Schéma d'Euler explicite

Définition 1 *Étant donnés un pas de temps Δt et une suite d'instants $(t^n = t^0 + n\Delta t)_{n\in\mathbb{N}}$, le* schéma d'Euler explicite *associé à l'équation différentielle*

$$\frac{\mathrm{d}u}{\mathrm{d}t} \;=\; f(t,u)\,, \tag{1}$$

où f est une fonction continue de $\mathbb{R}^+ \times U$ dans $\mathbb{R}^d$, est donné par la relation de récurrence

$$v^{n+1} \;=\; v^n + \Delta t\, f(t^n, v^n)\,. \tag{2}$$

Une « solution » de ce schéma est un N-uplet $(v^n)_{n\in\{0,\dots,N\}}$ de vecteurs de U vérifiant (2) *pour tout $n \in \{0,\dots,N-1\}$.*

Ce schéma approche $u(t^{n+1})$ par

$$u(t^n) \;+\; \Delta t\;\, f(t^n, u(t^n))\,,$$

ce qui revient à remplacer f par la fonction constante égale à $f(t^n, u(t^n))$ sur l'intervalle $[t^n, t^{n+1}]$ dans la formule intégrale

$$u(t^{n+1}) \;=\; u(t^n) \;+\; \int_{t^n}^{t^{n+1}} f(s, u(s))\,\mathrm{d}s\,.$$

Théorème 1 *Si $f : \mathbb{R}^+ \times U \to \mathbb{R}^d$ est continue, quel que soient $t_0 \in \mathbb{R}^+$, $u_0 \in U$, il existe $T > t_0$ tel que, si pour $0 < \Delta t < T - t_0$ et $N\Delta \le T - t_0$, la donnée initiale $v^0 = u_0$ et le schéma* (2) *définissent un unique N-uplet $(v^n)_{n\in\{0,\dots,N\}}$ de vecteurs de U. Soit alors la fonction $v_{\Delta t} : [t_0, T] \to U$, continue affine par morceaux valant v^n en t^n, c'est-à-dire que*

$$v_{\Delta t}(t) \;=\; v^n \;+\; (v^{n+1} - v^n)\,\frac{t - t^n}{\Delta t}\,, \quad \textit{si}\;\; t \in [t^n, t^{n+1}[\,, \quad n \in \{0,\dots,N-1\}\,.$$

Il existe une suite $(h_k)_{k\in\mathbb{N}}$, strictement décroissante et tendant vers 0 telle que v_{h_k} tend vers une fonction $u \in \mathscr{C}(t_0, T; U)$ solution de l'équation intégrale

$$u(t) \;=\; u_0 \;+\; \int_{t_0}^{T} f(s, u(s))\,\mathrm{d}s\,,$$

ce qui implique que u est solution de (1) *pour la donnée initiale $u(0) = u_0$.*

Ce théorème de convergence à sous-suite près du schéma d'Euler démontre « au passage » un théorème d'existence de solutions pour l'équation différentielle (1) en supposant seulement f continue : ce résultat est en général attribué à Cauchy, Peano et Arzelà (en ne retenant parfois que deux noms sur les trois). Sa démonstration repose sur un théorème d'analyse fonctionnelle, attribué à Arzelà et Ascoli, souvent appelé simplement théorème d'Ascoli.

Théorème 2 (Ascoli) *On considère l'espace $\mathscr{C}(K; R^d)$ des fonctions continues sur un intervalle compact K et à valeurs dans $\mathbb{R}^d$, muni de la norme sup :*

$$\|v\|_\infty \;:=\; \max_{t\in K} \|v(t)\|_{\mathbb{R}^d}\,.$$

Si une suite $(v_k)_{k\in\mathbb{N}}$ est bornée *dans $\mathscr{C}(K; R^d)$ et* équicontinue*, c'est-à-dire que pour tout $\varepsilon > 0$, il existe $\delta > 0$ tel que pour t, $s \in K$, si $|t - s| \le \delta$ alors $\|v_k(t) - v_k(s)\|_{\mathbb{R}^d} \;\le\; \varepsilon$ quel que soit $k \in \mathbb{N}$.*

Schémas explicites

Définition 2 *Étant donnés un pas de temps Δt et une suite d'instants $(t^n = t^0 + n\Delta t)_{n\in\mathbb{N}}$, un* schéma explicite *à un pas est une relation de récurrence de la forme*

$$v^{n+1} \;=\; v^n + \Delta t\, \phi(t^n, v^n, \Delta t)\,, \tag{3}$$

où ϕ est une fonction de $\mathbb{R}^+ \times U \times \mathbb{R}^+$ (avec U un ouvert de $\mathbb{R}^d$) dans $\mathbb{R}^d$.
Une « solution » du schéma est un N-uplet $(v^n)_{n\in\{0,\dots,N\}}$ de vecteurs de U vérifiant (3) *pour tout $n \in \{0,\dots,N-1\}$.*

Exemples :

- *Schéma d'Euler* : $\phi(t,w,\Delta t) = f(t,w)$ comme on l'a déjà vu.
- *Schéma de Runge* : $\phi(t,w,\Delta t) = f(t+\frac{1}{2}\Delta t, w+\frac{1}{2}\Delta t f(t,w))$. Ce schéma provient de l'approximation de l'intégrale

$$\int_{t^n}^{t^{n+1}} f(s,u(s))\,\mathrm{d}s$$

par la « formule des trapèzes », c'est-à-dire par

$$\Delta t\;\, f(t^{n+1/2}, u(t^{n+1/2}))$$

où $t^{n+1/2} = t^n + \frac{\Delta t}{2}$, en approchant $u(t^{n+1/2})$ par le schéma d'Euler entre t^n et $t^{n+1/2}$.

Définition 3 *Le schéma* (3) *est dit* consistant *avec l'équation différentielle* (1) *si pour toute solution $u \in \mathscr{C}^1([t_0,T];U)$ de* (1)*, pour tout $t \in [t_0,T]$, l'*erreur de consistance *:*

$$\mathcal{R}(t,u,\Delta t) \;:=\; \frac{u(t+\Delta t)-u(t)}{\Delta t} \;-\; \phi(t,u(t),\Delta t)$$

tend vers 0 lorsque Δt tend vers 0, uniformément pour $t \in [t_0,T]$. Le schéma (3) *est dit* consistant à l'ordre p *($p \in \mathbb{N}^*$), ou simplement d'*ordre p*, si de plus (en supposant u de classe $\mathscr{C}^{p+1}$) il existe $C > 0$, indépendant de Δt et de $t \in [t_0,T]$ tel que*

$$\|\mathcal{R}(t,u,\Delta t)\| \;\le\; C\,(\Delta t)^p\,.$$

Proposition 1 *Si ϕ est continue, le schéma* (3) *est consistant avec* (1) *si et seulement si $\phi(t,w,0) = f(t,w)$ quel que soit $(t,w) \in \mathbb{R}^+ \times U$.*

Si f est de classe $\mathscr{C}^p$, on définit $f^{[0]} := f$ et par récurrence, pour tout $k \in \{0,\dots,p-1\}$,

$$\begin{aligned} f^{[k+1]} :\; \mathbb{R}^+ \times U \;&\to\; \mathbb{R}^d \\ (t,w) \;&\mapsto\; f^{[k+1]}(t,w) \;:=\; \frac{\partial f^{[k]}}{\partial t}(t,w) \;+\; \sum_{j=1}^{d} \frac{\partial f^{[k]}}{\partial w_j}(t,w)\, f_j(t,w)\,, \end{aligned}$$

et si ϕ est de classe $\mathscr{C}^p$, alors le schéma (3) *est consistant avec* (1) *à l'ordre p si et seulement si pour tout $k \in \{1,\dots,p-1\}$,*

$$\frac{\partial^k \phi}{\partial(\Delta t)^k}(t,w,0) \;=\; \frac{1}{k+1}\, f^{[k]}(t,w)$$

quel que soit $(t,w) \in \mathbb{R}^+ \times U$.

Exemples :

- Le schéma d'Euler est d'ordre 1. En effet, pour $\phi(t,w,\Delta t) = f(t,w)$ on a évidemment $\phi(t,w,0) = f(t,w)$, mais $\frac{\partial\phi}{\partial(\Delta t)}(t,w,0) = 0$, ce qui n'est pas égal à $f^1(t,w)$ si la fonction f^1 n'est pas nulle. Donc en général, le schéma d'Euler n'est pas d'ordre 2.
- Le schéma de Runge est d'ordre 2. En effet, pour $\phi(t,w,\Delta t) = f(t+\frac{1}{2}\Delta t, w+\frac{1}{2}\Delta t f(t,w))$, $\phi(t,w,0) = f(t,w)$,

$$\frac{\partial\phi}{\partial(\Delta t)}(t,w,0) \;=\; \frac{1}{2}\left(\frac{\partial f}{\partial t}(t,w) \;+\; \sum_{j=1}^{d} \frac{\partial f}{\partial w_j}(t,w)\, f_j(t,w)\right) \;=\; \frac{1}{2}\, f^{[1]}(t,w)\,.$$

Donc le schéma de Runge est au moins d'ordre 1. Plus généralement,

$$\begin{aligned}\frac{\partial\phi}{\partial(\Delta t)}(t,w,\Delta t) \;=\;& \frac{1}{2}\left(\frac{\partial f}{\partial t}(t+\tfrac{1}{2}\Delta t, w+\tfrac{1}{2}\Delta t f(t,w))\right.\\ &\left. +\; \textstyle\sum_{j=1}^{d} \dfrac{\partial f}{\partial w_j}(t+\tfrac{1}{2}\Delta t, w+\tfrac{1}{2}\Delta t f(t,w))\, f_j(t,w)\right) \;=\end{aligned}$$

$$\begin{aligned}&\frac{1}{2}\, f^{[1]}(t+\tfrac{1}{2}\Delta t, w+\tfrac{1}{2}\Delta t f(t,w)) \;+\\ &\frac{1}{2}\, \textstyle\sum_{j=1}^{d} \dfrac{\partial f}{\partial w_j}(t+\tfrac{1}{2}\Delta t, w+\tfrac{1}{2}\Delta t f(t,w))\,(f_j(t,w) \;-\; f_j(t+\tfrac{1}{2}\Delta t, w+\tfrac{1}{2}\Delta t f(t,w))\,,\end{aligned}$$

d'où

$$\begin{aligned}\frac{\partial^2\phi}{\partial(\Delta t)^2}(t,w,0) \;&=\; \frac{1}{4}\left(f^{[2]}(t,w) \;-\; \textstyle\sum_{j=1}^{d} \dfrac{\partial f}{\partial w_j}(t,w)\, f_j^{[1]}(t,w)\right)\\ &=\; \frac{1}{4}\left(\frac{\partial f^{[1]}}{\partial t}(t,w) \;+\; \textstyle\sum_{j=1}^{d}\Big(\dfrac{\partial f^{[1]}}{\partial w_j}(t,w)\, f_j(t,w) \;-\; \dfrac{\partial f}{\partial w_j}(t,w)\, f_j^{[1]}(t,w)\Big)\right)\,,\end{aligned}$$

ce qui n'est pas égal à $f^{[2]}(t,w)$ en général. On voit cependant sur cet exemple que la vérification de l'ordre d'un schéma devient rapidement très technique.

Dém. [Proposition 1] Pour $u \in \mathscr{C}^1([t_0,T];U)$, pour tout $t \in [t_0,T]$,

$$\frac{u(t+\Delta t)-u(t)}{\Delta t} \;=\; \int_0^1 u'(t+\theta\Delta t)\,\mathrm{d}\theta$$

tend vers $u'(t)$ lorsque Δt tend vers 0, uniformément par rapport à t car u' est uniformément continue sur le compact $[t_0,T]$. Par suite, si u est solution de (1), l'erreur de consistance $\mathscr{R}(t,u,\Delta t)$ tend vers $f(t,u(t)) - \phi(t,u(t),0)$ (uniformément pour $t \in [0,T]$ puisque ϕ est uniformément continue sur les compacts de la forme $[0,T]\times u([0,T])\times[0,(\Delta t)_*]$). Par définition, le schéma (3) est consistant avec (1) si et seulement si cette limite est nulle quelle que soit la solution u. Ceci est par conséquent équivalent à $f(t,w) = \phi(t,w,0)$ quel que soit $(t,w) \in \mathbb{R}^+ \times U$.

Maintenant si $u \in \mathscr{C}^{p+1}([t_0,T];U)$, pour tout $t \in [t_0,T]$,

$$\frac{u(t+\Delta t)-u(t)}{\Delta t} \;=\; \sum_{k=1}^{p} \frac{1}{k!}\,(\Delta t)^{k-1}\, u^{(k)}(t) \;+\; \frac{1}{p!}\,(\Delta t)^p \int_0^1 (1-\theta)^p \;u^{(p+1)}(t+\theta\Delta t)\,\mathrm{d}\theta$$

Si de plus u est solution de (1), on vérifie par une récurrence immédiate que pour tout $k \in \{1,\dots,p+1\}$,

$$u^{(k)}(t) \;=\; f^{[k-1]}(t,u(t))\,.$$

Par ailleurs, si ϕ est de classe $\mathscr{C}^p$,

$$\phi(t,u(t),\Delta t) \;=\; \sum_{k=0}^{p-1} \frac{1}{k!}\,(\Delta t)^k\,\frac{\partial^k \phi}{\partial(\Delta t)^k}(t,u(t),0) + \frac{1}{(p-1)!}\,(\Delta t)^p \int_0^1 (1-\theta)^{p-1}\,\frac{\partial^p \phi}{\partial(\Delta t)^p}(t,u(t),0)\;\mathrm{d}\theta$$

donc l'erreur de consistance s'écrit

$$\mathscr{R}(t,u,\Delta t) \;=\; \sum_{k=0}^{p-1} \frac{1}{k!}\,(\Delta t)^k \left(\frac{1}{k+1}\,f^{[k]}(t,u(t)) \;-\; \frac{\partial^k \phi}{\partial(\Delta t)^k}(t,u(t),0)\right) +$$

$$\frac{1}{(p-1)!}\,(\Delta t)^p \int_0^1 (1-\theta)^{p-1} \left(\frac{1-\theta}{p}\,f^{[p]}(t,u(t)) \;-\; \frac{\partial^p \phi}{\partial(\Delta t)^p}(t,u(t),0)\right) \mathrm{d}\theta\,.$$

Si tous les termes de la somme sont nuls, alors il existe bien $C>0$ tel que $\|\mathcal{R}(t,u,\Delta t)\| \;\leq\; C\,(\Delta t)^p$. Inversement, si l'un des termes de la somme (autre que le premier) est non nul, soit $k \in \{1,\ldots,p-1\}$ le plus petit entier tel que

$$\frac{\partial^k \phi}{\partial(\Delta t)^k}(t,u(t),0) \;\neq\; \frac{1}{k+1}\,f^{[k]}(t,u(t))\,.$$

Alors il existe $(\Delta t)_*>0$ et $\gamma>0$ tels que pour $0<\Delta t<(\Delta t)_*$, $\|\mathcal{R}(t,u,\Delta t)\| \;\geq\; \gamma\,(\Delta t)^k$, ce qui interdit une majoration du type $\|\mathcal{R}(t,u,\Delta t)\| \;\leq\; C\,(\Delta t)^p$. □

Définition 4 *Le schéma* (3) *est dit* stable par rapport aux erreurs *sur l'intervalle* $[t_0,T]$ *s'il existe* $(\Delta t)_*>0$*, et* $C>0$ *tels que, pour* $N\in\mathbb{N}$ *tel que* $N \geq (T-t_0)/(\Delta t)_*$, $\Delta t = (T-t_0)/N$, $(\varepsilon^n)_{n\in\{0,\ldots,N\}} \in \mathbb{R}^{(N+1)d}$*, si* $(v^n)_{n\in\{0,\ldots,N\}}$ *est solution du schéma* (3) *et* $(u^n)_{n\in\{0,\ldots,N\}}$ *est solution du schéma perturbé*

$$u^{n+1} \;=\; u^n + \Delta t\,\phi(t^n,u^n,\Delta t) \;+\; \varepsilon^n, \tag{4}$$

alors pour tout $n\in\{0,\ldots,N\}$,

$$\|u^n - v^n\| \leq C\left(\|u^0-v^0\| \;+\; \sum_{m=0}^{n-1} \|\varepsilon^m\|\right).$$

Proposition 2 *S'il existe* $(\Delta t)_*>0$ *et* $\Gamma>0$ *tels que, pour tout* $t\in\mathbb{R}^+$*, pour* $u,\,v\in U$*,* $\Delta t<(\Delta t)_*$,

$$\|\phi(t,u,\Delta t)-\phi(t,v,\Delta t)\| \;\leq\; \Gamma\,\|u-v\|\,,$$

alors le schéma (3) *est stable par rapport aux erreurs sur tout intervalle* $[t_0,T]$.

Dém. Si $(v^n)_{n\in\{0,\ldots,N\}}$ vérifie (3) et $(u^n)_{n\in\{0,\ldots,N\}}$ vérifie (4), alors pour tout $n\in\{0,\ldots,N\}$,

$$u^{n+1}-v^{n+1} \;=\; u^n - v^n \;+\; \Delta t\,(\phi(t^n,u^n,\Delta t) \;-\; \phi(t^n,v^n,\Delta t)) \;+\; \varepsilon^n\,,$$

d'où

$$\|u^{n+1}-v^{n+1}\| \;\leq\; (1+\Gamma\Delta t)\,\|u^n-v^n\| \;+\; \|\varepsilon^n\|\,.$$

On en déduit par une récurrence facile que

$$\|u^n-v^n\| \;\leq\; (1+\Gamma\Delta t)^n\,\|u^0-v^0\| \;+\; \sum_{m=0}^{n-1}(1+\Gamma\Delta t)^{n-m}\|\varepsilon^m\|\,.$$

(Ce résultat est parfois énoncé sous l'appellation *lemme de Gronwall discret*, par analogie avec le résultat analogue pour les fonctions : si $w'(t) \leq cw(t) + \varepsilon(t)$ quel que soit t alors $w(t) \leq \mathrm{e}^{ct}w(0) + \int_0^t \mathrm{e}^{c(t-\tau)}\varepsilon(\tau)\mathrm{d}\tau$, voir l'appendice.) Or pour tout réel x, $1 + x \leq \mathrm{e}^x$. Donc

$$\|u^n - v^n\| \leq \mathrm{e}^{n\Gamma\Delta t}\left(\|u^0 - v^0\| + \sum_{m=0}^{n-1}\|\varepsilon^m\|\right),$$

et l'on a la majoration cherchée pour $C = \mathrm{e}^{\Gamma(T-t_0)}$. □

L'hypothèse de la proposition 2 est très forte, puisqu'elle demande à la fonction ϕ d'être globalement Lipschitienne par rapport à w : ceci semble par exemple exclure la fonction ϕ : $(t, w) \mapsto w^2$, correspondant au schéma d'Euler pour l'équation de *Riccati* $u' = u^2$. On peut énoncer un résultat plus faible, avec une limitation sur l'intervalle de résolution analogue à celle que l'on connaît pour les équations différentielles. (Pour l'équation de Riccati par exemple, les solutions « explosent » en temps fini : une solution ne s'annulant pas vérifie

$$\frac{u'(t)}{u(t)^2} = 1$$

d'où

$$-\frac{1}{u(t)} + \frac{1}{u(t_0)} = t - t_0$$

et donc $\lim_{t\to t_0+1/u(t_0)} |u(t)| = +\infty$.)

Proposition 3 *Supposons ϕ de classe $\mathscr{C}^1$. Quels que soient $t_0 \in \mathbb{R}^+$, $u^0 \in U$ et $v^0 \in U$ assez proches, il existe $(\Delta t)_* > 0$, $T > 0$, $\rho > 0$ et $C > 0$ tels que pour tout $N \in \mathbb{N}$ tel que $N \geq (T-t_0)/(\Delta t)_*$, pour $\Delta t = (T-t_0)/N$, si $(v^n)_{n\in\{0,\ldots,N\}}$ est solution de* (3) *et $(u^n)_{n\in\{0,\ldots,N\}}$ est solution de* (4) *avec $\sum_{n=0}^{N-1}\|\varepsilon^n\| \leq \rho$ pour tout n, alors pour tout $n \in \{0,\ldots,N\}$,*

$$\|u^n - v^n\| \leq C\left(\|u^0 - v^0\| + \sum_{m=0}^{n-1}\|\varepsilon^m\|\right).$$

Dém. Soient $u^0, v^0 \in U$: s'ils sont assez proches l'un de l'autre il existe une boule fermée B, de rayon $R > 0$, dans $\mathbb{R}^d$, incluse dans U et les contenant tous les deux. On notera B_3 la boule de même centre et de rayon $3R$. Soient $t_0 \in \mathbb{R}^+$, $T_* > t_0$ et $(\Delta t)_* \in]0, T_* - t_0[$. Alors sur le compact $[t_0, T_*] \times B_3 \times [0, (\Delta t)_*]$, la fonction ϕ est bornée, disons par $M > 0$, et uniformément Lipschitzienne par rapport à w.

Si l'on choisit $T \in]0, T_*]$ tel que $(T - t_0)\,M \leq R$, alors toute solution $(v^n)_{n\in\{0,\ldots,N\}}$ de (3) avec $N \geq (T - t_0)/(\Delta t)_*$ et $\Delta t = (T - t_0)/N$ vérifie $\|v^n - v^0\| \leq (n\Delta t)\,M \leq R$ pour tout $n \in \{0,\ldots,N\}$. (Ceci résulte d'une récurrence facile : c'est vrai pour $n = 0$; si c'est vrai pour n, $\|v^{n+1} - v^n\| \leq \Delta t M$ d'où par l'inégalité triangulaire, $\|v^{n+1} - v^0\| \leq (n + 1)\Delta t\,M$.) De la même manière, si $(u^n)_{n\in\{0,\ldots,N\}}$ est solution de (4) avec $\sum_{n=0}^{N-1}\|\varepsilon^n\| \leq R$, $\|u^n - u^0\| \leq (n\Delta t)\,M + \sum_{m=0}^{n-1}\|\varepsilon^m\| \leq 2R$.

Ainsi $(v^n)_{n\in\{0,\ldots,N\}}$ et $(u^n)_{n\in\{0,\ldots,N\}}$ restent dans le domaine de validité de la constante de Lipschitz de ϕ, et on peut donc faire le même calcul que dans la proposition 2. □

Dans ce qui suit on suppose la fonction f de classe $\mathscr{C}^1$, de sorte que le théorème de Cauchy-Lipschitz s'applique à l'équation différentielle (1).

Définition 5 *Étant donnée une solution $u \in \mathscr{C}^1([t_0,T];U)$ de* (1)*, on appelle* erreur globale *du schéma* (3) *:*

$$\mathscr{E}(\Delta t) \;:=\; \max_{0\le n\le T/(\Delta t)} \|v^n - u(t^n)\|\,.$$

Le schéma (3) *est dit* convergent *vers u sur l'intervalle $[t_0,T]$ si l'*erreur globale *tend vers 0 lorsque Δt tend vers 0.*

Théorème 3 *On suppose l'équation différentielle* (1) *admet une solution $u \in \mathscr{C}^1([t_0,T];U)$. On considère le schéma* (3) *avec la donnée initiale $u(0)=v^0$. S'il est stable par rapport aux erreurs sur $[t_0,T]$ et consistant avec l'équation différentielle* (1)*, alors il est convergent vers la solution exacte u. En outre il existe $C>0$ tel que*

$$\mathscr{E}(\Delta t) \;\le\; C\,(\Delta t)^p\,.$$

Dém. Par définition de l'erreur de consistance,

$$u(t^{n+1}) = u(t^n) \;+\; \Delta t\,\phi(t^n,u(t^n),\Delta) \;+\; \Delta t\,\mathscr{R}(t^n,u,\Delta t)\,.$$

Autrement dit $(u(t^n))$ est solution de (4) avec $\varepsilon^n = \Delta t\,\mathscr{R}(t^n,u,\Delta t)$. Donc d'après la stabilité du schéma,

$$\|u(t^n)-v^n\| \;\le\; C_0 \sum_{m=0}^{n-1} \|\mathscr{R}(t^m,u,\Delta t)\| \;\le\; C_0\,(n\Delta t)\,\max_m \|\mathscr{R}(t^m,u,\Delta t)\|\,.$$

Comme $n\Delta t \le T-t_0$, ceci tend vers 0 puisque l'erreur de consistance tend vers 0. Si de plus on a une majoration de l'erreur $\|\mathcal{R}(t,w,\Delta t)\| \;\le\; C\,(\Delta t)^p$, on en déduit

$$\|u(t^n)-v^n\| \;\le\; C_0\,(T-t_0)\,C\,(\Delta t)^p\,.$$

□

La question de la stabilité des schémas est en fait plus délicate qu'il n'y paraît, car la constante C de la définition 4 peut être *colossale*, de sorte qu'une petite erreur ($\|u^0-v^0\|$) sur la donnée initiale (dûe par exemple à l'imprécision d'une mesure physique, ou à l'arrondi dans l'ordinateur) entraîne une erreur démesurée sur la solution approchée. Ce problème de stabilité se pose également pour les solutions exactes de l'équation différentielle sous-jacente.

Prenons simplement le cas d'une équation linéaire

$$\frac{\mathrm{d}u}{\mathrm{d}t} \;=\; \lambda\,u\,, \tag{5}$$

où λ est un paramètre réel. S'il est positif, l'équation (5) modélise par exemple la croissance d'une population dont le nombre d'individus à l'instant t est $u(t)$ et dont le taux de natalité (c'est-à-dire le nombre de naissances par unité de temps et nombre d'individus) est λ. (Par exemple en France ce taux est de l'ordre de $14‰$ par an.) On sait que les solutions (exactes) de (5) sont de la forme $u(t) = u_0\,\mathrm{e}^{\lambda t}$. Ceci implique qu'une erreur commise à l'instant initial est multipliée par $\mathrm{e}^{\lambda T}$ au bout d'un temps T. (La solution du problème de Cauchy pour la donnée initiale $u(0)=u_0+\varepsilon$ est donnée par $u(t)=u_0\,\mathrm{e}^{\lambda t}+\varepsilon\,\mathrm{e}^{\lambda t}$.) Pour $\lambda = 14‰$ par exemple, $\mathrm{e}^{10\,\lambda} \simeq 1.15027$ est une valeur « raisonnable », mais $\mathrm{e}^{1000\,\lambda} \simeq 1202604$! Si le modèle est pertinent sur

quelques dizaines d'années, il perd son sens sur un millénaire. La même critique s'applique au modèle discret (toujours pour la croissance de cette population) s'écrivant simplement

$$v^{n+1} \,=\, (1+\lambda)\, v^n\,, \tag{6}$$

et se résolvant en $v^n = (1+\lambda)^n v^0$. Pour $\lambda = 14‰$, on trouve $(1+\lambda)^{10} \simeq 1.14916$, tandis que $(1+\lambda)^{1000} \simeq 1091327$!

Noter que le modèle discret (6) n'est rien d'autre que le schéma d'Euler (explicite) appliqué à (5) avec $\Delta t = 1$. Un tel « pas » peut paraître grossier, mais cela dépend de la valeur de λ et de l'intervalle de temps auquel on s'intéresse. Pour $\lambda = 14‰$ à nouveau, l'erreur relative en un an (c'est-à-dire à $t = \Delta t = 1 = n$) entre la solution du modèle continu (5) et la solution du modèle discret (6) est (pour des données initiales identiques) $\mathrm{e}^\lambda/(1+\lambda) - 1 \simeq 10^{-4}$. Les erreurs s'accumulant, on trouve une erreur relative de 10^{-3} au bout de dix ans, 10^{-2} à un siècle, 10^{-1} après un millénaire, etc. Mais le principal problème reste le facteur $(1+\lambda)^n$, qui devient assez rapidement gigantesque lorsque n augmente, puisque $1+\lambda > 1$.

Considérons maintenant le cas où λ est strictement négatif. L'équation différentielle (5) modélise alors par exemple une population en déclin, sans naissances et avec un taux de mortalité égal à $\mu := -\lambda$. Évidemment les solutions de (5), toujours de la forme $u(t) = u_0\,\mathrm{e}^{\lambda t} = u_0\,\mathrm{e}^{-\mu t}$, tendent vers 0 lorsque t tend vers $+\infty$. De même, si μ est effectivement un taux de mortalité, par nature inférieur à 1 et même strictement en général, les solutions $v^n = (1-\mu)^n v^0$ tendent vers 0 lorsque n tend vers $+\infty$. En revanche, si l'on applique le schéma d'Euler à (5) avec un pas Δt trop grand, les solutions prétendûment approchées $v^n = (1-\mu\Delta t)^n v^0$ divergent : en effet, si $2 < \mu\Delta t$ alors $|1-\mu\Delta t| > 1$ et donc $\lim_{n\to+\infty} \|v^n\| = +\infty$ si $v^0 \neq 0$! Une façon de remédier à ce problème est de considérer le schéma d'Euler *implicite*.

Avant d'introduire le schéma d'Euler implicite (pour l'équation (1) dans un premier temps), rappelons que pour tout $x \in \mathbb{R}$,

$$\lim_{n\to+\infty} \left(1+\frac{x}{n}\right)^n = \mathrm{e}^x\,.$$

Cette formule est même une façon (la plus élémentaire, vue en principe en L1) de définir la fonction exponentielle. Ce n'est pas hasard si elle est attribuée à Euler : elle explique pourquoi le schéma d'Euler explicite

$$v^{n+1} = (1+\lambda\Delta t)\, v^n$$

approche la solution de (5) : à $T = N\Delta t > 0$ fixé,

$$v^N = \left(1+\lambda\frac{T}{N}\right)^N u_0 \;\to\; \mathrm{e}^{\lambda T}\, u_0$$

lorsque $N \to +\infty$ (et par conséquent $\Delta t \to 0$). Maintenant on remarque que l'on a aussi

$$\left(1-\lambda\frac{T}{N}\right)^{-N} \;\to\; \frac{1}{\mathrm{e}^{-\lambda T}} \,=\, \mathrm{e}^{\lambda T}$$

lorsque $N \to +\infty$. Autrement dit, si l'on définit $(w^n)_{n\in\{0,\dots,N\}}$ par $w^0 = u_0$ et

$$(1-\lambda\Delta t)\, w^{n+1} = w^n$$

pour tout n, c'est-à-dire qu'en fait $w^n = (1-\lambda\Delta t)^{-n}\, u_0$, on a encore $w^N \to \mathrm{e}^{\lambda T}\, u_0$ lorsque $N \to +\infty$ (et $\Delta t \to 0$). Le calcul de w^n est une autre façon d'approcher la solution exacte. L'avantage de w^n sur v^n est que, à $\Delta t > 0$ fixé aussi grand qu'on veut, $\lim_{n\to+\infty} \|w^n\| = 0$, puisque $|1-\lambda\Delta t)|^{-1} < 1$ (λ étant strictement négatif).

Schéma d'Euler implicite

Définition 6 *Étant donnés un pas de temps Δt et une suite d'instants $(t^n = t^0 + n\Delta t)_{n\in\mathbb{N}}$, le* schéma d'Euler implicite *ou* rétrograde *associé à l'équation différentielle* (1) *est donné par la relation de récurrence* implicite

$$w^{n+1} \;=\; w^n + \Delta t\, f(t^{n+1}, w^{n+1})\,. \tag{7}$$

Une solution de ce schéma est un N-uplet $(w^n)_{n\in\{0,\dots,N\}}$ de vecteurs de U vérifiant (7) *pour tout $n \in \{0,\dots,N-1\}$.*

Contrairement aux schémas explicites, il n'est pas évident a priori que ce schéma ait des solutions. Cependant, si f est de classe $\mathscr{C}^1$ cela découle du théorème de Banach-Picard (pour l'existence) et du théorème des fonctions implicites (pour la régularité utile ensuite à la stabilité). Soient en effet $t_0 \in \mathbb{R}^+$, $u_0 \in U$, $T > t_0$, $R > 0$, $M > 0$ et $L > 0$ tels que $(T-t_0)\,M \;\leq\; R$ et pour tous $w, z \in \overline{B}(u_0; R)$ (la boule fermée de centre u_0 et de rayon R), $\|f(t,w)\| \;\leq\; M$, $\|f(t,w) - f(t,z)\| \;\leq L\,\|w - z\|$. Si $L\Delta t < 1$ et $N\Delta t \leq T - t_0$ alors la condition initiale $w^0 = u_0$ et le schéma (7) définissent de façon unique un N-uplet $(w^n)_{n\in\{0,\dots,N\}}$ de vecteurs de U tels que pour tout $n \in \{0,\dots,N\}$, $w^n \in \overline{B}(u_0; n\Delta t M) \subset \overline{B}(u_0; R)$. La preuve se fait bien sûr par récurrence. L'initialisation est évidente. Si l'on suppose $w^k \in \overline{B}(u_0; k\Delta t M)$ construits pour $k \leq n \leq N-1$, alors l'application

$$\psi_n : z \mapsto w^n + \Delta t\, f(t^{n+1}, z)$$

laisse la boule fermée $\overline{B}(u_0; (n+1)\Delta t M)$ invariante (car $\|\psi_n(z) - u_0\| \leq \|w^n - u_0\| + \Delta t M$) et elle est contractante dans cette boule (car $\|\psi_n(z) - \psi_n(w)\| \leq L\Delta t \|z - w\|$ et $L\Delta t < 1$). Donc d'après le théorème de Banach-Picard elle admet un unique point fixe w^{n+1}.

La fonction

$$\begin{array}{rccl} \Psi : & \mathbb{R}^+ \times \mathbb{R}^+ \times U \times U & \to & \mathbb{R}^d \\ & (\Delta t, s, w, z) & \mapsto & \Psi(\Delta t, s, w, z) \;:=\; z \;-\; w \;-\; \Delta t\, f(s, z)\,. \end{array}$$

est de classe $\mathscr{C}^1$ comme f, et pour tout $(s, w, z) \in \times\mathbb{R}^+ \times U \times U$, $\mathrm{d}_z\Psi(0, s, w, z) \;=\; I_d$. Or $\mathrm{d}_z\Psi$ uniformément continue sur tout compact et l'ensemble des isomorphismes de $\mathbb{R}^d$ est ouvert. Donc il existe $(\Delta t)_* > 0$ tel que pour $0 \leq \Delta t \leq (\Delta t)_*$ et $(s, w, z) \in \times[t_0, T] \times \overline{B}(u_0; R) \times \overline{B}(u_0; R)$, $\mathrm{d}_z\Psi(\Delta t, s, w, z)$ est inversible. Ceci montre que le théorème des fonctions implicites s'applique à Ψ en chaque point de $[0, (\Delta t)_*] \times [t_0, T] \times \overline{B}(u_0; R) \times \overline{B}(u_0; R)$ et par conséquent que l'application $\Phi : (\Delta t, s, w) \mapsto z = \Phi(\Delta t, s, w)$ telle que $z \;=\; w + \Delta t\, f(s, z)$ est de classe $\mathscr{C}^1$. Ainsi, le schéma (7) se réécrit, au prix du calcul de Ψ, comme un schéma du type (3) avec

$$\phi(s, w, \Delta t) \;:=\; f(s + \Delta t, \Phi(\Delta t, s + \Delta t, w))\,.$$

Par composition, ϕ est aussi de classe $\mathscr{C}^1$.

On vérifie que le schéma d'Euler implicite est d'ordre 1 en remarquant que, par définition, $\Phi(0, s, w) \;=\; w$ quel que soient s et w, d'où $\phi(t, w, 0) \;=\; f(t, w)$.

θ-schéma

Définition 7 *Étant donnés $\theta \in [0,1]$, un pas de temps Δt et une suite d'instants $(t^n = t^0 + n\Delta t)_{n\in\mathbb{N}}$, le θ-schéma associé à l'équation différentielle* (1) *est défini par la relation de récurrence* implicite

$$w^{n+1} = w^n + \Delta t\,(\theta\, f(t^{n+1}, w^{n+1}) + (1-\theta)\, f(t^n, w^n))\,. \tag{8}$$

Une solution de ce schéma est un N-uplet $(w^n)_{n\in\{0,\dots,N\}}$ de vecteurs de U vérifiant (8) *pour tout $n \in \{0,\dots,N-1\}$.*

Autrement dit, le θ-schéma est obtenu comme combinaison convexe des schémas d'Euler explicite et implicite. Comme pour le schéma d'Euler implicite, on montre qu'il admet des solutions par application d'un théorème de point fixe.

Le cas particulier $\theta = 1/2$ correspond à ce que l'on appelle le schéma de *Crank-Nicholson*. On vérifie que ce schéma est d'ordre 2.

Schémas de Runge-Kutta

Si l'on veut « monter en ordre » il faut imaginer des schémas plus sophistiqués. C'est le cas des schémas dits de *Runge-Kutta*. Ils généralisent le schéma de Runge, en raffinant l'approximation de l'intégrale

$$\int_{t^n}^{t^{n+1}} f(s,u(s))\,\mathrm{d}s\,.$$

Étant donnés des instants intermédiaires $t_i^n = t^n + c_i\Delta t$ pour $i \in \{1,\dots,q\}$, avec $c_i \in [0,1]$, et des coefficients b_j positifs ou nuls dont la somme vaut 1, on peut approcher l'intégrale ci-dessus par la combinaison convexe

$$\Delta t \ \sum_{j=1}^{q} b_j\, f(t_j^n, u(t_j^n))\,,$$

et $u(t_i^n)$ par une expression analogue entre t^n et t_i^n :

$$u(t_i^n) \simeq u(t^n) \ + \ \Delta t \sum_{j=1}^{q} a_{i,j}\, f(t_j^n, u(t_j^n))\,,$$

où les coefficients $a_{i,j}$ sont positifs ou nuls et tels que $\sum_{i=1}^{q} a_{i,j} \ = \ c_j$.

Définition 8 *Étant donnés un pas de temps Δt, une suite d'instants $(t^n = t^0 + n\Delta t)_{n\in\mathbb{N}}$, $c_i \in [0,1]$, $b_j \in [0,1]$ tels que $\sum_j b_j = 1$, $a_{i,j} \in [0,1]$ tels que $\sum_{j=1}^{q} a_{i,j} \ = \ c_i$, on définit un schéma de* Runge-Kutta à q étages *par $t_i^n = t^n + c_i\Delta t$ pour $i \in \{1,\dots,q\}$ et*

$$\text{(9)} \qquad \left\{ \begin{array}{l} v_i^n \ = \ v^n \ + \ \Delta t \displaystyle\sum_{j=1}^{q} a_{i,j}\, f(t_j^n, v_j^n)\,, \\ \\ v^{n+1} \ = \ v^n \ + \ \Delta t \displaystyle\sum_{j=1}^{q} b_j\, f(t_j^n, v_j^n)\,. \end{array} \right.$$

Une solution de ce schéma est un N-uplet $(v^n)_{n\in\{0,\dots,N\}}$ de vecteurs de U pour lequel il existe $v_i^n \in U$ vérifiant (9) *pour tous $i \in \{1,\dots,q\}$ et $n \in \{0,\dots,N-1\}$. Ce schéma est explicite si $a_{i,j} = 0$ pour $j \geq i$.*

Tous les schémas rencontrés jusqu'à présent entrent dans la catégorie des schémas de Runge-Kutta, à un étage (Euler) ou deux étages (Runge, θ-schéma). Parmi les autres schémas classiques on trouve celui à quatre étages, défini par

(10)
$$\left\{ \begin{array}{l} v_1^n \ = \ v^n\,, \\ v_2^n \ = \ v^n \ + \ \frac{1}{2}\,\Delta t\, f(t^n + \frac{1}{2}\Delta t, v_1^n)\,, \\ v_3^n \ = \ v^n \ + \ \frac{1}{2}\,\Delta t\, f(t^n + \frac{1}{2}\Delta t, v_2^n)\,, \\ v_4^n \ = \ v^n \ + \ \Delta t\, f(t^n + \Delta t, v_3^n)\,, \\ v^{n+1} \ = \ v^n \ + \ \frac{1}{6}\,\Delta t\, \left(f(t^n, v_1^n) + 2f(t^n + \frac{1}{2}\Delta t, v_2^n) + 2f(t^n + \frac{1}{2}\Delta t, v_3^n) + f(t^n, v_4^n)\right)\,, \end{array} \right.$$

dont on peut montrer qu'il est d'ordre 4 (ce qui est le maximum que l'on puisse espérer pour un schéma explicite à 4 étages).

Appendice : Lemme de Gronwall

Inéquations différentielles On appelle souvent lemme de Gronwall l'observation suivante. Supposons qu'une fonction $u \in \mathscr{C}^1(I;\mathbb{R})$ vérifie

$$u'(t) \,\leq\, a(t)\,u(t) \,+\, b(t)\,,$$

avec a et $b \in \mathscr{C}(I;\mathbb{R})$. Alors en s'inspirant de la résolution de l'équation différentielle $u' = a(t)\,u \,+\, b(t)$, on multiplie l'inégalité par le nombre strictement positif $\mathrm{e}^{-\int_{t_0}^t a(\tau)\,\mathrm{d}\tau}$ et l'on en déduit

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathrm{e}^{-\int_{t_0}^t a(\tau)\,\mathrm{d}\tau}\,u(t)\right) \,\leq\, \mathrm{e}^{-\int_{t_0}^t a(\tau)\,\mathrm{d}\tau}\,b(t)\,,$$

d'où en intégrant :

$$\mathrm{e}^{-\int_{t_0}^t a(\tau)\,\mathrm{d}\tau}\,u(t) \,-\, u(t_0) \,\leq\, \int_{t_0}^t \mathrm{e}^{-\int_{t_0}^s a(\tau)\,\mathrm{d}\tau}\,b(s)\,\mathrm{d}s\,,$$

et finalement :

$$u(t) \,\leq\, u(t_0)\,\mathrm{e}^{\int_{t_0}^t a(\tau)\,\mathrm{d}\tau} \,+\, \int_{t_0}^t \mathrm{e}^{\int_s^t a(\tau)\,\mathrm{d}\tau}\,b(s)\,\mathrm{d}s\,.$$

Inéquations intégrales Le « vrai » lemme de Gronwall est un peu plus subtil, puisqu'il suppose une inégalité intégrale et non une inégalité différentielle (la seconde impliquant la première mais pas l'inverse). Or les *estimations a priori* que l'on obtient en général sont plutôt du type intégral, d'où l'intérêt de ce lemme, dont la preuve est néanmoins élémentaire.

Lemme 1 (Gronwall) *Si* $u \in \mathscr{C}([0,T];\mathbb{R}^+)$ *est telle qu'il existe* $a \in \mathscr{C}([0,T];\mathbb{R}^+)$ *et* $c \in \mathscr{C}([0,T];\mathbb{R})$ *avec*

$$u(t) \,\leq\, c(t) \,+\, \int_0^t a(\tau)\,u(\tau)\,\mathrm{d}\tau\,, \quad \textit{pour tout } t \in [0,T]\,,$$

alors

$$u(t) \,\leq\, c(t) \,+\, \int_0^t c(\tau)\,a(\tau)\,\mathrm{e}^{\int_\tau^t a(s)\,\mathrm{d}s}\,\mathrm{d}\tau\,, \quad \textit{pour tout } t \in [0,T]\,.$$

Dém. L'astuce consiste à démontrer l'inégalité voulue non pas pour u directement mais pour le second membre de l'inégalité dont on dispose. Plus précisément, posons

$$v(t) \,=\, \int_0^t a(\tau)\,u(\tau)\,\mathrm{d}\tau\,.$$

Alors on a, et c'est ici qu'intervient l'hypothèse $a \geq 0$,

$$v'(t) \,=\, a(t)\,u(t) \,\leq\, a(t)\,(\,c(t) \,+\, v(t)\,)$$

pour tout $t \in [0,T]$. Ainsi v satisfait une inéquation différentielle comme au paragraphe précédent, avec $b(t) \,=\, a(t)\,c(t)$. En observant que $v(0) \,=\, 0$, on en déduit que

$$v(t) \,\leq\, \int_0^t a(\tau)\;\; c(\tau)\;\, \mathrm{e}^{\int_\tau^t a(s)\,\mathrm{d}s}\,\mathrm{d}\tau\,, \quad \text{pour tout } t \in [0,T]\,.$$

On conclut en injectant cette majoration de $v(t)$ dans l'inégalité de départ. □

Suivant le contexte, l'inégalité de Gronwall peut se simplifier ou s'exprimer différemment. Si c est constante par exemple, on obtient la majoration :

$$u(t) \leq c\, \mathrm{e}^{\int_0^t a(s)\,\mathrm{d}s}.$$

D'autre part, on peut écrire une formule analogue avec un point t_0 quelconque à la place de 0. (Mais attention aux bornes des intégrales que l'on manipule, il faut penser à mettre des valeurs absolues au bon endroit.) Enfin, si c est dérivable on peut donner une autre version de l'inégalité de Gronwall : en intégrant par parties on obtient en effet

$$u(t) \leq c(0)\, \mathrm{e}^{\int_0^t a(s)\,\mathrm{d}s} + \int_0^t c'(\tau)\, \mathrm{e}^{\int_\tau^t a(s)\,\mathrm{d}s}\,\mathrm{d}\tau\,, \quad \text{pour tout } t \in [0,T]\,.$$