1 Schémas numériques pour Équations Différentielles Ordinaires

1.1 Le Problème de Cauchy

On s'intéresse à la résolution numérique du problème de Cauchy (P) suivant :

$$(P) \quad \begin{cases} x'(t) = f(t, x(t)), \quad t \in [t_0, t_0 + T] \\ x(t_0) = x_0 \end{cases}$$

où $f : [t_0, t_0 + T] \times \mathbb{R}^d \to \mathbb{R}^d$ est une fonction donnée, $x_0 \in \mathbb{R}^d$ est la condition initiale, et $x(t) \in \mathbb{R}^d$ est la solution recherchée.

1.2 Schémas à un pas explicites

Pour approcher la solution de (P), on construit des schémas numériques. Un schéma à un pas explicite (S) se présente sous la forme générale :

$$(S) \quad \begin{cases} x_{n+1} = x_n + \Delta t \Phi(t_n, x_n, \Delta t) \\ x_0 \text{ donné (souvent } x_0 = x(t_0)) \end{cases}$$

où :

- $t_n = t_0 + n\Delta t$ sont les points de la discrétisation en temps, pour $n = 0, \ldots, N$.
- $\Delta t = T/N$ est le pas de temps, avec N le nombre total de pas.
- x_n est une approximation de la solution $x(t_n)$.
- Φ est la fonction d'increment qui caractérise le schéma.

1.3 Exemples de schémas à un pas explicites

1.3.1 Schéma d'Euler explicite

Pour ce schéma, la fonction d'increment est $\Phi(t_n, x_n, \Delta t) = f(t_n, x_n)$. Le schéma s'écrit :

$$x_{n+1} = x_n + \Delta t f(t_n, x_n)$$

1.3.2 Schéma du Point Milieu (Runge-Kutta d'ordre 2)

La fonction d'increment est $\Phi(t_n, x_n, \Delta t) = f\left(t_n + \frac{\Delta t}{2}, x_n + \frac{\Delta t}{2} f(t_n, x_n)\right)$. Le schéma s'écrit :

$$x_{n+1} = x_n + \Delta t f\left(t_n + \frac{\Delta t}{2}, x_n + \frac{\Delta t}{2} f(t_n, x_n)\right)$$

1.3.3 Schéma de Heun (Runge-Kutta d'ordre 2)

La fonction d'increment est $\Phi(t_n, x_n, \Delta t) = \frac{1}{2}\left[f(t_n, x_n) + f(t_n + \Delta t, x_n + \Delta t f(t_n, x_n))\right]$. Le schéma s'écrit :

$$x_{n+1} = x_n + \frac{\Delta t}{2}\left[f(t_n, x_n) + f(t_n + \Delta t, x_n + \Delta t f(t_n, x_n))\right]$$

2 Étude de convergence pour les EDO

On s'intéresse à savoir si la suite $(x_n)_{0 \le n \le N}$ générée par un schéma (S) converge vers la solution $x(t)$ du problème (P) lorsque $\Delta t \to 0$.

2.1 Convergence

Definition 2.1 (Convergence d'un schéma)**.** Le schéma (S) est dit convergent si, pour toute solution x de (P) et pour toute suite $(x_n)_{0\leq n\leq N}$ construite par (S) avec $x_0 = x(t_0)$, on a :

$$\lim_{\Delta t\to 0}\left(\max_{0\leq n\leq N}\|x_n - x(t_n)\|\right) = 0$$

Remark 2.2. Si la suite $(x_n)_{0\leq n\leq N}$ est "bien" construite (par exemple, si le problème (P) admet une solution unique x et si f est de classe C^p assurant une régularité suffisante à x), la convergence peut être établie. L'erreur globale $\max_{0\leq n\leq N}\|x_n - x(t_n)\|$ dépend de Δt. En général, la convergence d'un schéma numérique est la conséquence de sa consistance et de sa stabilité (Théorème de Lax-Richtmyer, adapté aux EDO).

2.2 Consistance et Ordre d'un schéma

Remark 2.3. Pour analyser un schéma numérique pour EDO, on procède typiquement en deux étapes :

1. On étudie la consistance du schéma (quelle est l'erreur commise en un seul pas ?).
2. On examine sa stabilité (comment les erreurs se propagent-elles ?).

Ces deux propriétés permettent ensuite de conclure sur la convergence.

2.2.1 Erreur de consistance (Erreur locale de troncature)

Definition 2.4 (Erreur de consistance)**.** Soit $x(t)$ la solution exacte du problème (P). On appelle erreur locale de troncature (ou erreur de consistance) du schéma (S) à l'instant t_{n+1}, la quantité e_{n+1} définie par :

$$e_{n+1} = x(t_{n+1}) - (x(t_n) + \Delta t\Phi(t_n, x(t_n), \Delta t))$$

Cette erreur mesure à quel point la solution exacte $x(t)$ échoue à satisfaire l'équation du schéma numérique.

2.2.2 Ordre d'un schéma

Definition 2.5 (Ordre d'un schéma)**.** Le schéma (S) est dit consistant d'ordre (au moins) $q \geq 1$ si, pour toute solution exacte $x(t)$ de (P) suffisamment régulière, il existe une constante $C > 0$ (indépendante de Δt) telle que :

$$\max_{0\leq n\leq N-1}\|e_{n+1}\| \leq C(\Delta t)^{q+1}$$

Autrement dit, $e_{n+1}(\Delta t) = O((\Delta t)^{q+1})$. Si un schéma est consistant, alors $q \geq 1$. Le schéma est dit d'ordre q s'il est d'ordre au moins q et pas d'ordre au moins $q+1$. Cela signifie que $e_{n+1}(\Delta t) = K(\Delta t)^{q+1} + O((\Delta t)^{q+2})$ avec $K \neq 0$.

Example 2.6 (Étude de consistance du schéma d'Euler explicite)**.** Pour le schéma d'Euler explicite, $\Phi(t, y, \Delta t) = f(t, y)$. L'erreur locale de troncature est, pour une solution $x(t)$ suffisamment régulière :

$$e(t, \Delta t) = x(t + \Delta t) - (x(t) + \Delta t f(t, x(t)))$$

Comme $x(t)$ est solution de (P), on a $x'(t) = f(t, x(t))$. Donc :

$$e(t, \Delta t) = x(t + \Delta t) - x(t) - \Delta t x'(t)$$

En effectuant un développement de Taylor de $x(t+\Delta t)$ autour de t, en supposant $x \in C^2([t_0, t_0+T])$:

$$x(t+\Delta t) = x(t) + \Delta t x'(t) + \frac{(\Delta t)^2}{2} x''(t) + O((\Delta t)^3)$$

En substituant ce développement dans l'expression de $e(t, \Delta t)$:

$$\begin{aligned} e(t, \Delta t) &= \left(x(t) + \Delta t x'(t) + \frac{(\Delta t)^2}{2} x''(t) + O((\Delta t)^3) \right) - x(t) - \Delta t x'(t) \\ &= \frac{(\Delta t)^2}{2} x''(t) + O((\Delta t)^3) \end{aligned}$$

Ainsi, pour le schéma d'Euler explicite, l'erreur locale de troncature à l'instant t_n est $e_{n+1} = \frac{(\Delta t)^2}{2} x''(t_n) + O((\Delta t)^3)$. On a $q+1 = 2$, donc $q = 1$. Le schéma d'Euler explicite est d'ordre (au moins) 1 si $x \in C^2$. Si de plus $x \in C^3$ et $x''(t_n) \neq 0$ pour au moins un t_n, alors le schéma est exactement d'ordre 1.

2.2.3 Conditions d'ordre pour une fonction d'increment Φ

Remark 2.7 (Régularité de Φ). Soit $\Phi : I \times \mathbb{R}^d \times [0, \Delta t_0] \to \mathbb{R}^d$, où $I = [t_0, t_0+T]$ et $\Delta t_0 > 0$. La fonction Φ est dite de classe C^p si toutes ses dérivées partielles par rapport à t, x (ou y), et Δt jusqu'à l'ordre p existent et sont continues sur $I \times \mathbb{R}^d \times [0, \Delta t_0]$.

Pour une fonction d'increment Φ de classe C^q, le schéma $x_{n+1} = x_n + \Delta t \Phi(t_n, x_n, \Delta t)$ est consistant d'ordre au moins q si et seulement si les conditions suivantes sont vérifiées pour tout (t, y) dans le domaine de définition :

1. $\Phi(t, y, 0) = f(t, y)$ (Condition pour l'ordre 0, implique $q \geq 1$ pour la consistance)
2. $\dfrac{\partial^k \Phi}{\partial (\Delta t)^k}(t, y, 0) = \dfrac{1}{k+1} D_t^k f(t, y)$, pour $k = 1, \ldots, q-1$.

où $D_t f(t, y)$ est l'opérateur de dérivée totale par rapport à t le long des solutions de l'EDO:

$$D_t f(t, y) = \frac{\partial f}{\partial t}(t, y) + \frac{\partial f}{\partial y}(t, y) \cdot f(t, y)$$

et $D_t^0 f = f$, $D_t^k f = D_t(D_t^{k-1} f)$ pour $k \geq 1$.

Example 2.8 (Étude de consistance du schéma de Runge-Kutta d'ordre 2 (Point Milieu)). Le schéma est $x_{n+1} = x_n + \Delta t f(t_n + \frac{\Delta t}{2}, x_n + \frac{\Delta t}{2} f(t_n, x_n))$. La fonction d'increment est $\Phi(t, y, \Delta t) = f\left(t + \frac{\Delta t}{2}, y + \frac{\Delta t}{2} f(t, y)\right)$. Vérifions les conditions d'ordre :

1. Condition pour $q \geq 1$:
$$\Phi(t, y, 0) = f\left(t + 0, y + 0 \cdot f(t, y)\right) = f(t, y)$$
Cette condition est vérifiée. Donc le schéma est au moins d'ordre 1.

2. Condition pour $q \geq 2$ (correspond à $k = 1$ dans la formule générale): On doit vérifier si $\dfrac{\partial \Phi}{\partial (\Delta t)}(t, y, 0) = \dfrac{1}{2} D_t f(t, y)$. Calculons $\dfrac{\partial \Phi}{\partial (\Delta t)}$: Soient $u_1(t, \Delta t) = t + \frac{\Delta t}{2}$ et $u_2(y, \Delta t) = y + \frac{\Delta t}{2} f(t, y)$. Alors $\Phi(t, y, \Delta t) = f(u_1, u_2)$.

$$\begin{aligned} \frac{\partial \Phi}{\partial (\Delta t)}(t, y, \Delta t) &= \frac{\partial f}{\partial u_1}(u_1, u_2) \cdot \frac{\partial u_1}{\partial (\Delta t)} + \frac{\partial f}{\partial u_2}(u_1, u_2) \cdot \frac{\partial u_2}{\partial (\Delta t)} \\ &= \frac{\partial f}{\partial t}\left(t + \frac{\Delta t}{2}, y + \frac{\Delta t}{2} f(t, y)\right) \cdot \frac{1}{2} + \frac{\partial f}{\partial y}\left(t + \frac{\Delta t}{2}, y + \frac{\Delta t}{2} f(t, y)\right) \cdot \left(\frac{1}{2} f(t, y)\right) \end{aligned}$$

Évaluons en $\Delta t = 0$:

$$\begin{aligned}\frac{\partial \Phi}{\partial(\Delta t)}(t,y,0) &= \frac{1}{2}\frac{\partial f}{\partial t}(t,y) + \frac{1}{2}\frac{\partial f}{\partial y}(t,y)f(t,y)\\ &= \frac{1}{2}\left(\frac{\partial f}{\partial t}(t,y) + \frac{\partial f}{\partial y}(t,y)f(t,y)\right)\\ &= \frac{1}{2}D_t f(t,y)\end{aligned}$$

Cette condition est vérifiée. Donc le schéma est au moins d'ordre 2.

3. Condition pour $q \geq 3$ (correspond à $k = 2$ dans la formule générale): On devrait vérifier si $\frac{\partial^2 \Phi}{\partial(\Delta t)^2}(t,y,0) = \frac{1}{3}D_t^2 f(t,y)$. En général, cette condition n'est pas satisfaite pour une fonction f arbitraire. Pour montrer que le schéma n'est pas d'ordre au moins 3 (et donc qu'il est exactement d'ordre 2), il faudrait calculer $\frac{\partial^2 \Phi}{\partial(\Delta t)^2}(t,y,0)$ et montrer que cette expression est différente de $\frac{1}{3}D_t^2 f(t,y)$, ou que le terme $g(t,y)$ dans $e_{n+1} = g(t,y)(\Delta t)^3 + O((\Delta t)^4)$ est non nul.

Le schéma du point milieu est donc d'ordre 2.

2.3 Stabilité

Definition 2.9 (Stabilité d'un schéma)**.** Le schéma (S) est dit stable pour une classe de fonctions f s'il existe une constante $S > 0$, indépendante de Δt, telle que pour toutes suites $(x_n)_{0\leq n\leq N}$ et $(y_n)_{0\leq n\leq N}$ vérifiant :

$$\begin{aligned}x_{n+1} &= x_n + \Delta t \Phi(t_n, x_n, \Delta t)\\ y_{n+1} &= y_n + \Delta t \Phi(t_n, y_n, \Delta t) + \delta_n \quad \text{(perturbation)}\end{aligned}$$

pour $n = 0, \ldots, N-1$, où les δ_n sont des perturbations, on ait la majoration :

$$\max_{0\leq n\leq N} \|x_n - y_n\| \leq S\left(\|x_0 - y_0\| + \sum_{j=0}^{N-1}\|\delta_j\|\right)$$

La stabilité garantit que de petites perturbations (erreurs initiales ou erreurs introduites à chaque pas) n'entraînent pas une divergence incontrôlée des solutions numériques.

Proposition 2.10 (Stabilité du schéma d'Euler explicite)**.** Si la fonction $f(t,y)$ est Lipschitzienne par rapport à y, uniformément en t, c'est-à-dire s'il existe $L_f > 0$ telle que $\|f(t,y_1) - f(t,y_2)\| \leq L_f\|y_1 - y_2\|$ pour tous t, y_1, y_2, alors le schéma d'Euler explicite est stable.

Preuve. Soient les suites (x_n) et (y_n) définies par :

$$\begin{aligned}x_{n+1} &= x_n + \Delta t f(t_n, x_n)\\ y_{n+1} &= y_n + \Delta t f(t_n, y_n) + \delta_n\end{aligned}$$

Soustrayons les deux équations :

$$x_{n+1} - y_{n+1} = (x_n - y_n) + \Delta t(f(t_n, x_n) - f(t_n, y_n)) - \delta_n$$

En prenant la norme et en utilisant l'inégalité triangulaire et la condition de Lipschitz sur f:

$$\begin{aligned}\|x_{n+1}-y_{n+1}\| &\leq \|x_n - y_n + \Delta t(f(t_n,x_n)-f(t_n,y_n))\| + \|-\delta_n\| \\ &\leq \|x_n-y_n\| + \Delta t\|f(t_n,x_n)-f(t_n,y_n)\| + \|\delta_n\| \\ &\leq \|x_n-y_n\| + \Delta t L_f\|x_n-y_n\| + \|\delta_n\| \\ &= (1+L_f\Delta t)\|x_n-y_n\| + \|\delta_n\|\end{aligned}$$

Posons $e_n = \|x_n - y_n\|$ et $\gamma_n = \|\delta_n\|$. On a $e_{n+1} \leq (1+L_f\Delta t)e_n + \gamma_n$. On utilise une version du lemme de Gronwall discret.

Lemma 2.11 (Lemme de Gronwall discret)**.** Soient $(\alpha_n)_{0\leq n\leq N}$, $(\gamma_n)_{0\leq n\leq N-1}$ des suites de réels positifs ou nuls, et $(\beta_n)_{0\leq n\leq N-1}$ une suite de réels positifs ou nuls. Si $\alpha_{n+1} \leq (1+\beta_n)\alpha_n + \gamma_n$ pour $n = 0,\dots,N-1$, alors :

$$\alpha_n \leq \left(\prod_{j=0}^{n-1}(1+\beta_j)\right)\alpha_0 + \sum_{j=0}^{n-1}\left(\prod_{k=j+1}^{n-1}(1+\beta_k)\right)\gamma_j$$

En utilisant $1+u \leq e^u$, on a $\prod(1+\beta_j) \leq \exp(\sum \beta_j)$. Ainsi, $\alpha_n \leq \exp\left(\sum_{j=0}^{n-1}\beta_j\right)\left(\alpha_0 + \sum_{j=0}^{n-1}\gamma_j\right)$ (en majorant $\exp(-\sum\beta_k) \leq 1$ dans une version plus précise). Plus précisément, si $\alpha_{n+1} \leq (1+B)\alpha_n + \Gamma$ pour $B \geq 0$, alors $\alpha_n \leq (1+B)^n\alpha_0 + \Gamma\sum_{j=0}^{n-1}(1+B)^j \leq e^{nB}\alpha_0 + \Gamma\frac{e^{nB}-1}{B}$. Si Γ_j varie: $\alpha_n \leq e^{\sum_{j=0}^{n-1}\beta_j}\left(\alpha_0 + \sum_{j=0}^{n-1}\gamma_j e^{-\sum_{k=0}^{j}\beta_k}\right)$.

Dans notre cas, $\beta_n = L_f\Delta t$. Donc $\sum_{j=0}^{n-1}\beta_j = nL_f\Delta t$. Comme $n\Delta t \leq N\Delta t = T$:

$$e^{\sum_{j=0}^{n-1}L_f\Delta t} = e^{nL_f\Delta t} \leq e^{L_fT}$$

Aussi, $e^{-\sum_{k=0}^{j}L_f\Delta t} \leq 1$. Donc, pour tout $n \in \{0,\dots,N\}$:

$$\|x_n - y_n\| \leq e^{L_fT}\left(\|x_0-y_0\| + \sum_{j=0}^{N-1}\|\delta_j\|\right)$$

Le schéma d'Euler explicite est donc stable avec une constante de stabilité $S = e^{L_fT}$. $\square$

Proposition 2.12 (Condition suffisante de stabilité pour Φ Lipschitzienne)**.** Si la fonction d'increment $\Phi(t,y,\Delta t)$ est Lipschitzienne par rapport à y, uniformément en t et Δt, c'est-à-dire s'il existe une constante $\Lambda > 0$ telle que pour tous $y_1, y_2 \in \mathbb{R}^d$, pour tout $t \in [t_0, t_0+T]$ et pour tout $\Delta t \in [0, \Delta t_0]$:

$$\|\Phi(t,y_1,\Delta t) - \Phi(t,y_2,\Delta t)\| \leq \Lambda\|y_1-y_2\|$$

alors le schéma (S) $x_{n+1} = x_n + \Delta t\Phi(t_n,x_n,\Delta t)$ est stable, avec une constante de stabilité $S = e^{\Lambda T}$.

Preuve. La preuve est analogue à celle pour le schéma d'Euler explicite. On a :

$$\begin{aligned}\|x_{n+1}-y_{n+1}\| &\leq \|x_n - y_n + \Delta t(\Phi(t_n,x_n,\Delta t)-\Phi(t_n,y_n,\Delta t))\| + \|\delta_n\| \\ &\leq \|x_n-y_n\| + \Delta t\|\Phi(t_n,x_n,\Delta t)-\Phi(t_n,y_n,\Delta t)\| + \|\delta_n\| \\ &\leq \|x_n-y_n\| + \Delta t\Lambda\|x_n-y_n\| + \|\delta_n\| \\ &= (1+\Lambda\Delta t)\|x_n-y_n\| + \|\delta_n\|\end{aligned}$$

L'application du lemme de Gronwall discret donne le résultat avec $S = e^{\Lambda T}$. □

2.4 Convergence des schémas à un pas explicite

Le théorème suivant (parfois appelé Théorème de Lax pour les EDO) relie consistance, stabilité et convergence.

Theorem 2.13 (Convergence des schémas à un pas explicites)**.** Si la méthode à un pas explicite (S) est stable et consistante d'ordre $q \geq 1$, alors elle est convergente. De plus, si la solution $x(t)$ du problème de Cauchy (P) est de classe C^{q+1} et si la suite $(x_n)_{0\leq n\leq N}$ est générée par (S) avec $x_0 = x(t_0)$, alors il existe une constante $K > 0$ telle que :

$$\max_{0\leq n\leq N} \|x(t_n) - x_n\| \leq K(\Delta t)^q$$

Plus précisément, si C_{cons} est la constante de l'erreur de consistance (i.e. $\|e_{n+1}\| \leq C_{cons}(\Delta t)^{q+1}$) et S est la constante de stabilité, on a :

$$\max_{0\leq n\leq N} \|x(t_n) - x_n\| \leq S \sum_{j=0}^{N-1} \|e_{j+1}\| \approx S \cdot N \cdot C_{cons}(\Delta t)^{q+1} = S \cdot \frac{T}{\Delta t} \cdot C_{cons}(\Delta t)^{q+1} = S \cdot C_{cons} \cdot T \cdot (\Delta t)^q$$

L'erreur globale est donc en $O((\Delta t)^q)$.

Preuve (Idée de la preuve)**.** Soit $x(t)$ la solution exacte. On a $x(t_{n+1}) = x(t_n) + \Delta t\Phi(t_n, x(t_n), \Delta t) + e_{n+1}$, où e_{n+1} est l'erreur locale de troncature. Le schéma numérique est $x_{n+1} = x_n + \Delta t\Phi(t_n, x_n, \Delta t)$. Soit $\varepsilon_n = x(t_n) - x_n$ l'erreur globale au temps t_n. $\varepsilon_{n+1} = x(t_{n+1}) - x_{n+1} = (x(t_n) - x_n) + \Delta t(\Phi(t_n, x(t_n), \Delta t) - \Phi(t_n, x_n, \Delta t)) + e_{n+1}$. En utilisant la condition de Lipschitz pour Φ (qui découle de la stabilité pour de nombreux schémas ou est une hypothèse pour la stabilité plus générale): $\|\varepsilon_{n+1}\| \leq \|\varepsilon_n\| + \Delta t\Lambda\|\varepsilon_n\| + \|e_{n+1}\| = (1 + \Lambda\Delta t)\|\varepsilon_n\| + \|e_{n+1}\|$. Puisque $x_0 = x(t_0)$, $\varepsilon_0 = 0$. L'erreur locale $\|e_{n+1}\| \leq C_{cons}(\Delta t)^{q+1}$. Par le lemme de Gronwall, avec $\alpha_n = \|\varepsilon_n\|$, $\beta_n = \Lambda\Delta t$, $\gamma_n = C_{cons}(\Delta t)^{q+1}$:

$$\|\varepsilon_n\| \leq e^{n\Lambda\Delta t} \left(\|\varepsilon_0\| + \sum_{j=0}^{n-1} C_{cons}(\Delta t)^{q+1} e^{-(j+1)\Lambda\Delta t} \right)$$

En majorant $e^{-(j+1)\Lambda\Delta t} \leq 1$ et $e^{n\Lambda\Delta t} \leq e^{\Lambda T}$:

$$\|\varepsilon_n\| \leq e^{\Lambda T} \sum_{j=0}^{n-1} C_{cons}(\Delta t)^{q+1} = e^{\Lambda T} n C_{cons}(\Delta t)^{q+1}$$

Comme $n\Delta t \leq T$, on a $n \leq T/\Delta t$.

$$\|\varepsilon_n\| \leq e^{\Lambda T} \frac{T}{\Delta t} C_{cons}(\Delta t)^{q+1} = (e^{\Lambda T} T C_{cons})(\Delta t)^q$$

Ceci montre que $\max_{0\leq n\leq N} \|\varepsilon_n\| = O((\Delta t)^q)$. La constante S du théorème est $e^{\Lambda T}$ et la constante c est C_{cons}. □