Cours d'analyse numérique de licence de mathématiques

Roland Masson

5 décembre 2013

1 Equations différentielles ordinaires

Equations différentielles ordinaires: plan

- Schéma à un pas
- Consistance, stabilité, convergence du schéma
- Schéma d'Euler implicite

Equations différentielles ordinaires

Soit $f \in C^1\left(\mathbb{R}^n \times \mathbb{R}, \mathbb{R}^n\right)$, et $T > 0$, on cherche à approcher numériquement la solution $x \in C^2\left([0, T[, \mathbb{R}^n\right)$ de l'équation différentielle ordinaire (EDO) suivante avec condition initiale en $t = 0$ (problème de Cauchy):

$$\begin{cases} x'(t) = f(x(t), t),\ t \in [0, T[, \\ x(0) = x_0. \end{cases}$$

On rappelle que cette EDO admet une solution unique maximale au sens où il existe T_M et $x \in C^2\left([0, T_M[, \mathbb{R}^n\right)$ solution tels que si il existe T et $y \in C^2\left([0, T[, \mathbb{R}^n\right)$ solution alors $T \leq T_M$ et $y = x$ sur $[0, T[$. De plus si $T_M < +\infty$, alors $\lim_{t \to T_M} \|x(t)\| = +\infty$ (théorème d'explosion en temps fini).

Exemples

Exemple 1: $n = 1$

$$\begin{cases} x'(t) = x^2(t),\ t > 0, \\ x(0) = 1. \end{cases}$$

La solution maximale est $x(t) = \frac{1}{1-t}$ sur l'intervalle $[0, 1[$, $(T_M = 1)$ et la solution explose en $t = 1$.

Exemple 2: le système du second ordre sur $\mathbb{R}$ avec $f \in C^1\Big(\mathbb{R} \times \mathbb{R} \times \mathbb{R}, \mathbb{R}\Big)$

$$\begin{cases} x''(t) = f(x(t), x'(t), t),\ t > 0, \\ x(0) = x_0, \\ x'(0) = y_0, \end{cases}$$

se réécrit comme un système du premier ordre sur $\mathbb{R}^2$

$$\begin{cases} x'(t) = y(t),\ t > 0, \\ y'(t) = f(x, y, t),\ t > 0, \\ x(0) = x_0, \\ y(0) = y_0. \end{cases}$$

Schéma à un pas: notations

- On note $\bar{x} \in C^2\Big([0, T_M[, \mathbb{R}^n\Big)$ la solution maximale de l'EDO et on considère $T \in]0, T_M[$.
- On discrétise l'intervalle $[0, T]$:

$$t_0 = 0 < t_1 \cdots < t_k \cdots < t_m = T.$$

- On note $\Delta t_k = t_{k+1} - t_k$, $k = 0, \cdots, m-1$ et

$$\Delta t = \max_{k=1,\cdots,m-1} \Delta t_k.$$

- On cherche les solutions approchées x_k de $\bar{x}_k = \bar{x}(t_k)$ pour $k = 0, \cdots, m$.
- On note $e_k = \bar{x}_k - x_k$, $k = 0, \cdots, m$, les erreurs d'approximation.

Schéma à un pas

Soit $\phi : \mathbb{R}^n \times \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}^n$, et $x_0 \in \mathbb{R}^n$ une approximation de $\bar{x}_0$, le schéma s'écrit:

$$\frac{x_{k+1} - x_k}{\Delta t_k} = \phi\Big(x_k, t_k, \Delta t_k\Big), \ \ k = 0, \cdots, m-1.$$

Exemple 1: Schéma d'Euler explicite

$$\frac{x_{k+1} - x_k}{\Delta t_k} = f\Big(x_k, t_k\Big), \ \ k = 0, \cdots, m-1,$$

Exemple 2: schéma d'Euler implicite

$$\frac{x_{k+1} - x_k}{\Delta t_k} = f\Big(x_{k+1}, t_{k+1}\Big), \ \ k = 0, \cdots, m-1,$$

Consistance du schéma

On définit l'erreur de consistance du schéma:

$$r_k = \frac{\bar{x}_{k+1} - \bar{x}_k}{\Delta t_k} - \phi\Big(\bar{x}_k, t_k, \Delta t_k\Big), \ \ k = 0, \cdots, m-1,$$

- Le schéma est dit consistant ssi $\max_{k=0,\cdots,m-1} \|r_k\| \to 0$ lorsque $\Delta t \to 0$ (et donc nécessairement $m \to +\infty$).
- Le schéma est dit consistant à l'ordre p ssi il existe une constante β ne dépendant que de f, x_0, T (mais pas de Δt) telle que

$$\|r_k\| \leq \beta \Delta t^p, \ \ k = 0, \cdots, m-1.$$

Condition suffisante de consistance du schéma

Lemme: Si $\phi \in C^0\left(\mathbb{R}^n \times \mathbb{R}_+ \times \mathbb{R}_+, \mathbb{R}^n\right)$ et si pour tout $z \in \mathbb{R}^n$ et $t \in [0, T]$ on a $\phi(z, t, 0) = f(z, t)$, alors le schéma est consistant.

Preuve: comme $\bar{x}(t_{k+1}) - \bar{x}(t_k) = \int_{t_k}^{t_{k+1}} f(\bar{x}(s), s)ds = \int_{t_k}^{t_{k+1}} \phi(\bar{x}(s), s, 0)ds$ on a

$$r_k = \frac{1}{\Delta t_k} \int_{t_k}^{t_{k+1}} \Big(\phi(\bar{x}(s), s, 0) - \phi(\bar{x}(t_k), t_k, \Delta t_k)\Big) ds$$

Par uniforme continuité de $\phi(\bar{x}(s), s, h)$ sur $[0, T] \times [0, H]$ pour tout $H > 0$, alors pour tout $\epsilon > 0$ il existe $h > 0$ tel que si $\Delta t < h$ on a

$\|\phi(\bar{x}(s), s, 0) - \phi(\bar{x}(t_k), t_k, \Delta t_k)\| < \epsilon$ pour tout $s \in [t_k, t_{k+1}]$ et $k = 0, \cdots, m-1$

et donc $\|r_k\| < \epsilon$ pour tout $k = 0, \cdots, m-1$.

Consistance: exemple du schéma d'Euler

On a

$$r_k = \frac{\bar{x}_{k+1} - \bar{x}_k}{\Delta t_k} - f(\bar{x}_k, t_k) = \frac{1}{\Delta t_k} \int_{t_k}^{t_{k+1}} \Big(f(\bar{x}(s), s) - f(\bar{x}(t_k), t_k) \Big) ds$$

Soit $g(s) = f(\bar{x}(s), s) \in C^1\Big([0, T_M[, \mathbb{R}^n\Big)$, on note $M = \sup_{s \in [0,T]} \|g'(s)\|$, on a

$$\|r_k\| \leq \frac{M}{2} \Delta t_k,$$

le schéma d'Euler explicite est donc consistant à l'ordre 1. On obtient de même que le schéma d'Euler implicite est consistant à l'ordre 1.

Exemples de schémas consistants à l'ordre 2

Ils s'obtiennent en approchant l'intégrale $\int_{t_k}^{t_{k+1}} f(\bar{x}(s), s)ds$ par la formule du point milieu ou par celle du trapèze.

Exemple 1: Schéma d'Euler explicite d'ordre 2 (point milieu):

$$\frac{x_{k+1} - x_k}{\Delta t_k} = f\left(x_k + \frac{\Delta t_k}{2} f(x_k, t_k), t_k + \frac{\Delta t_k}{2}\right), \ k = 0, \cdots, m-1,$$

Exemple 2: Schéma de Heun (trapèze):

$$\frac{x_{k+1} - x_k}{\Delta t_k} = f(x_k, t_k)/2 + f\left(x_k + \Delta t_k f(x_k, t_k), t_k + \Delta t_k\right)/2, \ k = 0, \cdots, m-1.$$

Exercice: montrer que ces schémas sont consistants à l'ordre 2.

Convergence du schéma

On note $e_k = \bar{x}_k - x_k$ l'erreur de discrétisation pour $k = 0, \cdots, m$.

- On dit que le schéma converge si, en supposant $e_0 = 0$, on a

$$\max_{k=0,\cdots,m} \|e_k\| \to 0 \text{ lorsque } \Delta t \to 0.$$

- On dit que le schéma converge à l'ordre p si il existe une constante C ne dépendant que de f, $\bar{x}_0$, T et telle que si $e_0 = 0$ alors

$$\max_{k=0,\cdots,m} \|e_k\| \leq C\Delta t^p.$$

Stabilité du schéma par rapport aux erreurs

On dit que le schéma est stable par rapport aux erreurs si il existe $\Delta t^* > 0$ et $K \geq 0$ ne dépendant que de $f, \bar{x}_0$ et T tels que si $\Delta t \leq \Delta t^*$ et

$$\begin{cases} \frac{x_{k+1}-x_k}{\Delta t_k} = \phi\Big(x_k, t_k, \Delta t_k\Big), \ k = 0, \cdots, m-1, \ x_0 \in \mathbb{R}^n, \\ \frac{y_{k+1}-y_k}{\Delta t_k} = \phi\Big(y_k, t_k, \Delta t_k\Big) + \epsilon_k, \ k = 0, \cdots, m-1, \ y_0 \in \mathbb{R}^n, \end{cases}$$

pour $\epsilon_k \in \mathbb{R}^n, k = 0, \cdots, m-1$ donnés, alors

$$\|x_k - y_k\| \leq K\Big(\|x_0 - y_0\| + \sum_{i=0}^{k-1} \Delta t_i \|\epsilon_i\|\Big), \ k = 1, \cdots, m.$$

Condition suffisante de stabilité du schéma par rapport aux erreurs

Si il existe $\Delta t^* > 0$ et $M > 0$ tel que pour tout $(x, y) \in \mathbb{R}^n \times \mathbb{R}^n$, $0 \leq h < \Delta t^*$ et $t \in [0, T]$ on ait

$$\|\phi(x, t, h) - \phi(y, t, h)\| \leq M\|y - x\|,$$

alors le schéma est stable.

Preuve: on a pour tous $k = 0, \cdots m - 1$:

$$\|y_{k+1} - x_{k+1}\| \leq (1 + \Delta t_k M)\|y_k - x_k\| + \Delta t_k \|\epsilon_k\| \leq e^{\Delta t_k M}\|y_k - x_k\| + \Delta t_k \|\epsilon_k\|$$

on en déduit par récurrence que

$$\|y_k - x_k\| \leq e^{t_k M}\|y_0 - x_0\| + \sum_{i=0}^{k-1} e^{(t_k - t_{i+1})M} \Delta t_i \|\epsilon_i\| \leq e^{TM}\left(\|y_0 - x_0\| + \sum_{i=0}^{k-1} \Delta t_i \|\epsilon_i\|\right)$$

Difficulté: on n'a pas en général la propriété de Lipchitzité de ϕ sur $\mathbb{R}^n \times \mathbb{R}^n$ mais seulement sur des bornés. Par exemple pour le schéma d'Euler, il faut en général pour appliquer le résultat, tenir compte du fait que sur $[0, T]$, la solution $\bar{x}$ reste dans un compact B et modifier f en dehors de B pour vérifier les hypothèses de Lipchitzité sur $\mathbb{R}^n \times \mathbb{R}^n$.

Théorème de convergence du schéma à un pas

Si le schéma est consistant d'ordre p et stable par rapport aux erreurs alors il est convergeant à l'ordre p.

Preuve: le schéma étant consistant à l'ordre p on a

$$\bar{x}_{k+1} - \bar{x}_k = \Delta t_k \phi\Big(\bar{x}_k, t_k, \Delta t_k\Big) + \Delta t_k r_k, \;\; k = 0, \cdots, m-1,$$

avec $\|r_k\| \leq \beta \Delta t^p$. En utilisant la stabilité du schéma par rapport aux erreurs on a donc pour tout $k = 1, \cdots, m$:

$$\|e_k\| \leq K\Big(\|e_0\| + \beta(\sum_{i=0}^{k-1} \Delta t_k)\Delta t^p\Big) \leq K\Big(\|e_0\| + \beta T \Delta t^p\Big).$$

Théorème de convergence général du schéma à un pas

Théorème: On note $B_A = \{x, \|x\| \leq A\}$. Soit $T \in]0, T_M[$ et

$$A^* = sup_{t\in[0,T]}\|\bar{x}(t)\|.$$

On suppose que le schéma est consistant à l'ordre $p > 0$ au sens où il existe $\beta > 0$ tel que

$$\|r_k\| \leq \beta \Delta t^p, \ \ k = 0, \cdots, m-1.$$

On suppose qu'il existe $\Delta t^* > 0$ tel que:
pour tout $A > 0$ il existe $M_A > 0$ tel que pour tout $(x, y) \in B_A \times B_A$, $0 \leq h < \Delta t^*$ et $t \in [0, T]$ on ait

$$\|\phi(y, t, h) - \phi(x, t, h)\| \leq M_A\|y - x\|.$$

Alors il existe $\Delta t^{**} > 0$, et $K > 0$ et $\epsilon > 0$, tels que si $\|e_0\| \leq \epsilon$, et $0 \leq \Delta t \leq \Delta t^{**}$ alors

- $x_k \in B_{2A^*}$ pour tout $k = 0, \cdots, m$ (stabilité du schéma),
- $\|e_k\| \leq K\Big(\Delta t^p + \|e_0\|\Big)$ (convergence du schéma).

Preuve du théorème de convergence général du schéma à un pas

On choisit $\Delta t^{**} \in]0, \Delta t^*[$ et $\epsilon > 0$ tels que

$$\beta e^{T(M_{2A^*}+1)}(\Delta t^{**})^p \leq A^*/2, \quad e^{TM_{2A^*}}\epsilon \leq A^*/2.$$

On va montrer par récurrence que

$$x_k \in B_{2A^*}$$

et

$$\|e_k\| \leq \beta e^{t_k(M_{2A^*}+1)}\Delta t^p + e^{t_k M_{2A^*}}\|e_0\|.$$

- Pour $k = 0$ on a $x_0 = \bar{x}_0 - e_0$. Comme $\epsilon \leq A^*/2$ on a $\|x_0\| \leq A^* + A^*/2 \leq 2A^*$ et donc $x_0 \in B_{2A^*}$. On a bien sûr aussi $\|e_0\| \leq \beta \Delta t^p + \|e_0\|$ donc la proposition est vraie pour $k = 0$.
- Supposons qu'elle est vérifiée pour tout $k \geq 0$ et montrons qu'elle reste vraie pour $k+1$: on a

$$e_{k+1} = e_k + \Delta t_k \Big(\phi(\bar{x}_k, t_k, \Delta t_k) - \phi(x_k, t_k, \Delta t_k)\Big) + \Delta t_k r_k,$$

Preuve du théorème de convergence général du schéma à un pas (suite)

On a donc

$$\left\{\begin{array}{rl} \|e_{k+1}\| & \leq (1+M_{2A^*}\Delta t_k)\|e_k\| + \Delta t_k \beta \Delta t^p \\ & \leq e^{M_{2A^*}\Delta t_k}\left(\beta e^{t_k(M_{2A^*}+1)}\Delta t^p + e^{t_k M_{2A^*}}\|e_0\|\right) + \Delta t_k \beta \Delta t^p, \\ & \leq \beta e^{t_{k+1}(M_{2A^*}+1)}\Delta t^p + e^{t_{k+1}M_{2A^*}}\|e_0\|, \end{array}\right.$$

en utilisant $1+u \leq e^u$ et

$$\Delta t_k + e^{t_{k+1}(M_{2A^*}+1)-\Delta t_k} \leq (1+\Delta t_k)e^{t_{k+1}(M_{2A^*}+1)-\Delta t_k} \leq e^{t_{k+1}(M_{2A^*}+1)}.$$

Par ailleurs on a

$$\|x_{k+1}\| \leq \|\bar{x}_{k+1}\| + \|e_{k+1}\| \leq A^* + A^*/2 + A^*/2 = 2A^*,$$

ce qui achève la preuve par récurrence.

Etude du schéma d'Euler implicite

Soit $f \in C^1\left(\mathbb{R}^n \times \mathbb{R}, \mathbb{R}^n\right)$ telle que

$$\left(f'_x(x,t)\xi, \xi\right) \leq 0 \text{ pour tout } x, \xi \in \mathbb{R}^n \times \mathbb{R}^n, \text{ et } t \in [0, T].$$

On note $\bar{x} \in C^2\left([0, T_M[, \mathbb{R}^n\right)$ la solution maximale de $x'(t) = f(x(t), t)$, $x(0) = \bar{x}_0$ et on choisit $T \in [0, T_M[$.

Théorème: Sous les hypothèses précédentes, le schéma d'Euler implicite

$$\frac{x_{k+1} - x_k}{\Delta t_k} = f\left(x_{k+1}, t_{k+1}\right), \ k = 0, \cdots, m-1,$$

est bien défini et vérifie

$$\|e_k\|_2 \leq \|e_0\|_2 + \Delta t \int_0^{t_k} \|\bar{x}''(s)\|_2 ds, \ k = 0, \cdots, m.$$

Etude du schéma d'Euler implicite

Preuve de l'existence et unicité de la suite x_k:
On commence par montrer que le système non linéaire à x_k, t_k et $h > 0$ donnés

$$F(x, h) = x - x_k - hf(x, t_k + h) = 0,$$

admet une unique solution pour tout $h \geq 0$.

Unicité: Soit $\varphi(s) = f(x(1-s) + ys, t)$, on a
$\varphi'(s) = f'_x(x(1-s) + ys, t)(y - x)$ et donc

$$\Big(f(y, t) - f(x, t), y - x\Big) = \int_0^1 \Big(f'_x(x(1-u) + yu, t)(y - x), y - x\Big) du \leq 0,$$

pour tout $x, y \in \mathbb{R}^n$ et $t \in [0, T]$. Soient deux solutions x_1 et x_2 de $F(x, h) = 0$, d'après l'inégalité qui précède, elles vérifient nécessairement l'inégalité

$$\Big(x_2 - x_1, x_2 - x_1\Big) \leq 0 \text{ et donc } x_2 = x_1.$$

Etude du schéma d'Euler implicite

Existence: Pour $h = 0$ on a l'unique solution $x = x_k$. Soit

$$I = \{\bar{h}, \text{ tel que } F(x, h) = 0 \text{ admette une solution pour tout } 0 \leq h < \bar{h}\}.$$

Supposons que $H = \sup(I) < +\infty$. On va montrer que $H = \max(I)$. Soit $h_i \in I, i \in \mathbb{N}$, telle que $\lim_{i \to +\infty} h_i = H$ et x_i la solution de $F(x_i, h_i) = 0$. On a

$$\Big(x_i - x_k, x_i\Big) = h_i\Big(f(x_i, t_k + h_i) - f(0, t_k + h_i), x_i\Big) + h_i(f(0, t_k + h_i), x_i\Big)$$

et donc $\|x_i\|_2 \leq \|x_k\|_2 + H \sup_{h \in [0,H]} \|f(0, t_k + h)\|_2$. On déduit que la suite $(x_i)_{i \in \mathbb{N}}$ est bornée et donc, par le théorème de Bolzano Weierstrass, il existe une sous suite $(x_{i_j})_{j \in \mathbb{N}}$ qui converge vers z. Par continuité de f on a donc $F(z, H) = 0$. En notant que $F'_x(z, H) = I - H\ f'_x(z, t_k + H)$ est inversible, le théorème des fonctions implicites montre qu'il existe un voisinage V de H tel que l'équation $F(x, h) = 0$ admette une solution pour tout $h \in V$, et donc on conclut par l'absurbe que $sup(I) = +\infty$.

Preuve de l'estimation: On a

$$\begin{cases} r_k = \frac{\bar{x}(t_{k+1})-\bar{x}(t_k)}{\Delta t_k} - f(\bar{x}(t_{k+1}), t_{k+1}) \\ \quad = \frac{1}{\Delta t_k}\int_{t_k}^{t_{k+1}} \Big(f(\bar{x}(s), s) - f(\bar{x}_{k+1}, t_{k+1})\Big) ds = \frac{1}{\Delta t_k}\int_{t_k}^{t_{k+1}} (\bar{x}'(s) - \bar{x}'(t_{k+1})) ds \\ \quad = \frac{1}{\Delta t_k}\int_{t_k}^{t_{k+1}} \Big((\bar{x}'(s) - \bar{x}'(t_{k+1}))(s - t_k)\Big)' ds - \frac{1}{\Delta t_k}\int_{t_k}^{t_{k+1}} \bar{x}''(s)(s - t_k) ds \end{cases}$$

et donc $\|r_k\|_2 \leq \int_{t_k}^{t_{k+1}} \|\bar{x}''(s)\|_2 ds$.

Montrons par récurrence que $\|e_k\|_2 \leq \|e_0\|_2 + \Delta t \int_0^{t_k} \|\bar{x}''(s)\|_2 ds$ pour tout $k = 0, \cdots, m$.

Pour $k = 0$ c'est immédiat. Supposons que l'estimation est vérifiée pour $k \geq 0$. On a

$$e_{k+1} = e_k + \Delta t_k \Big(f(\bar{x}_{k+1}, t_{k+1}) - f(x_{k+1}, t_{k+1})\Big) + \Delta t_k r_k$$

puis

$$(e_{k+1}, e_{k+1}) = (e_k, e_{k+1}) + \Delta t_k (r_k, e_{k+1}) + \Delta t_k \Big(f(\bar{x}_{k+1}, t_{k+1}) - f(x_{k+1}, t_{k+1}), e_{k+1}\Big)$$

comme le dernier terme est négatif, on a par Cauchy Schwarz l'inégalité

$\|e_{k+1}\|_2 \leq \|e_k\|_2 + \Delta t_k \|r_k\|_2 \leq \|e_0\|_2 + \Delta t \int_0^{t_{k+1}} \|\bar{x}''(s)\|_2 ds$, ce qui achève la preuve par récurrence.