t)$. Dans ce cas, la méthode de Crank-Nicolson est

$$\frac{U_i^{(j+1)} - U_i^{(j)}}{(\Delta t)} = \frac{c^2}{2} \left[\frac{U_{i+1}^{(j)} - 2U_i^{(j)} + U_{i-1}^{(j)}}{(\Delta x)^2} + \frac{U_{i+1}^{(j+1)} - 2U_i^{(j+1)} + U_{i-1}^{(j+1)}}{(\Delta x)^2} \right].$$

Cette méthode est différente de la première méthode présentée au début du chapitre. Elle est implicite plutôt qu'explicite.

En effet, l'équation ci-dessus est équivalente à

$$-\frac{\sigma}{2}U_{i+1}^{(j+1)} + (1 + \sigma)U_i^{(j+1)} - \frac{\sigma}{2}U_{i-1}^{(j+1)} = \frac{\sigma}{2}U_{i+1}^{(j)} + (1 - \sigma)U_i^{(j)} + \frac{\sigma}{2}U_{i-1}^{(j)} \quad (\text{éq. [6]})$$

pour $1 \leq i \leq (N-1)$ et $j \geq 0$. Il est aussi possible de généraliser la méthode de Crank-Nicolson en considérant au lieu de la moyenne arithmétique des approximations de la dérivée partielle d'ordre 2 par rapport à x un point quelconque entre ces deux valeurs approximatives. Nous parlerons alors de la θ -méthode. Ainsi

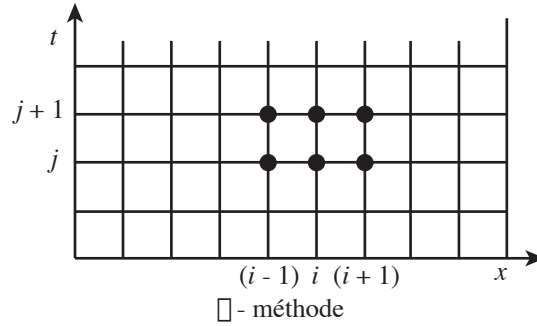
$$\frac{U_i^{(j+1)} - U_i^{(j)}}{(\Delta t)} = c^2 \left[(1 - \theta) \left(\frac{U_{i+1}^{(j)} - 2U_i^{(j)} + U_{i-1}^{(j)}}{(\Delta x)^2} \right) + \theta \left(\frac{U_{i+1}^{(j+1)} - 2U_i^{(j+1)} + U_{i-1}^{(j+1)}}{(\Delta x)^2} \right) \right],$$

pour $1 \leq i \leq (N-1)$ et $j \geq 0$ et où $0 \leq \theta \leq 1$ dans cette dernière méthode. Cette équation est équivalente à

$$-\theta\sigma U_{i+1}^{(j+1)} + (1 + 2\theta\sigma)U_i^{(j+1)} - \theta\sigma U_{i-1}^{(j+1)} = (1 - \theta)\sigma U_{i+1}^{(j)} + (1 - 2(1 - \theta)\sigma)U_i^{(j)} + (1 - \theta)\sigma U_{i-1}^{(j)} \quad (\text{éq. [7]})$$

pour $1 \leq i \leq (N-1)$ et $j \geq 0$. Il nous faut aussi ajouter les conditions: $U_0^{(j)} = U_N^{(j)} = 0$ pour $j \geq 0$ et $U_i^{(0)} = f(i \Delta x)$ pour $1 \leq i \leq (N-1)$. Si $\theta = (1/2)$, nous obtenons la méthode de Crank-Nicolson, i.e. l'équation [6], alors que si $\theta = 0$, nous avons la méthode explicite, i.e. l'équation [3]

Nous avons illustré ci-dessous les points du maillage qui interviennent dans les équations [6] et [7].



Il faut donc résoudre un système d'équations linéaires pour déterminer $U_i^{(j)}$ dans la θ -méthode (lorsque $0 < \theta \leq 1$). C'est l'explication du qualificatif "implicite" pour décrire ces méthodes. Plus précisément, en supposant que $U_i^{(j)}$ connu pour tout $0 \leq i \leq N$, alors nous avons à résoudre un système de $(N-1)$ équations linéaires dont les inconnues sont $U_1^{(j+1)}, U_2^{(j+1)}, \dots, U_{N-1}^{(j+1)}$. Ce système est de la forme suivante:

$$\begin{pmatrix} (1 + 2\theta\sigma) & -\theta\sigma & 0 & \dots & \dots & \dots & 0 \\ -\theta\sigma & (1 + 2\theta\sigma) & -\theta\sigma & 0 & \dots & \dots & 0 \\ 0 & -\theta\sigma & (1 + 2\theta\sigma) & -\theta\sigma & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & -\theta\sigma & (1 + 2\theta\sigma) & -\theta\sigma & 0 \\ 0 & \dots & \dots & 0 & -\theta\sigma & (1 + 2\theta\sigma) & -\theta\sigma \\ 0 & \dots & \dots & \dots & 0 & -\theta\sigma & (1 + 2\theta\sigma) \end{pmatrix} \begin{pmatrix} U_1^{(j+1)} \\ U_2^{(j+1)} \\ \vdots \\ U_{N-1}^{(j+1)} \end{pmatrix} = \mathbf{V}^{(j)}$$

où $\mathbf{V}^{(j)}$ est un vecteur colonne $(N-1) \times 1$ fonction des valeurs $U_i^{(j)}$, $1 \leq i \leq (N-1)$.

Ce système d'équations linéaires peut être résolu sans trop de peine. La matrice du système est tridiagonale et diagonalement dominante, i.e. chaque entrée sur la diagonale est strictement supérieure à la somme des valeurs absolues des autres entrées sur la ligne ou la colonne la contenant. Cette θ -méthode pour $0 < \theta \leq 1$ nécessite plus de travail que la méthode explicite ($\theta = 0$), mais elle a des avantages quant à la stabilité. C'est ce que nous allons maintenant étudier.

Proposition 1

(a) Soit $(1/2) \leq \theta \leq 1$. Alors la θ -méthode est numériquement stable. Nous disons alors que la θ -méthode est inconditionnellement stable.

(b) Soit $0 \leq \theta < (1/2)$. Si $\sigma \leq (1/2(1-2\theta))$, alors la θ -méthode est numériquement stable. Nous disons alors que la θ -méthode est conditionnellement stable.

Preuve: Le cas où $\theta = 0$ a été étudié précédemment, il s'agit de la méthode explicite, et nous avons montré que la méthode est numériquement stable si $\sigma \leq (1/2)$. Donc la partie (b) de la proposition pour $\theta = 0$ est vérifiée. Par la suite, nous supposons que $\theta > 0$.

Pour étudier la stabilité, nous procéderons comme pour la méthode explicite. Considérons le problème intermédiaire suivant:

$$(\spadesuit) \quad \begin{cases} -\theta\sigma U_{i+1}^{(j+1)} + (1+2\theta\sigma)U_i^{(j+1)} - \theta\sigma U_{i-1}^{(j+1)} = (1-\theta)\sigma U_{i+1}^{(j)} + (1-2(1-\theta)\sigma)U_i^{(j)} + (1-\theta)\sigma U_{i-1}^{(j)} \\ \text{pour } 1 \leq i \leq (N-1) \text{ et } j \leq 0 \text{ avec la condition } U_0^{(j)} = U_N^{(j)} = 0 \text{ pour tout } j \geq 0 \end{cases}$$

Par rapport à la θ -méthode, nous avons laissé tomber la condition initiale $U_i^{(0)} = f(i\Delta x)$ pour $1 \leq i \leq (N-1)$. Nous allons déterminer des solutions non triviales du système (\spadesuit) de la forme $U_i^{(j)} = F(i)G(j)$. En remplaçant dans l'équation, nous obtenons

$$-\theta\sigma F(i+1)G(j+1) + (1+2\theta\sigma)F(i)G(j+1) - \theta\sigma F(i-1)G(j+1)$$

est égal

$$(1-\theta)\sigma F(i+1)G(j) + (1+2\theta\sigma-2\sigma)F(i)G(j) + (1-\theta)\sigma F(i-1)G(j).$$

Nous pouvons séparer ces deux fonctions et nous obtenons

$$\begin{aligned} \frac{G(j+1)}{G(j)} &= \frac{(1-\theta)\sigma F(i+1) + (1+2\theta\sigma-2\sigma)F(i) + (1-\theta)\sigma F(i-1)}{-\theta\sigma F(i+1) + (1+2\theta\sigma)F(i) - \theta\sigma F(i-1)} \\ &= \left[\frac{\sigma F(i+1) - 2\sigma F(i) + \sigma F(i-1)}{-\theta\sigma F(i+1) + (1+2\theta\sigma)F(i) - \theta\sigma F(i-1)} \right] + 1. \end{aligned}$$

Le terme de gauche est une fonction de j et celui de droite est une fonction de i . Pour que l'équation ci-dessus soit vérifiée, il faut que chacun des termes soit constant. Donc

$$\frac{G(j+1)}{G(j)} = \left[\frac{\sigma F(i+1) - 2\sigma F(i) + \sigma F(i-1)}{-\theta\sigma F(i+1) + (1+2\theta\sigma)F(i) - \theta\sigma F(i-1)} \right] + 1 = \lambda$$

où λ est une constante. Posons $\lambda' = \lambda - 1$. Nous obtenons ainsi deux équations:

$$\sigma(1+\theta\lambda')F(i+1) - (2\sigma+\lambda'+2\theta\sigma\lambda')F(i) + \sigma(1+\theta\lambda')F(i-1) = 0 \quad \text{et} \quad G(j+1) = \lambda G(j)$$

pour $1 \leq i \leq (N-1)$ et $j \leq 0$. À ceci, il faut ajouter les conditions suivantes:

$$U_0^{(j)} = F(0)G(j) = 0, \quad \forall j \geq 0 \quad \Rightarrow \quad F(0) = 0 \quad \text{et} \quad U_N^{(j)} = F(N)G(j) = 0, \quad \forall j \geq 0 \quad \Rightarrow \quad F(N) = 0.$$

En résumé, nous devons résoudre

$$\sigma(1 + \theta\lambda')F(i+1) - (2\sigma + \lambda' + 2\theta\sigma\lambda')F(i) + \sigma(1 + \theta\lambda')F(i-1) = 0 \quad \text{avec } F(0) = F(N) = 0 \text{ et} \\ G(j+1) = \lambda G(j)$$

pour $1 \leq i \leq (N-1)$ et $j \leq 0$.

Si $(1 + \theta\lambda') = 0$, i.e. $\lambda' = -\theta^{-1}$, alors la première équation équivalente à $F(i) = 0$ pour tout $1 \leq i \leq (N-1)$. Comme nous cherchons des solutions non triviales, nous pouvons donc supposer par la suite que $(1 + \theta\lambda') \neq 0$, i.e. $\lambda' \neq -\theta^{-1}$.

Si $(1 + \theta\lambda') \neq 0$, i.e. $\lambda' \neq -\theta^{-1}$, alors la première équation est d'ordre 2 et elle est équivalente à

$$F(i+1) - \left[2 + \frac{\lambda'}{\sigma(1 + \theta\lambda')}\right] F(i) + F(i-1) = 0$$

après avoir divisé par $\sigma(1 + \theta\lambda')$. Il nous faut étudier les racines du polynôme

$$x^2 - \left[2 + \frac{\lambda'}{\sigma(1 + \theta\lambda')}\right] x + 1 = 0.$$

Nous avons trois cas: (i) deux racines réelles distinctes; (ii) une racine réelle double; (iii) deux racines complexes non réelles distinctes. Dans les deux premiers cas (i) et (ii), à cause de la condition $F(0) = F(N) = 0$, nous obtenons que $F(i) = 0$ pour $1 \leq i \leq (N-1)$ et il nous faut rejeter ces deux cas. Il nous faut donc considérer seulement le cas (iii). Ces deux racines sont

$$1 + \frac{\lambda'}{2\sigma(1 + \theta\lambda')} + \sqrt{\frac{\lambda'}{\sigma(1 + \theta\lambda')} \left[\frac{\lambda'}{\sigma(1 + \theta\lambda')} + 4\right]} \quad \text{et} \quad 1 + \frac{\lambda'}{2\sigma(1 + \theta\lambda')} - \sqrt{\frac{\lambda'}{\sigma(1 + \theta\lambda')} \left[\frac{\lambda'}{\sigma(1 + \theta\lambda')} + 4\right]}.$$

Parce que ces racines sont complexes non réelles, nous devons avoir que

$$\frac{\lambda'}{\sigma(1 + \theta\lambda')} \left[\frac{\lambda'}{\sigma(1 + \theta\lambda')} + 4\right] < 0 \quad \iff \quad -4 < \frac{\lambda'}{\sigma(1 + \theta\lambda')} < 0$$

Si ρ est la norme de ces deux racines, alors comme pour le cas explicite et à cause de la condition $F(0) = F(N) = 0$, nous obtenons que

$$F(i) = \rho^i \sin\left(\frac{n\pi i}{N}\right) \quad \text{avec } n \in \mathbf{N}, n > 0, n \not\equiv 0 \pmod{N}$$

Nous pourrions calculer explicitement ρ , mais ceci ne sera pas nécessaire pour la stabilité. Pour la deuxième équation $G(j+1) = \lambda G(j)$, nous obtenons que la solution est $G(j) = A\lambda^j$. Nous obtenons donc que

$$U_i^{(j)} = A\rho^i \sin\left(\frac{n\pi i}{N}\right) \lambda^j$$

avec $n \in \mathbf{N}$, $n > 0$, $n \not\equiv 0 \pmod{N}$, est une solution du système (\spadesuit), où $0 \leq i \leq N$ et $j \geq 0$. Ici ρ et λ dépendent de n et nous écrirons ρ_n et λ_n pour souligner cette dépendance. Il est possible de restreindre n entre 1 et $N-1$. Finalement nous obtenons que la solution de (\spadesuit) est de la forme

$$U_i^{(j)} = \sum_{n=1}^{N-1} a_n \rho_n^i \sin\left(\frac{n\pi i}{N}\right) \lambda_n^j.$$

Pour la stabilité, il faut s'assurer que $|\lambda_n| < 1$. Rappelons que $\lambda_n = \lambda'_n + 1$. Conséquemment pour la stabilité, il faut que s'assurer que $-2 < \lambda'_n < 0$. Nous avons montré ci-dessus que

$$-4 < \frac{\lambda'_n}{\sigma(1 + \theta\lambda'_n)} < 0 \quad \text{(inégalité [1])}$$

Noter que $\sigma > 0$. Si $(1 + \theta\lambda'_n) > 0$, i.e. $\lambda'_n > -\theta^{-1}$, alors nous obtenons de l'inégalité [1]:

$$-4\sigma(1 + \theta\lambda'_n) < \lambda'_n < 0 \quad \Rightarrow \quad -4\sigma < (1 + 4\sigma\theta)\lambda'_n < 4\sigma\theta\lambda'_n \quad \Rightarrow \quad -\frac{4\sigma}{(1 + 4\sigma\theta)} < \lambda'_n < 0.$$

Si $(1 + \theta\lambda'_n) < 0$, i.e. $\lambda'_n < -\theta^{-1}$, alors nous obtenons de l'inégalité [1]: $-4\sigma(1 + \theta\lambda'_n) > \lambda'_n > 0$. Nous avons alors une contradiction, car $\theta > 0 \Rightarrow \lambda'_n < 0$ et aussi $\lambda'_n > 0$. Ainsi nous avons l'inégalité

$$-\frac{4\sigma}{(1 + 4\sigma\theta)} < \lambda'_n < 0.$$

Si maintenant

$$-2 \leq -\frac{4\sigma}{(1 + 4\sigma\theta)},$$

alors nous aurons la stabilité de la θ -méthode. Nous avons les équivalences suivantes

$$-2 \leq -\frac{4\sigma}{(1 + 4\sigma\theta)} \quad \Leftrightarrow \quad -2(1 + 4\sigma\theta) \leq -4\sigma \quad \Leftrightarrow \quad -1 \leq -2\sigma(1 - 2\theta). \quad (\text{inégalité [2]})$$

Montrons maintenant (a). Soit $(1/2) \leq \theta \leq 1$. Alors $(1 - 2\theta) \leq 0$ et $-2\sigma(1 - 2\theta) \geq 0$. Conséquemment l'inégalité [2] est vérifiée et par le fait même la θ -méthode est numériquement stable.

Montrons maintenant (b). Soit $0 < \theta < (1/2)$. Si

$$\sigma \leq \frac{1}{2(1 - 2\theta)} \quad \Rightarrow \quad 2\sigma(1 - 2\theta) \leq 1 \quad \Rightarrow \quad -2\sigma(1 - 2\theta) \geq -1$$

car $(1 - 2\theta) > 0$. Ainsi l'inégalité [2] est vérifiée et par le fait même la θ -méthode est numériquement stable si

$$\sigma \leq \frac{1}{2(1 - 2\theta)}.$$

□