Cours 3

Erreur de troncature

1) Problématique
2) Schéma d'Euler explicite
3) Schéma d'Euler implicite
4) Schéma de Crank-Nicolson
5) Schéma de Heun
6) Définition générale
7) Erreur et erreur de troncature

François Dubois, mars 2003, édition 20 juillet 2006, 15 pages.

ch(3) Erreur de troncature

1) Problématique

- Nous cherchons à connaître les solutions $t \mapsto u(t) \in \mathbb{R}^m$ de l'équation différentielle

$$(1) \quad \frac{du}{dt} = f(u), \quad t \geq 0$$

au moins de manière approchée aux instants

$$(2) \quad t^k = k\Delta t, \quad k \in \mathbb{N}, \quad \Delta t > 0 \text{ fixé}.$$

Pour cela, nous avons à notre disposition les quatre schémas proposés au chapitre précédent (et il y en a aussi beaucoup d'autres!) qui fournissent une valeur approchée u^k à l'instant t^k, que nous devons comparer à la valeur $u(t^k)$.

- Plus précisément, fixons $T > 0$, relativement "grand" devant un temps caractéristique de variation des solutions de l'équation (1). Nous notons $u^k_{\Delta t}$ la valeur numérique approchée obtenue avec un schéma numérique. L'écart entre la solution exacte $u(k\Delta t)$ et cette valeur approchée $u^k_{\Delta t}$ est noté $\varepsilon^k_{\Delta t}$:

$$(3) \quad \varepsilon^k_{\Delta t} = u^k_{\Delta t} - u(k\Delta t), \quad k\Delta t \leq T.$$

Nous voulons que cet écart soit "petit", ce pour tous les instants discrets $k\Delta t$, jusqu'au temps T. On introduit donc l'erreur $\delta(\Delta t)$:

$$(4) \qquad \delta(\Delta t) = \sup_{0 \leq k\Delta t \leq T} |\varepsilon^k_{\Delta t}|$$

et on souhaite que $\delta(\Delta t)$ soit "petit" pour $\Delta t > 0$ "assez petit". En d'autres termes, on souhaite que $\delta(\Delta t)$ tende vers zéro lorsque Δt tend vers zéro?

- Comment aborder l'étude de ce problème? En effet, si on sait a priori calculer $u^k_{\Delta t}$ avec l'algorithmique du schéma numérique (et sa mise en œuvre sur ordinateur), on ignore tout de $u(k\Delta t)$, valeur de la solution exacte de l'équation (1) à l'instant discret $k\Delta t$. L'astuce consiste à renverser les rôles, ce qui conduit à la notion d'erreur de troncature. Nous verrons en fin de chapitre que si le schéma est stable, l'erreur de tronca. donne une bonne idée de l'erreur $\delta(\Delta t)$.

2) Schéma d'Euler explicite

- Nous l'écrivons

$$(5) \quad \frac{1}{\Delta t}\left(u^{k+1} - u^k\right) - f(u^k) = 0 .$$

A partir d'une valeur u^k à l'instant discret $k\Delta t$, le schéma d'Euler explicite reconstitue une valeur u^{k+1} qui veut être une valeur approchée de $u((k+1)\Delta t)$.
Imaginons qu'on parte de la valeur exacte $u(k\Delta t)$, ie qu'on suppose $u^k = u(k\Delta t)$
Le schéma numérique (5) calcule une valeur u^{k+1} différente de $u(k\Delta t)$; puisque le schéma numérique (5) n'est qu'une approximation de la relation exacte

$$(6) \quad \frac{1}{\Delta t}\left[u((k+1)\Delta t) - u(k\Delta t)\right] - \frac{1}{\Delta t}\int_{k\Delta t}^{(k+1)\Delta t} f(u(t))\,dt = 0$$

- Nous injectons $u^k = u(t^k)$, valeur exacte dans le schéma (5). Nous avons alors

$$u^{k+1} = u(k\Delta t) + \Delta t\, f(u(k\Delta t)) \neq u((k+1)\Delta t)$$

soit en d'autres termes,

$$(7) \quad \frac{1}{\Delta t}\left(u((k+1)\Delta t) - u(k\Delta t)\right) - f(u(k\Delta t)) \neq 0$$

La solution exacte de l'équation (1) n'est en général pas solution du schéma numér.

que (5). C'est cet écart qu'on appelle "erreur de troncature". Nous définissons cette erreur à l'ordre d'un temps $t > 0$ arbitraire, d'un pas de temps $\Delta t > 0$ quelconque également et pour une solution $u(\cdot)$ de l'équation différentielle (1). Nous posons pour le schéma d'Euler explicite :

$$(8)\quad \mathcal{E}(\Delta t, t; u(\cdot)) = \frac{1}{\Delta t}\big(u(t+\Delta t) - u(t)\big) - f(u(t)).$$

Cette erreur de troncature mesure "en quoi le schéma est mal vérifié pour une solution exacte de l'équation à résoudre". Si elle est grande, on a peu d'espoir. Si elle est "petite", très petite pour Δt assez petit, on imagine que le schéma "résout bien" l'équation et qu'en conséquence c'est l'erreur $\delta(\Delta t)$ qui sera petite !

- Même si on ne connaît pas la solution $u(\cdot)$ de l'équation (1), on peut faire le développement limité de $\mathcal{E}(\Delta t, t; u(\cdot))$ lorsque le pas de temps Δt tend vers zéro. Grâce à la formule de Taylor

$$(9)\quad u(t+\Delta t) = u(t) + \Delta t\, \frac{du(t)}{dt} + \frac{1}{2}\Delta t^2 \frac{d^2u}{dt^2} + o(\Delta t^3)$$

Si $u(\cdot)$ est assez régulière, nous tirons de (8) :

10) $\mathcal{E}(\Delta t, t; u(\cdot)) = \left[\frac{du}{dt}(t) - f(u(t))\right] + \frac{\Delta t}{2}\frac{d^2u}{dt^2} + O(\Delta t^2)$,

ce qui constitue un développement limité de l'erreur de troncature. Puisque $u(\cdot)$ est solution du système dynamique (1), le premier terme du membre de droite de la relation (10) est nul. Nous en déduisons, puisqu'a priori d^2u/dt^2 est non nul:

$$(11) \qquad \mathcal{E}(\Delta t, t; u(\cdot)) = O(\Delta t).$$

- Nous avons mis en évidence l'ordre asymptotique de convergence de l'erreur de troncature du schéma d'Euler explicite. Il est de la forme $O(\Delta t^1)$, avec la valeur "unité" comme exposant de Δt. Pour cette raison, on dit que le schéma d'Euler explicite est <u>d'ordre 1</u>.

3) Schéma d'Euler rétrograde

- Nous procédons pour le schéma

$$(12) \qquad \frac{1}{\Delta t}\left(u^{k+1} - u^k\right) - f(u^{k+1}) = 0$$

comme pour le schéma d'Euler explicite. Nous introduisons une solution $u(\cdot)$ de l'équation (1), et injectons les valeurs

$u(t)$, $u(t+\Delta t)$ dans l'expression (12) du schéma, remplaçant le temps discret $t^k = k\Delta t$ par le temps continu t :

$$(13)\quad \mathcal{C}(\Delta t, t; u(\cdot)) = \frac{1}{\Delta t}\left(u(t+\Delta t) - u(t)\right) - f(u(t+\Delta t))$$

Nous définissons ainsi l'erreur de troncature du schéma implicite. Afin de connaître son comportement asymptotique pour Δt tendant vers zéro, nous avons besoin de développer $f(u(t+\Delta t))$. Nous l'effectuons au troisième ordre de précision.

Lemme (1) Développement limité.

Pour $t \mapsto u(t)$ régulière et $u \mapsto f(u)$ régulière, nous avons

$$(14)\quad \begin{cases} f(u(t+\Delta t)) = f(u(t)) + \Delta t \dfrac{du}{dt} f'(u(t)) + \dfrac{\Delta t^2}{2}\dfrac{d^2u}{dt^2} \\ \qquad + \dfrac{1}{2}\left(\Delta t \dfrac{du}{dt}\right)^2 f''(u(t)) + O(\Delta t^3) \qquad f''(u(t)) \end{cases}$$

- Preuve du lemme 1.

Nous écrivons la formule de Taylor pour $f(u(t)+v)$, pour un infiniment petit v a priori arbitraire :

$$(15)\quad f(u(t)+v) = f(u(t)) + v f'(u(t)) + \frac{v^2}{2} f''(u(t)) + O(v^3)$$

puis nous particularisons v compte tenu du développement donné en (9) :

$$(16)\quad v = \Delta t \frac{du}{dt} + \frac{1}{2}\Delta t^2 \frac{d^2u}{dt^2} + O(\Delta t^3) .$$

On a donc

$$(17)\quad \frac{1}{2} v^2 = \frac{1}{2}\left(\Delta t \frac{du}{dt}\right)^2 + O(\Delta t^3)$$

$$(18)\quad O(v^3) = O(\Delta t^3) .$$

On injecte les relations (16) à (18) au sein du développement (15). On remarque que

$$f(u(t+\Delta t)) = f(u(t) + v + O(\Delta t^3))$$
$$= f(u(t)+v) + O(\Delta t^3) ,$$

donc

$$f(u(t+\Delta t)) = f(u(t)) + \left(\Delta t \frac{du}{dt} + \frac{1}{2}\Delta t^2 \frac{d^2u}{dt^2}\right) f'(u(t))$$
$$+ \frac{1}{2}\left(\Delta t \frac{du}{dt}\right)^2 f''(u(t)) + O(\Delta t^3) ,$$

ce qui constitue exactement le développement (14) annoncé. □

- Avec la définition (13) de l'erreur de troncature et le développement (14), on a

$$\mathcal{E}(\Delta t, t; u(\cdot)) = \frac{du}{dt} + \frac{\Delta t}{2}\frac{d^2u}{dt^2} + O(\Delta t^2) - \left[f(u(t)) + \Delta t \frac{du}{dt} f'(u(t))\right] + O(\Delta t^2)$$

$$= \left(\frac{du}{dt} - f(u(t))\right) + \Delta t\left(\frac{1}{2}\frac{d^2u}{dt^2} - \frac{du}{dt} f'(u(t))\right) + O(\Delta t^2) .$$

Si $u(\cdot)$ est solution de l'équation différentielle (1), on a $\frac{du}{dt} = f(u(t))$ et par dérivation par rapport au temps :

$$\frac{d^2u}{dt^2} = \frac{d}{dt}\left(f(u(t))\right) = f'(u(t)) \cdot \frac{du}{dt}$$

donc le développement de l'erreur de troncature s'écrit :

(19) $$\mathcal{E}(\Delta t, t; u(\cdot)) = -\frac{1}{2}\Delta t \frac{d^2u}{dt^2} + O(\Delta t^2)$$

ce qui montre que le schéma d'Euler rétrograde est d'ordre 1.

4) Schéma de Crank-Nicolson.

- C'est en quelque sorte la "moyenne" entre les deux schémas d'Euler (5) et (12) :

(20) $$\frac{1}{\Delta t}(u^{k+1} - u^k) - \frac{1}{2}\left[f(u^k) + f(u^{k+1})\right] = 0 .$$

De manière analogue aux deux autres schémas, on introduit une solution de l'équation différentielle (1), on remplace u^k par $u(t)$ et u^{k+1} par $u(t+\Delta t)$ dans l'expression (20) du schéma, et le résultat obtenu définit l'erreur de troncature :

(21) $$\mathcal{E}(\Delta t, t; u(\cdot)) = \frac{1}{\Delta t}\left(u(t+\Delta t) - u(t)\right) - \frac{1}{2}\left[f(u(t)) + f(u(t+\Delta t))\right] .$$

• le développement limité de l'erreur de troncature (21) du schéma de Crank-Nicolson s'obtient en rapportant les développements (9) et (14). Nous obtenons :

$$\mathcal{C} = \frac{du}{dt} + \frac{1}{2}\Delta t \frac{d^2u}{dt^2} + \frac{1}{6}\Delta t^2 \frac{d^3u}{dt^3} + O(\Delta t^3) - \frac{1}{2}\Big[f(u(t)) + f(u(t)) + \Delta t \frac{du}{dt} f'(u(t)) + \frac{\Delta t^2}{2}\frac{d^2u}{dt^2} f'(u(t)) + \frac{1}{2}\left(\Delta t \frac{du}{dt}\right)^2 f''(u(t)) + O(\Delta t^3)$$

$$(22)\quad \mathcal{C} = \begin{cases} \left[\frac{du}{dt} - f(u(t))\right] + \frac{\Delta t}{2}\left[\frac{d^2u}{dt^2} - \frac{du}{dt} f'(u(t))\right] \\ + \Delta t^2\left[\frac{1}{6}\frac{d^3u}{dt^3} - \frac{1}{4}\frac{d^2u}{dt^2} f'(u(t)) - \frac{1}{4}\left(\frac{du}{dt}\right)^2 f''(u(t))\right] \end{cases} + O(\Delta t^3)$$

• Nous constatons que le terme constant en Δt du développement (22) est nul car $u(\cdot)$ est solution de l'équation (1). Quand on dérive une fois cette relation, nous avons :

$$(23)\quad \frac{d^2u}{dt^2} = f'(u(t))\frac{du}{dt},$$

ce qui montre que le coefficient du terme en Δt dans le développement (22) est nul, donc que le schéma de Crank-Nicolson est au moins d'ordre deux. Par dérivation en temps de la relation (23), nous avons :

$$(24)\quad \frac{d^3u}{dt^3} = f''(u(t))\left(\frac{du}{dt}\right)^2 + f'(u(t))\frac{d^2u}{dt^2}$$

donc le coefficient de Δt^2 dans le développement (22) vaut $\left(\frac{1}{6}-\frac{1}{4}\right) d^3u/dt^3 = -\frac{1}{12}\frac{d^3u}{dt^3}$; il est en général non nul. Nous retenons

(25) $\mathcal{E}(\Delta t, t; u(\cdot)) = -\frac{1}{12}\frac{d^3u}{dt^3}\Delta t^2 + o(\Delta t^3)$, Crank-Nicolson est l'ordre de troncature du schéma de Crank-Nicolson tend vers zéro comme $O(\Delta t^2)$. On dit pour cette raison qu'il est d'ordre deux.

5) Schéma de Heun

- on rappelle qu'il prend la forme

(26) $\frac{1}{\Delta t}(u^{k+1}-u^k) - \frac{1}{2}\left[f(u^k) + f(u^k + \Delta t f(u^k))\right] = 0$.

L'ordre de troncature se définit comme dans les cas précédents

(27) $\mathcal{E}(\Delta t, t; u(\cdot)) = \begin{cases} \frac{1}{\Delta t}(u(t+\Delta t) - u(t)) - \frac{1}{2} f(u(t)) \\ -\frac{1}{2} f(u(t) + \Delta t f(u(t))) . \end{cases}$

- Pour déterminer l'ordre de ce schéma, on développe l'erreur de troncature (27), sans oublier les relations (1), (23) et (24) qui lui sont dérivées. De manière analogue au lemme 1, on a la relation (15), à appliquer avec $v = \Delta t f(u(t))$ cette fois, pour lequel nous avons $\frac{v^2}{2} = \frac{1}{2}(\Delta t)^2 (f(u(t)))^2$, $O(v^3) = O(\Delta t^3)$. Il vient

(28) $f(u(t)+\Delta t f(u(t))) = f(u(t)) + \Delta t\, f(u(t))\, f'(u(t)) + \frac{1}{2}\Delta t^2 f^2(u(t)) f''(u(t)) + O(\Delta t^3)$

on reporte cette expression au second membre de la relation (27), et on tient compte du développement (9). Il vient

$$\mathcal{C} = \frac{du}{dt} + \frac{\Delta t}{2}\frac{d^2u}{dt^2} + \frac{\Delta t^2}{6}\frac{d^3u}{dt^3} + O(\Delta t^3) - f(u(t)) - \frac{1}{2}\left[\Delta t\, f(u(t))\, f'(u(t)) + \frac{\Delta t^2}{2} f^2(u(t)) f''(u(t))\right] + O(\Delta t^3,$$

(29)
$$\mathcal{C} = \begin{cases} \dfrac{du}{dt} - f(u(t)) + \dfrac{\Delta t}{2}\left[\dfrac{d^2u}{dt^2} - f(u(t)) f'(u(t))\right] \\ + \Delta t^2\left(\dfrac{1}{6}\dfrac{d^3u}{dt^3} - \dfrac{1}{4} f^2(u(t)) f''(u(t))\right) + O(\Delta t^3). \end{cases}$$

- Le terme en Δt^0 dans le développement (29) est identiquement nul compte tenu de la relation (1). Le terme en Δt l'est également, compte tenu de (23) et de (1). On a aussi suite à (24) :

(30) $$\frac{d^2u}{dt^3} = f''(u(t))\left(f(u(t))\right)^2 + \left(f'(u(t))\right)^2 f(u(t))$$

donc

(31) $$\mathcal{C}(\Delta t, t; u(\cdot)) = \left[-\frac{1}{12} f''(u(t)) f^2(u(t)) + \frac{1}{6}\left(f'(u(t))\right)^2 f(u(t))\right]\Delta t^2 + O(\Delta t^3)$$

et le coefficient du terme d'ordre deux dans le développement (31) est en général non nul. Le schéma de Heun est d'ordre deux.

6) Définition générale

- Dans le cas d'un schéma général qui s'écrit par exemple sous la forme

$$(32)\quad \frac{1}{\Delta t}(u^{k+1}-u^k)-\Phi(u^k,u^{k+1},f(u^k),f(u^{k+1}))=0,$$

nous définissons l'erreur de troncature $\mathcal{T}(\Delta t,t;u(\cdot))$ par

$$(33)\quad \mathcal{T}(\Delta t,t;u(\cdot))=\begin{cases}\frac{1}{\Delta t}\big(u(t+\Delta t)-u(t)\big)\\ -\Phi\big(u(t),u(t+\Delta t),f(u(t)),f(u(t+\Delta t))\big)\end{cases}$$

pour une solution $u(\cdot)$ de l'équation (1).

- on dit que le schéma (32) est <u>consistant</u> avec l'équation (1) lorsque l'erreur de troncature $\mathcal{T}$ définie en (33) <u>tend vers zéro</u> si $\Delta t\to 0$. On peut vérifier sans peine que c'est le cas si et seulement si

$$(34)\quad \Phi(u(t),u(t),f(u(t)),f(u(t)))=f(u(t)),\ \forall u(t).$$

- on dit que le schéma est <u>d'ordre p</u> si l'erreur de troncature (33) admet le développement

$$(35)\quad \mathcal{T}(\Delta t,t,u(\cdot))=O(\Delta t^p)$$

lorsque Δt tend vers zéro.

7) Erreur et erreur de troncature

- Nous illustrons sur un exemple un résultat général qui énonce qu'un schéma stable et consistant est alors convergent. Nous fixons $\lambda > 0$ et étudions l'équation modèle

(36) $$\frac{du}{dt} + \lambda u = 0, \quad t > 0.$$

Nous la discrétisons avec un schéma d'Euler explicite

(37) $$\frac{1}{\Delta t}\left(u^{k+1} - u^k\right) + \lambda u^k = 0.$$

L'erreur de troncature est définie par

(38) $$\mathcal{T} = \frac{1}{\Delta t}\left(u((k+1)\Delta t) - u(k\Delta t)\right) + \lambda u(k\Delta t)$$

et on a vu à la relation (10) qu'on a :

(39) $$\mathcal{T}(\Delta t, t^k; u(\cdot)) = \frac{\Delta t}{2}\frac{d^2u}{dt^2}(k\Delta t) + O(\Delta t^2).$$

- Comme à la relation (3), nous introduisons l'erreur $\varepsilon_{\Delta t}^k$ par :

(40) $$\varepsilon_{\Delta t}^k = u_{\Delta t}^k - u(k\Delta t)$$

Par soustraction de (38) de (37), on a :

(41) $$\frac{1}{\Delta t}\left(\varepsilon_{\Delta t}^{k+1} - \varepsilon_{\Delta t}^k\right) + \lambda\, \varepsilon_{\Delta t}^k = -\mathcal{T}(\Delta t, k\Delta t; u(\cdot))$$

L'erreur vérifie une équation analogue à celle du schéma numérique, avec comme source l'erreur de troncature.

- on suppose $0 \leq k\Delta t \leq T$, avec T fixé. on peut donc majorer uniformément la dérivée seconde $\frac{d^2u}{dt^2}$ sur cet intervalle :

(42) $\left| \frac{d^2u}{dt^2}(k\Delta t) \right| \leq C$, $\forall k \in \mathbb{N}$ tq $k\Delta t \leq T$.

on tire alors de (39) et (42)

(43) $|G(\Delta t, k\Delta t; u(\cdot))| \leq C\Delta t$, $k\Delta t \leq T$.

- on suppose de plus le schéma d'Euler stable, ie

(44) $0 < \lambda\Delta t \leq 1$.

on peut alors écrire la relation (41) sous la forme

(45) $\varepsilon_{\Delta t}^{k+1} = (1-\lambda\Delta t)\,\varepsilon_{\Delta t}^{k} - \Delta t\, G(\Delta t, k\Delta t; u)$.

on tire alors de (44). $0 < 1-\lambda\Delta t < 1$
et en prenant les valeurs absolues de part et d'autre de (45):

$$|\varepsilon_{\Delta t}^{k+1}| \leq |\varepsilon_{\Delta t}^{k}| + \Delta t\, |G(\Delta t, k\Delta t; u(\cdot))|$$

donc compte tenu de (43):

(46) $|\varepsilon_{\Delta t}^{k+1}| \leq |\varepsilon_{\Delta t}^{k}| + C\Delta t^2$, $k\Delta t \leq T$.

- on écrit la chaîne d'inégalités (46), depuis $\varepsilon_{\Delta t}^0 \equiv 0$ jusqu'à $\varepsilon_{\Delta t}^k$. Il vient

$$|\varepsilon_{\Delta t}^k| \leq |\varepsilon_{\Delta t}^{k-1}| + C\Delta t^2$$
$$|\varepsilon_{\Delta t}^{k-1}| \leq |\varepsilon_{\Delta t}^{k-2}| + C\Delta t^2$$
$$\vdots$$
$$|\varepsilon_{\Delta t}^1| \leq |\varepsilon_{\Delta t}^0| + C\Delta t^2 .$$

Puisons ajoute toutes ces inégalités, ce qui est possible grâce à la stabilité (44). On en déduit, puisque $k\Delta t \leq T$:

$$(47) \quad |\varepsilon_{\Delta t}^k| \leq Ck\Delta t^2 \leq CT\Delta t \quad , \quad k\Delta t \leq T.$$

Nous venons d'établir que l'erreur est majorée par une constante multipliée par Δt; elle est d'ordre un en Δt, comme l'erreur de troncature. Si nous introduisons l'erreur $\delta(\Delta t)$ comme en (4), on tire de la relation (47) l'estimation

$$(48) \qquad \delta(\Delta t) \leq CT\Delta t ,$$

ce qui confirme bien le résultat établi.

D, mars 2003.