Résolution numérique d'EDO : méthode d'Euler implicite

On s'intéresse à l'équation différentielle

$$y' = f(t, y), \quad t \in I,\, y \in U \subset \mathbb{R}^n \tag{1}$$

d'inconnue $y : I \to \mathbb{R}^n$. Comme précédemment, on suppose que le problème de Cauchy de condition initiale $y(t_0) = y_0$ admet une solution maximale unique notée z. Plutôt que d'approximer la différence

$$z(t_{k+1}) - z(t_k) = \int_{t_k}^{t_{k+1}} f(t, z(t))dt$$

par la méthode des rectangles à gauche, qui amène à la méthode d'Euler explicite, on peut utiliser la méthode des rectangles à droite, conduisant à la formule suivante

$$y_{k+1} = y_k + h_k f(t_{k+1}, y_{k+1}). \tag{2}$$

C'est la méthode d'Euler dite *implicite*, car elle nécessite la résolution d'une équation non linéaire à chaque étape, par exemple en utilisant l'algorithme de Newton. Elle est en conséquence plus coûteuse que la méthode d'Euler explicite. Les définitions de consistance, de stabilité et de convergence s'étendent naturellement aux méthodes implicites, et on montre, de la même manière que pour Euler explicite, que la méthode implicite est aussi d'ordre 1.

Proposition 1. *Si f est de classe $\mathcal{C}^1$, alors la méthode d'Euler implicite est consistante d'ordre 1.*

L'avantage de cette méthode n'est donc pas un ordre élevé, mais, comme on va le voir, c'est sa stabilité pour certaines équations particulières, appelées parfois équation différentielle *raide*. Pour illustrer cette notion avec un exemple, considérons l'équation différentielle très simple

$$y' = -\lambda y,$$

avec $\lambda > 0$. Les solutions exactes sont les $y(t) = y_0 e^{-\lambda t}$, qui sont des fonctions qui décroissent rapidement vers 0 lorsque t croît. Étant donné un pas de temps h, la solution approchée donnée par la méthode d'Euler explicite est

$$y_{k+1} = y_k - h\lambda y_k,$$

donc

$$y_k = y_0(1 - h\lambda)^k,$$

dont l'allure est similaire à une solution exacte que si $|1 - h\lambda| < 1$, et donc $0 < h < 2/\lambda$. Si λ est grand, cela implique que le pas h doit être très petit, ce qui n'est pas toujours possible pour une application pratique. La méthode d'Euler implicite, quant à elle, donne la solution approchée

$$y_k = \frac{y_0}{(1 + h\lambda)^k},$$

qui converge vers 0 inconditionnellement, ce qui est numériquement préférable. Ce genre de comportements très différents entre les schémas implicite et explicite survient typiquement dans des problèmes physiques à plusieurs échelles, comme par exemple plusieurs réactions chimiques simultanées dont certaines sont lentes et d'autres rapides. Dans ce genre de cas, on préférera des méthodes implicites.

Citons pour finir un cas particulier de système où la méthode d'Euler implicite est particulièrement stable.

Définition 2. On dit que le système est dissipatif si pour tout $t \geqslant 0$ et $z_1, z_2 \in \mathbb{R}^n$,

$$(f(t, z_1) - f(t, z_2),\, z_1 - z_2) \geqslant 0.$$

De manière équivalente, si f est de classe $\mathcal{C}^1$,

$$(\nabla_y f(t, z_1)z_2,\, z_2) \geqslant 0.$$

Si z_1 et z_2 sont deux solutions de l'équation différentielle (1), alors

$$\frac{d}{dt}|z_1 - z_2|^2 = 2(f(t, z_1) - f(t, z_2),\, z_1 - z_2) \leqslant 0.$$

En particulier, tous les points d'équilibre du système sont stables. C'est une condition assez forte sur f, qui n'est vérifiée que dans certains cas particuliers. On peut penser à des systèmes physiques où l'énergie décroît – elle est dissipée –, comme par exemple le mouvement d'une particule chargée de vitesse v, soumise à un champ électrique E et un champ magnétique B constants, ainsi qu'une force de frottement $-kv$, où $k \geqslant 0$. Si l'on normalise sa masse et sa charge, l'équation du mouvement s'écrit

$$\frac{dv}{dt} = (E + v \times B) - kv,$$

et alors

$$(f(v_1) - f(v_2),\, v_1 - v_2) = -k|v_1 - v_2|^2 \leqslant 0.$$

On dispose dans ce cas particulier du théorème suivant.

Théorème 3. *On suppose que f est de classe $\mathcal{C}^1$, et que le système est dissipatif. Alors*

1. *pour tout pas $h_{max} > 0$, la suite $(y_k)_{k\geqslant 0}$ définie par la méthode d'Euler implicite (2) est bien définie, et*
2. *la méthode est consistante d'ordre 1, et stable avec la constante de stabilité $S = 1$.*

Démonstration. 1. Commençons par montrer la bonne définition de la suite (y_k). Étant donné y_k, on introduit la fonction $G(h, z) = z - y_k - hf(t_k + h, z)$. Trouver y_{k+1} revient alors à démontrer l'existence d'un point z tel que $G(h_k, z) = 0$. Notons que l'unicité n'est pas obligatoire, mais elle est en tout les cas garantie par le caractère dissipatif du système : on montre en fait que pour tout $h \geqslant 0$, $z \mapsto G(h, z)$ est injective. Si $z_1, z_2 \in \mathbb{R}^n$,

$$\begin{aligned} G(h, z_1) = G(h, z_2) &\implies z_1 - z_2 = h(f(t_k + h, z_1) - f(t_k + h, z_2)) \\ &\implies |z_1 - z_2|^2 = h(f(t_k + h, z_1) - f(t_k + h, z_2),\, z_1 - z_2) \leqslant 0 \\ &\implies z_1 = z_2 \end{aligned}$$

Soit $E = \{h^* \geqslant 0$ tel que $\forall h \in [0, h^*], \exists z \in \mathbb{R}^n, G(h, z) = 0\}$ On va montrer que $E = \mathbb{R}_+$, ce qui prouvera la bonne définition de (y_k). Comme $G(0, y_k) = 0$, E est non vide. Par ailleurs, G est de classe $\mathcal{C}_1$, et

$$\nabla_y G(h, z) = I_n - h\nabla_y f(t_k + h, z),$$

où ∇_y désigne le gradient par rapport à la deuxième variable. Ainsi,

$$\begin{aligned} (G(h, z_1)z_2,\, z_2) &= |z_2|^2 - h(\nabla_y f(t_k + h, z_1)z_2,\, z_2) \\ &\geqslant |z_2|^2, \end{aligned}$$

et donc $\nabla_y G(h, z)$ est inversible pour tout (h, z). Par conséquent, le théorème des fonctions implicites assure que si (h, z) vérifie $G(h, z) = 0$, il existe des voisinages ouverts $V \subset [0, +\infty[$ de h et Ω de (h, z), ainsi qu'une fonction ϕ de classe $\mathcal{C}^1$ définie sur V tels que

$$((h, z) \in \Omega \text{ et } G(h, z) = 0) \iff (h \in V \text{ et } z = \phi(h)).$$

En particulier, E est ouvert dans $[0, +\infty[$. Supposons maintenant que $\sup E = \bar{h}^* < +\infty$. Par définition, il existe une suite (h_p) croissante et convergeante vers h^*, ainsi qu'une suite $(z_p) \subset \mathbb{R}^n$ telle que $G(h_p, z_p) = 0$. Mais alors

$$\begin{aligned} z_p &= y_k + h_p f(t_k + h_p, z_p) \\ &= y_k + h_p\big(f(t_k + h_p, z_p) - f(t_k + h_p, y_k)\big) + h_p f(t_k + h_p, y_k) \end{aligned}$$

et donc, par dissipativité,

$$\begin{aligned}|z_p - y_k|^2 &\leqslant h_p(f(t_k + h_p, y_k),\, z_p - y_k)\\ &\leqslant h_p|f(t_k + h_p, y_k)||z_p - y_k|,\end{aligned}$$

et on en déduit que la suite $(z_p - y_k)$ est bornée. On conclut en extrayant une sous-suite convergente, et la continuité de f implique que $h_* \in E$, et donc que E est fermé, ce qui montre que E est nécessairement égal à $\mathbb{R}_+$ tout entier.

2. On a déjà montré que la méthode était consistante, il reste donc à montrer qu'elle est stable. Soit (ε_k) une suite de points dans $\mathbb{R}^n$, et soient (y_k) et $(\tilde{y}_k)$ les suites définies par

$$\begin{cases} y_{k+1} = y_k + h_k f(t_{k+1}, y_{k+1}) \\ \tilde{y}_{k+1} = \tilde{y}_k + h_k f(t_{k+1}, \tilde{y}_{k+1}) + \varepsilon_k. \end{cases}$$

Notons que la suite $(\tilde{y}_k)$ est bien définie, car les résultats du premier point s'appliquent. On obtient immédiatement

$$\begin{aligned}|y_{k+1} - \tilde{y}_{k+1}|^2 &\leqslant (y_k - \tilde{y}_k - \varepsilon_k n y_{k+1} - \tilde{y}_{k+1})\\ &\leqslant (|y_k - \tilde{y}_k| + |\varepsilon_k|)|y_{k+1} - \tilde{y}_{k+1}|,\end{aligned}$$

et donc $|y_{k+1} - \tilde{y}_{k+1}| \leqslant |y_k - \tilde{y}_k| + |\varepsilon_k|$. Finalement, avec une récurrence immédiate, on en conclut que

$$\max_{0 \leqslant k \leqslant N-1} |y_k - \tilde{y}_k| \leqslant |y_0 - \tilde{y}_0| + \sum_{k=0}^{N-1} |\varepsilon_k|,$$

ce qui prouve la stabilité avec constante 1. □