

# Equations Différentielles II



UE11, Mines Paris - PSL\*

29 octobre 2024

## Table des matières

<b>Introduction</b>	<b>2</b>
<b>Objectifs du cours</b>	<b>3</b>
<b>Limites du schéma d'Euler</b>	<b>3</b>
Systèmes raides . . . . .	4
Systèmes hamiltoniens . . . . .	5
<b>Méthodes à un pas</b>	<b>6</b>
Principe . . . . .	6
Exemples . . . . .	6
Définition implicite de $\Phi$ . . . . .	7
<b>Analyse d'erreur</b>	<b>8</b>
Erreur de troncature locale . . . . .	8
Consistance . . . . .	9
Condition nécessaire et suffisante de consistance . . . . .	9
Condition nécessaire et suffisante de consistance d'ordre $\geq p$ . . . . .	11
Stabilité . . . . .	12
Stabilité . . . . .	12
Condition suffisante de stabilité . . . . .	12
Convergence . . . . .	12
Convergence . . . . .	13
Théorème de Lax . . . . .	13
Condition suffisante de convergence . . . . .	13
Erreurs d'arrondi et pas optimal . . . . .	14
<b>Annexe – Choix du pas de temps</b>	<b>14</b>
Pas fixe . . . . .	15
Adaptation du pas de temps . . . . .	15

---

\*Ce document est un des produits du projet  paulinebernard/CDIS issu de la collaboration de (P)auline Bernard (CAS) et (T)homas Romary (GEOSCIENCES). Il dérive du projet  boisgera/CDIS, initié par la collaboration de (S)ébastien Boisgérault (CAOR), (T)homas Romary et (E)milie Chautru (GEOSCIENCES), (P)auline Bernard (CAS), avec la contribution de Gabriel Stoltz (Ecole des Ponts ParisTech, CERMICS). Il est mis à disposition selon les termes de la licence Creative Commons “attribution – pas d'utilisation commerciale – partage dans les mêmes conditions” 4.0 internationale.

<b>Exercices</b>	<b>16</b>
Consistance et ordre de schémas . . . . .	16
Convergence de schémas . . . . .	17
Explicite ou implicite ? . . . . .	17
Euler symplectique . . . . .	17
<b>Corrections</b>	<b>18</b>
Consistance et ordre de schémas . . . . .	18
Convergence de schémas . . . . .	21
Explicite ou implicite ? . . . . .	22
Euler symplectique . . . . .	23
<b>Références</b>	<b>24</b>

## Introduction

Ce chapitre est consacré à la résolution numérique d'équations différentielles

$$\dot{x} = f(t, x) \quad , \quad x(t_0) = x_0 .$$

La nécessité de développer des méthodes d'intégration numériques vient du constat que seule une infime partie des équations différentielles sont résolubles exactement. Or, on a parfois besoin de connaître le plus précisément possible le comportement futur d'un système dynamique :

- soit en temps fini, par exemple pour déterminer la trajectoire d'une fusée pour la mise en orbite d'un satellite;
- soit en temps *long*, par exemple pour déterminer un cycle limite asymptotique (dynamique de population) ou bien se prononcer sur la stabilité de notre système solaire.

La méthode la plus connue est la *méthode d'Euler* datant de 1768, qui consiste à implémenter

$$x^{j+1} = x^j + \Delta t f(t_j, x^j) \quad x^0 = x_0$$

pour un pas de temps  $\Delta t$  suffisamment petit. Cette méthode appartient à la famille des méthodes *explicites*, c'est-à-dire que  $x^{j+1}$  est directement et explicitement défini en fonction de  $x^j$ . En 1824, Cauchy montre la convergence de cette méthode lorsque le pas de temps  $\Delta t$  tend vers 0, et prouve ainsi l'existence et l'unicité des solutions (en fait, il utilise plutôt la version *implicite* de la méthode d'Euler).

Même si la méthode d'Euler suffit dans les cas simples, elle exige parfois de recourir à des pas très faibles pour obtenir une précision acceptable sur des temps longs (voir Systèmes raides (p. 4)). Parfois, le compromis entre précision à chaque itération et accumulation des erreurs d'arrondis devient même impossible. De plus, cette méthode n'est pas adaptée à la simulation de certains systèmes dont certaines propriétés cruciales (comme la conservation de l'énergie) ne sont pas préservées (voir Systèmes Hamiltoniens (p. 5)). Au cours des derniers siècles, les scientifiques ont donc progressivement développé des méthodes de plus en plus complexes et performantes : schémas multi-pas d'ordre supérieur, méthodes implicites, variation du pas, schémas symplectiques etc.

En fait, dans l'histoire des équations différentielles, c'est souvent la mécanique céleste qui a été motrice des plus grandes avancées. Au milieu du XIX<sup>e</sup> siècle, les astronomes Adams et Le Verrier prédisent mathématiquement l'existence et la position de la planète Neptune et l'on entend parler pour la première fois de méthodes multi-pas. Ensuite, les progrès se sont enchaînés au rythme des modèles physiques. La première tendance a été de rechercher des schémas permettant toujours plus de précision à pas plus grand. Parmi les dates clés, on peut citer la publication en 1895 de la première méthode de Runge-Kutta par Runge, puis en 1901, de la populaire méthode de Runge-Kutta d'ordre 4 par Kutta, et ensuite en 1910, de l'*extrapolation de Richardson* permettant la montée en ordre et donc le recours à des pas plus grand pour une même précision. Mais au milieu du XX<sup>e</sup> siècle, on découvre des systèmes, dits *raides* (Hirschfelder, 1952), pour lesquels cette montée en ordre ne suffit pas et pour lesquels il faut repenser de nouveaux schémas (Dalquist, 1968). Enfin, à partir des années 80, les scientifiques développent l'intégration numérique *géométrique*, c'est-à-dire qui préservent les propriétés structurelles du système (symétrie, conservation d'énergie etc.), utile en particulier pour la simulation des systèmes hamiltoniens.

## Objectifs du cours

Ce cours a pour but de sensibiliser aux problèmes apparaissant lors de la simulation numérique des solutions d'équations différentielles, et de donner les bases d'analyse d'erreur numérique. Pour un exposé plus approfondi, on pourra par exemple se référer à (Demailly 2006).

En première lecture :

- comprendre les limites d'un schéma d'Euler.
- comprendre qu'un schéma numérique à un pas consiste à discrétiser une intégrale, en connaître quelques-uns autres que le schéma d'Euler.
- comprendre les notions de consistance/convergence d'un schéma et leur ordre. Savoir montrer que le schéma d'Euler explicite est convergent d'ordre 1.

En deuxième lecture :

- comprendre comment fonctionne un schéma implicite et comment l'implémenter.
- comprendre que la convergence est la combinaison de deux concepts : la consistance et la stabilité ; savoir utiliser ces notions pour évaluer l'impact des erreurs d'arrondi.
- savoir calculer l'ordre de consistance et montrer la convergence de schémas de base.
- comprendre l'apport de schémas symplectiques pour les systèmes hamiltoniens.

## Limites du schéma d'Euler

La première limite du schéma d'Euler est qu'il est d'ordre 1, c'est-à-dire qu'il produit une erreur en  $\Delta t^2$  à chaque pas. Nous verrons dans la suite des algorithmes

d'ordre supérieur qui permettent d'utiliser un pas plus grand pour une précision donnée. Mais au delà de cette problématique, il existe des systèmes pour lesquels de telles méthodes (même d'ordre supérieur) échouent. En voici deux exemples célèbres.

## Systèmes raides

La dénomination *systèmes raides* a été introduite en 1952 par Hirschfelder pour désigner des systèmes comprenant des dynamiques aux constantes de temps très différentes. Dans ce cas, le pas nécessaire pour simuler avec précision les dynamiques très rapides est si petit, qu'il est alors impossible de simuler assez longtemps pour observer les parties lentes. La particularité de ces systèmes est que cette décroissance du pas apparaît alors que la solution est parfaitement régulière, et non pas proche de singularités. C'est le cas des systèmes linéaires

$$\dot{x} = Ax + b$$

avec  $A$  de Hurwitz quand le rapport entre les parties réelles maximales et minimales des valeurs propres devient très grand. Ce phénomène peut notamment apparaître dans un simple système masse/ressort

$$m\ddot{y} = -\rho\dot{y} - ky$$

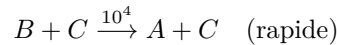
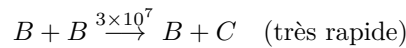
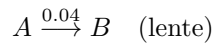
qui se met sous la forme précédente avec  $x = (y, \dot{y}) \in \mathbb{R}^2$  et  $A = \begin{pmatrix} 0 & 1 \\ -\frac{k}{m} & -\frac{\rho}{m} \end{pmatrix}$ .

Lorsque les valeurs propres sont réelles (i.e.  $\rho > 2\sqrt{mk}$ ), leur rapport est donné par

$$\frac{1 + \sqrt{1 - 4\frac{mk}{\rho^2}}}{1 - \sqrt{1 - 4\frac{mk}{\rho^2}}}$$

qui explose lorsque  $\frac{mk}{\rho^2}$  tend vers 0. Par exemple, lorsque les frottements sont très grands par rapport à la raideur du ressort, ou bien lorsque  $\rho$  et  $k$  sont du même ordre de grandeur et très grands.

Plus généralement, la coexistence de dynamiques très lentes à très rapides apparaît en cinétique chimique ou en biologie. La réaction de Robertson (1966)



modélisée par

$$\dot{x}_a = -0.04x_b + 10^4x_bx_c$$

$$\dot{x}_b = 0.04x_a - 10^4x_bx_c - 3 \times 10^7x_b^2$$

$$\dot{x}_c = 3 \times 10^7x_b^2$$

en est un exemple classique, souvent utilisée pour tester les schémas numériques. Il s'avère que pour ces systèmes, des schémas dits *implicites* performant beaucoup

mieux car ils autorisent l'utilisation de pas plus grands pour une même précision et plus de stabilité (voir l'exercice *Explicite ou Implicite?* (p. 17)). Pour plus de détails voir (Hairer and Wanner 1996). L'impossibilité de trouver un pas approprié avec un schéma d'Euler explicite pour ce système est aussi illustré dans le notebook Equations Differentielles II.ipynb.

## Systèmes hamiltoniens

La mécanique hamiltonienne permet typiquement de modéliser le comportement de systèmes dont une certaine énergie est conservée au cours du temps. Il peut s'agir par exemple de planètes en interaction gravitationnelle, de particules en interaction électromagnétique, etc.

Par exemple, dans un problème à  $N$  corps en interaction gravitationnelle, l'hamiltonien s'écrit <sup>1</sup>

$$H(q, p) = \sum_{i=1}^N \frac{1}{2m_i} p_i^\top p_i - \sum_{1 \leq i < k \leq N} G \frac{m_i m_k}{\|q_i - q_k\|}$$

où  $q_i \in \mathbb{R}^3$  désigne la position de chaque corps,  $m_i$  sa masse, et  $p_i = m_i \dot{q}_i \in \mathbb{R}^3$  sa quantité de mouvement. Le comportement de chaque corps est alors régi par la dynamique hamiltonienne <sup>2</sup>

$$\begin{aligned} \dot{q}_i &= \nabla_{p_i} H(q, p) = \frac{1}{m_i} p_i \\ \dot{p}_i &= -\nabla_{q_i} H(q, p) = -G \sum_{k \neq i} \frac{m_i m_k}{\|q_i - q_k\|^3} (q_i - q_k) \end{aligned}$$

On a alors le long des trajectoires

$$\frac{d}{dt} H(q(t), p(t)) = \langle \nabla_q H(t, q(t), p(t)), \dot{q} \rangle + \langle \nabla_p H(t, q(t), p(t)), \dot{p} \rangle = 0$$

et l'énergie  $H(q, p)$  est donc conservée.

Or, lorsqu'on essaye de simuler le système solaire avec un schéma d'Euler (explicite), l'énergie augmente peu à peu à chaque révolution et les trajectoires sont des spirales divergentes. Avec un schéma d'Euler implicite, Jupiter et Saturne s'effondrent vers le soleil et sont éjectées du système solaire! Même des schémas d'ordre supérieur ne permettent pas de simuler correctement ce système sur des temps "courts" sur l'échelle de temps astronomique (à moins de prendre des pas déraisonnablement petits). En fait, le problème c'est que ces méthodes

1. Pour obtenir l'hamiltonien, on commence par définir le lagrangien  $L(t, q, \dot{q})$ , puis la quantité de mouvement  $p = \nabla_{\dot{q}} L(t, q, \dot{q})$ , et enfin l'hamiltonien  $H(t, q, p)$  est obtenu par transformée de Legendre. Notons que dans ce cas général où  $H$  peut dépendre explicitement du temps (par exemple si de l'énergie est injectée ou prélevée par une action extérieure au système), on a  $\frac{d}{dt} H(t, q(t), p(t)) = \nabla_t H(t, q(t), p(t))$ , donc l'hamiltonien varie selon cet effet extérieur, et n'est plus constant.

2. L'application des lois de Newton donnerait directement

$$m_i a_i = m_i \ddot{q}_i = \sum_{k \neq i} F_k = -G \sum_{k \neq i} \frac{m_i m_k}{\|q_i - q_k\|^2} \frac{(q_i - q_k)}{\|q_i - q_k\|}$$

où  $F_k$  sont les forces de gravitation exercées par chaque corps  $k$  sur le corps  $i$ .

d'intégration ne préservent pas les propriétés structurelles des solutions telles que la conservation de l'énergie. Il faut donc développer des schémas particuliers, appelés *symplectiques*, comme illustré sur un simple oscillateur dans l'exercice *Schéma symplectique* (p. 17). Pour aller plus loin sur ces méthodes, voir (Hairer, Lubich, and Wanner 2010). Un exemple simple de système à deux corps est aussi donné dans le notebook *Equations Différentielles II.ipynb*.

## Méthodes à un pas

### Principe

Pour approximer les solutions d'une équation différentielle sur un intervalle  $[0, T]$ , les méthodes numériques à un pas se basent sur la représentation intégrale

$$x(t) = x_0 + \int_{t_0}^t f(s, x(s)) ds = x_0 + \sum_{j=0}^{J-1} \int_{t_j}^{t_{j+1}} f(s, x(s)) ds$$

où  $t_0 < t_1 < \dots < t_J$  avec  $t_J = T$ . L'idée est d'approximer les intégrales  $\int_{t_j}^{t_{j+1}} f(s, x(s))$  sur des intervalles  $[t_j, t_{j+1}]$  suffisamment petits.

Dans la suite, on note  $x^j$  l'approximation au temps  $t_j$  de la valeur exacte  $x(t_j)$  et  $\Delta t_j = t_{j+1} - t_j$  le  $j$ ème pas de temps. L'idée est de calculer récursivement

$$x^{j+1} = x^j + \Delta t_j \Phi(t_j, x^j, \Delta t_j)$$

où  $\Phi(t_j, x^j, \Delta t_j)$  doit donc approximer

$$\frac{1}{t_{j+1} - t_j} \int_{t_j}^{t_{j+1}} f(s, x(s)) ds .$$

Les différentes méthodes de quadrature, i.e. d'approximation de l'intégrale, peuvent donc être mises à profit. La difficulté ici est que seule la valeur initiale  $f(t_j, x(t_j))$  de  $f$  est connue (ou du moins estimée) à l'itération  $j$ , par  $f(t_j, x^j)$ . On distingue donc les méthodes *explicites* où  $\Phi(t_j, x^j, \Delta t_j)$  est écrite directement explicitement en fonction de la valeur initiale  $x^j$ , et les méthodes *implicites* où cette expression n'est connue qu'implicitement et des étapes intermédiaires de calcul sont nécessaires.

### Exemples

#### 1. Méthodes explicites:

- Euler explicite: l'intégrale est approximée par l'aire d'un rectangle déterminé par la valeur initiale de  $f$  à gauche de l'intervalle, i.e.

$$x^{j+1} = x^j + \Delta t_j f(t_j, x^j) .$$

- méthode de Heun : l'intégrale est approximée par l'aire d'un trapèze déterminé par la valeur initiale de  $f$  et une approximation de sa valeur finale, i.e.,

$$x^{j+1} = x^j + \frac{\Delta t_j}{2} \left( f(t_j, x^j) + f(t_{j+1}, x^j + \Delta t_j f(t_j, x^j)) \right) .$$

— schéma de Runge–Kutta d'ordre 4:

$$\begin{cases} F_1 = f(t_j, x^j) \\ F_2 = f\left(t_j + \frac{\Delta t_j}{2}, x^j + \frac{\Delta t_j}{2} F_1\right) \\ F_3 = f\left(t_j + \frac{\Delta t_j}{2}, x^j + \frac{\Delta t_j}{2} F_2\right) \\ F_4 = f(t_j + \Delta t_j, x^j + \Delta t_j F_3), \end{cases}$$

et on pose

$$x^{j+1} = x^j + \Delta t \frac{F_1 + 2F_2 + 2F_3 + F_4}{6}.$$

2. Méthodes implicites:

— Euler implicite : l'intégrale est approximée par l'aire d'un rectangle déterminé par la valeur finale de  $f$  à droite de l'intervalle, i.e.

$$x^{j+1} = x^j + \Delta t_j f(t_{j+1}, x^{j+1}).$$

— méthode des trapèzes (ou Crank–Nicolson) : l'intégrale est approximée par l'aire du trapèze déterminé par les valeurs initiales et finales de  $f$ , i.e.

$$x^{j+1} = x^j + \frac{\Delta t_j}{2} \left( f(t_j, x^j) + f(t_{j+1}, x^{j+1}) \right).$$

— méthode du point milieu : l'intégrale est approximée par l'aire d'un rectangle déterminé par une approximation de la valeur de  $f$  au milieu de l'intervalle, i.e.

$$x^{j+1} = x^j + \Delta t_j f\left(\frac{t_j + t_{j+1}}{2}, \frac{x^j + x^{j+1}}{2}\right).$$

On peut bien sûr construire des méthodes plus compliquées et plus précises pour des méthodes de Runge–Kutta d'ordre supérieur (explicites ou implicites).

### Définition implicite de $\Phi$

Dans les schémas implicites, l'application  $\Phi$  est définie de manière implicite. Par exemple, pour le schéma d'Euler, on a :

$$\Phi(t_j, x^j, \Delta t_j) = f\left(t_j + \Delta t_j, x^j + \Delta t_j \Phi(t_j, x^j, \Delta t_j)\right).$$

Il faut donc s'assurer que  $\Phi$  est bien définie, c'est-à-dire qu'il existe bien  $x^{j+1}$  tel que

$$x^{j+1} = x^j + \Delta t_j f(t_{j+1}, x^{j+1}).$$

Pour cela, nous pouvons voir  $x^{j+1}$  comme le point fixe de l'application  $F_j$  définie par

$$F_j(x) = x^j + \Delta t_j f(t_{j+1}, x).$$

à  $x^j$ ,  $\Delta t_j$ ,  $t_{j+1}$  fixés. L'existence (et l'unicité) de ce point fixe peut alors être démontrée par le théorème de point fixe de Banach. Si  $x \mapsto f(t_{j+1}, x)$  est Lipschitzienne, c'est-à-dire s'il existe  $L_j$  tel que

$$\|f(t_{j+1}, x_a) - f(t_{j+1}, x_b)\| \leq L_j \|x_a - x_b\| \quad \forall (x_a, x_b) \in \mathbb{R}^n \times \mathbb{R}^n,$$

alors  $F_j : \mathbb{R}^n \rightarrow \mathbb{R}^n$  est contractante pour un pas de temps  $\Delta t_j$  suffisamment petit puisque

$$\|F_j(x_a) - F_j(x_b)\| \leq \Delta t_j L_j \|x_a - x_b\| .$$

Puisque  $\mathbb{R}^n$  est complet, on déduit par le théorème du point fixe que  $x^{j+1}$  existe bien.

En pratique, on peut utiliser la méthode itérative de construction de ce point fixe donnée par la preuve du théorème pour approcher  $x^{j+1}$ . Une stratégie est de partir de la valeur donnée par le schéma d'Euler explicite

$$x^{j,0} = x^j + \Delta t_j f(t_j, x^j)$$

et affiner ensuite par l'algorithme du point fixe en itérant

$$x^{j,k+1} = F(x^{j,k})$$

jusqu'à ce que l'évolution relative

$$\frac{x^{j,k+1} - x^{j,k}}{x^{j,0}}$$

devienne inférieure à un seuil choisi par l'utilisateur. Puisque la suite  $(x^{j,k})_{k \in \mathbb{N}}$  est de Cauchy, on sait que cette algorithme s'arrête en un nombre fini d'itérations.

Un tel schéma est plus lourd en terme de calculs qu'un algorithme explicite mais il apporte en général plus de stabilité et permet souvent d'utiliser un pas plus grand. C'est en particulier utile pour les systèmes raides, comme illustré dans l'exercice *Explicite ou Implicite?* (p. 17).

## Analyse d'erreur

L'objectif de l'analyse d'erreur *a priori* est de donner une estimation de l'erreur commise par la méthode numérique en fonction des paramètres du problème (temps d'intégration, pas de temps, propriétés de  $f$ ). L'idée générale est de remarquer qu'à chaque pas de temps, on commet une erreur d'intégration locale (erreur de troncature dans la discrétisation de l'intégrale, à laquelle s'ajoutent souvent des erreurs d'arrondi), et que ces erreurs locales s'accumulent au fil des pas. Le contrôle de cette accumulation demande l'introduction d'une notion de stabilité adéquate, alors que les erreurs locales sont liées à une notion de consistance. L'alliance de stabilité et de consistance donne une propriété de convergence qui est souhaitée lors de l'implémentation de méthodes numériques.

### Erreur de troncature locale

L'erreur de troncature locale à l'itération  $j$  est l'erreur que l'on commet en une seule itération lors de l'approximation de l'intégrale pour passer de  $x^j$  à  $x^{j+1}$ . C'est donc l'erreur théorique que l'on obtiendrait si l'on appliquait le schéma numérique à la solution exacte  $x(t_j)$ . Elle est ainsi définie comme

$$\eta^{j+1} := \frac{x(t_{j+1}) - x(t_j) - \Delta t_j \Phi(t_j, x(t_j), \Delta t_j)}{\Delta t_j}.$$



### Définition – Consistance

On note  $\Delta t = \max_{0 \leq j \leq J-1} \Delta t_j$  le pas de temps maximal. On dit qu'une méthode numérique est *consistante* si

$$\lim_{\Delta t \rightarrow 0} \left( \max_{1 \leq j \leq J} \|\eta^j\| \right) = 0 ,$$

et qu'elle est *consistante d'ordre  $\geq p$*  s'il existe une constante  $c_s$  telle que, pour tout  $0 \leq j \leq J-1$ ,

$$\|\eta^{j+1}\| \leq c_s (\Delta t_j)^p .$$

Une méthode est donc *consistante d'ordre  $p$*  si elle est consistante d'ordre  $\geq p$ , mais pas  $\geq p+1$ .

### Théorème – Condition nécessaire et suffisante de consistance

Si  $\Phi$  est continue, alors le schéma est consistant si et seulement si

$$\Phi(t, x, 0) = f(t, x) \quad \forall (t, x) \in [0, T] \times \mathbb{R}^n .$$

**Démonstration** Soit  $C$  un ensemble fermé et borné tel que  $x(t) \in C$  pour tout  $t \in [0, T]$ . On note toujours  $\Delta t = \max_{0 \leq j \leq J-1} \Delta t_j$ . Par la représentation intégrale des solutions,

$$x(t_{j+1}) = x(t_j) + \int_{t_j}^{t_{j+1}} f(x(s), s) ds$$

l'erreur de troncature locale s'écrit

$$\eta^{j+1} = \frac{1}{\Delta t_j} \int_{t_j}^{t_{j+1}} \left( f(s, x(s)) - \Phi(t_j, x(t_j), \Delta t_j) \right) ds .$$

Si le schéma est consistant alors cette erreur tend vers 0 lorsque  $\Delta t$  tend vers 0 (pour n'importe quel système et n'importe quelle trajectoire). Or,

$$\lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t_j} \int_{t_j}^{t_j + \Delta t} \left( f(s, x(s)) - \Phi(t_j, x(t_j), \Delta t_j) \right) ds = f(t_j, x(t_j)) - \Phi(t_j, x(t_j), 0)$$

qui doit donc être nul.

Réciproquement, supposons  $\Phi(\cdot, \cdot, 0) = f$ . On doit montrer que l'erreur de consistance tend vers 0 lorsque  $\Delta t$  tend vers 0 (uniformément en  $j = 1, \dots, N$ ). Soit  $\varepsilon > 0$ . Par la continuité de  $\Phi$  et  $f$  et puisque  $\Phi(\cdot, \cdot, 0) = f$ , il existe  $\Delta_1 > 0$  tel que si  $\Delta t \leq \Delta_1$ , alors

$$\|\Phi(t, x, \Delta t_j) - f(t, x)\| \leq \varepsilon \quad \forall (t, x) \in [0, T] \times C \quad \forall j = 1, \dots, N .$$

Donc

$$\|\eta^{j+1}\| \leq \varepsilon + \frac{1}{\Delta t_j} \int_{t_j}^{t_{j+1}} \|f(s, x(s)) - f(t_j, x(t_j))\| ds .$$

Puisque  $s \mapsto f(s, x(s))$  est continue sur l'intervalle fermé et borné  $[0, T]$ , elle y est uniformément continue, donc il existe  $\Delta_2 > 0$  tel que si  $\Delta t \leq \Delta_2$ ,

$$\|f(s, x(s)) - f(t_j, x(t_j))\| \leq \varepsilon \quad \forall s \in [t_j, t_{j+1}] \quad \forall j = 1, \dots, N$$

et donc  $\|\eta^{j+1}\| \leq 2\varepsilon$  pour tout  $j$ . Le schéma est donc bien consistant. ■

**Exemple – Schémas consistants** Reprenons les exemples donnés plus haut.

- Euler explicite :  $\Phi(t, x, \Delta t) = f(t, x)$  indépendamment de  $\Delta t$  donc la condition est trivialement satisfaite.
- Méthode de Heun :  $\Phi(t, x, \Delta t) = \frac{f(t, x) + f(t + \Delta t, x + \Delta t f(t, x))}{2}$  donne bien  $f(t, x)$  si  $\Delta t = 0$ .
- Runge Kutta d'ordre 4 : lorsque  $\Delta t = 0$ ,  $F_1 = F_2 = F_3 = F_4 = f(t, x)$  donc  $\Phi(t, x, 0) = \frac{F_1 + 2F_2 + 2F_3 + F_4}{6} = f(t, x)$ .

De même, la consistance des méthodes implicites s'obtiennent en remarquant que  $x^{j+1} = x^j$  lorsque  $\Delta t = 0$ .

Cette condition suffisante permet donc de prouver facilement le caractère consistant d'un schéma. Cependant, en pratique, on s'intéresse surtout à son ordre de consistance. Pour cela, l'erreur de consistance se calcule souvent par des développements de Taylor des solutions lorsque celles-ci sont suffisamment régulières, et la constante  $c_s$  s'exprime alors comme une borne sur les dérivées des solutions. En fait, on remarque que lorsque  $f$  est continue, la solution est  $C^1$  (par définition de nos solutions). Mais puisque  $\dot{x}(t) = f(t, x(t))$ ,  $\dot{x}$  hérite de la régularité de  $f$ : si  $f$  est  $C^k$  alors les solutions  $x$  sont  $C^{k+1}$ . Le calcul de l'ordre de consistance dans le cas du schéma d'Euler explicite est donné ci-dessous. Pour les autres schémas, voir l'exercice *Consistance de schémas* (p. 16)

**Exemple – Ordre de consistance du schéma d'Euler explicite** L'erreur de troncature s'écrit

$$\eta^{j+1} = \frac{x(t_j + \Delta t_j) - (x(t_j) + \Delta t_j f(t_j, x(t_j)))}{\Delta t_j}.$$

Or, si  $f$  est  $C^1$ , alors  $x$  est  $C^2$  et par application la formule de Taylor avec reste intégral, on a

$$x(t_j + \Delta t_j) = x(t_j) + \Delta t_j f(t_j, x(t_j)) + \Delta t_j^2 \int_0^1 \ddot{x}(t_j + s\Delta t_j)(1-s)ds,$$

en utilisant  $\dot{x}(t_j) = f(t_j, x(t_j))$ . Ceci donne donc

$$\|\eta^{j+1}\| \leq \Delta t_j \int_0^1 \ddot{x}(t_j + s\Delta t_j)(1-s)ds \leq \frac{\Delta t_j}{2} \max_{t \in [t_j, t_{j+1}]} \|\ddot{x}(t)\| \leq \frac{\Delta t_j}{2} \max_{t \in [0, T]} \|\ddot{x}(t)\|.$$

Le schéma d'Euler explicite est donc consistant d'ordre  $\geq 1$  avec

$$c_s = \frac{\max_{t \in [0, T]} \|\ddot{x}(t)\|}{2}.$$

Notons qu'en utilisant  $\dot{x}(t) = f(t, x(t))$ ,

$$\ddot{x}(t) = \partial_t f(t, x(t)) + \partial_x f(t, x(t)) \cdot f(t, x(t)),$$

et on peut exprimer  $c_s$  en fonction de bornes sur  $x$  et sur les dérivées de  $f$ . Plus généralement, en dérivant successivement lorsque  $f$  est  $C^k$ , notons  $f^{[k]} : \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}^n$  la fonction dépendant des dérivées successives de  $f$  telle que

$$x^{(k+1)}(t) = f^{[k]}(t, x(t)).$$

On a alors le théorème suivant, généralisant les calculs précédents.

**Théorème – Condition nécessaire et suffisante de consistance d'ordre  $\geq p$**

Si  $\Phi$  et  $f$  sont  $C^p$  alors le schéma est consistant d'ordre  $\geq p$  si et seulement si

$$\frac{\partial^k \Phi}{\partial \Delta t^k}(t, x, 0) = \frac{1}{k+1} f^{[k]}(t, x) \quad \forall 0 \leq k \leq p-1, \quad \forall (t, x) \in [0, T] \times \mathbb{R}^n.$$

**Démonstration** L'erreur de troncature s'écrit

$$\eta^{j+1} = \frac{x(t_j + \Delta t_j) - \left( x(t_j) + \Delta t_j \Phi(t_j, x(t_j), \Delta t_j) \right)}{\Delta t_j}.$$

Or, si  $f$  est  $C^p$ , alors  $x$  est  $C^{p+1}$  et par application la formule de Taylor avec reste intégral, on a

$$\begin{aligned} x(t_j + \Delta t_j) &= x(t_j) + \sum_{k=1}^p \frac{\Delta t_j^k}{k!} x^{(k)}(t_j) + \frac{\Delta t_j^{p+1}}{p!} \int_0^1 x^{(p+1)}(t_j + s\Delta t_j) (1-s)^p ds \\ &= x(t_j) + \sum_{k=0}^{p-1} \frac{\Delta t_j^{k+1}}{(k+1)!} f^{[k]}(t_j, x(t_j)) + \frac{\Delta t_j^{p+1}}{p!} \int_0^1 x^{(p+1)}(t_j + s\Delta t_j) (1-s)^p ds \end{aligned}$$

Par ailleurs, puisque  $\Phi$  est  $C^p$ ,

$$\Phi(t, x, \Delta t) = \sum_{k=0}^{p-1} \frac{\Delta t^k}{k!} \frac{\partial^k \Phi}{\partial \Delta t^k}(t, x, 0) + \frac{\Delta t^p}{(p-1)!} \int_0^1 \frac{\partial^p \Phi}{\partial \Delta t^p}(t, x, t_j + s\Delta t_j) (1-s)^{p-1} ds$$

Il s'ensuit que

$$\eta^{j+1} = \sum_{k=0}^{p-1} \frac{\Delta t_j^k}{k!} \left[ \frac{1}{k+1} f^{[k]}(t_j, x(t_j)) - \frac{\partial^k \Phi}{\partial \Delta t^k}(t_j, x(t_j), 0) \right] + \Delta t_j^p R_j$$

où  $R_j$  est borné (uniformément en  $\Delta t_j$ ) par continuité de  $\frac{\partial^p \Phi}{\partial \Delta t^p}$  et  $x^{(p+1)}$ .

Maintenant, si le schéma est d'ordre  $\geq p$ , alors  $\frac{\eta^{j+1}}{\Delta t_j^p}$  doit rester borné lorsque  $\Delta t_j$  tend vers 0 et on obtient bien la condition du théorème. Réciproquement, si la condition du théorème est vérifiée, on obtient directement que  $\eta^{j+1}$  est borné en  $\Delta t_j^p$ . ■

**Exemple – Ordre de consistance de schémas** On a vu que si  $f$  est  $C^1$ , le schéma d'Euler explicite est consistant d'ordre  $\geq 1$ . On peut le retrouver ici en appliquant le critère précédent pour  $p = 1$  puisque  $\Phi(\cdot, \cdot, 0) = f$ . Ensuite, si  $f$  est  $C^2$ , on constate que

$$\frac{\partial \Phi}{\partial \Delta t}(t, x, 0) = 0 \neq f^{[1]}(t, x) = \partial_t f(t, x) + \partial_x f(t, x) \cdot f(t, x)$$

donc le schéma d'Euler explicite est consistant d'ordre exactement 1.

Par contre, toujours si  $f$  est  $C^2$ , on constate que l'on a bien pour le schéma de Heun

$$\frac{\partial \Phi}{\partial \Delta t}(t, x, 0) = f^{[1]}(t, x) = \partial_t f(t, x) + \partial_x f(t, x) \cdot f(t, x)$$

donc ce schéma est d'ordre  $\geq 2$ . On peut vérifier que si  $f$  est  $C^3$ , le critère n'est pas vérifié à l'ordre supérieur donc il est consistant d'ordre égal à 2.

## Stabilité

Une fois que l'on a étudié l'erreur locale commise en une itération, on s'intéresse à la manière dont elle va se propager au fur et à mesure des itérations. Pour cela, la notion de stabilité quantifie la robustesse de l'approximation numérique par rapport à l'accumulation des erreurs locales et perturbations.

### Définition – Stabilité

On dit qu'une méthode numérique est *stable* s'il existe une constante  $S(T) > 0$  (indépendante des  $\Delta t_j$ ) telle que, pour toutes suites  $x = \{x^j\}_{1 \leq j \leq J}$  et  $z = \{z^j\}_{1 \leq j \leq J}$  vérifiant

$$\begin{cases} x^{j+1} = x^j + \Delta t_j \Phi(t_j, x^j, \Delta t_j), \\ z^{j+1} = z^j + \Delta t_j \Phi(t_j, z^j, \Delta t_j) + \delta^{j+1}, \end{cases}$$

on ait

$$\max_{1 \leq j \leq J} \|x^j - z^j\| \leq S(T) \left( \|x^0 - z^0\| + \sum_{j=1}^J \|\delta^j\| \right).$$

### Théorème – Condition suffisante de stabilité

Si  $\Phi$  sont Lipschitziennes en  $x$ , c'est-à-dire il existe  $L > 0$  tel que pour tout  $0 \leq j \leq J$ ,

$$\|\Phi(t_j, x_a, \Delta t_j) - \Phi(t_j, x_b, \Delta t_j)\| \leq L \|x_a - x_b\| \quad \forall (x_a, x_b) \in \mathbb{R}^n \times \mathbb{R}^n$$

alors le schéma est stable avec  $S(T) = e^{LT}$ .

**Démonstration** On a alors

$$\|x^{j+1} - z^{j+1}\| \leq \|\delta^{j+1}\| + (1 + \Delta t_j L) \|x^j - z^j\| \leq \|\delta^{j+1}\| + e^{\Delta t_j L} \|x^j - z^j\|$$

puisque  $1 + x \leq e^x$  pour tout  $x \in \mathbb{R}$ . Par récurrence, on montre alors que pour tout  $1 \leq j \leq J$ ,

$$\|x^j - z^j\| \leq e^{(t_j - t_0)L} \|x^0 - z^0\| + \sum_{k=1}^j e^{(t_j - t_k)L} \|\delta^k\|.$$

Il s'ensuit que

$$\|x^j - z^j\| \leq e^{TL} \left( \|x^0 - z^0\| + \sum_{k=1}^j \|\delta^k\| \right),$$

ce qui donne le résultat. ■

## Convergence

La combinaison de consistance et de stabilité donne une propriété dite de *convergence* qui dit que l'erreur commise par le schéma par rapport à la vraie solution converge vers 0 lorsque le pas de temps converge vers 0. C'est une propriété cruciale pour un schéma numérique.

### Définition – Convergence

Soit  $\Delta t = \max_{0 \leq j \leq J-1} \Delta t_j$ . Un schéma numérique est *convergent* si

$$\lim_{\Delta t \rightarrow 0} \max_{1 \leq j \leq J} \|x^j - x(t_j)\| = 0$$

lorsque  $x^0 = x(t_0)$ . S'il existe  $p \in \mathbb{N}_{>0}$  et  $c_v > 0$  (indépendent de  $\Delta t$ ) tel que

$$\max_{1 \leq j \leq J} \|x^j - x(t_j)\| \leq c_v (\Delta t)^p$$

on dit que le schéma est *convergent à l'ordre  $p$* .

### Théorème – Théorème de Lax

Une méthode stable et consistante (à l'ordre  $p$ ) est convergente (à l'ordre  $p$ ).

**Démonstration** Notons  $z^j = x(t_j)$ . On remarque que

$$z^{j+1} = z^j + \Delta t_j \Phi(t_j, z^j, \Delta t_j) + \Delta t_j \eta^{j+1},$$

où  $\eta$  est l'erreur de consistance. D'après la propriété de stabilité, on a donc

$$\|x^j - x(t_j)\| \leq S(T) \sum_{j=1}^J \Delta t_{j-1} \|\eta^j\|,$$

et par consistance

$$\|x^j - x(t_j)\| \leq S(T) c_s \sum_{j=1}^J \Delta t_{j-1} (\Delta t_{j-1})^p \leq c_s S(T) T (\Delta t)^p.$$

■

### Théorème – Condition suffisante de convergence

L'inconvénient du théorème de Lax est qu'il faut prouver la stabilité pour obtenir la convergence. Or la seule condition suffisante dont nous disposons à cet effet, est le caractère globalement Lipschitzien de  $x \mapsto \Phi(t, x, \Delta t)$ . Mais il s'agit d'une condition très forte. En fait, il est possible de prouver la convergence sous la condition plus faible que  $x \mapsto \Phi(t, x, \Delta t)$  est "localement Lipschitzienne" :

Si

1. le schéma est consistant d'ordre  $p$ ,
2. pour tout boule fermée  $B$  de  $\mathbb{R}^n$ , il existe  $L > 0$ ,  $\Delta t_m > 0$  tels que pour tout  $t \in [0, T]$  et pour tout  $\Delta t \in [0, \Delta t_m]$ ,

$$\|\Phi(t_j, x_a, \Delta t_j) - \Phi(t_j, x_b, \Delta t_j)\| \leq L \|x_a - x_b\| \quad \forall (x_a, x_b) \in B \times B$$

Alors il existe un pas de temps maximal  $\Delta t_{\max} > 0$  tel que le schéma est convergent d'ordre  $p$ .

L'hypothèse 2. est en particulier vérifiée si  $x \mapsto \Phi(t, x, \Delta t)$  est  $C^1$  d'après une version un peu plus générale du théorème des accroissements finis.

**Exemple – Convergence du schéma d’Euler explicite** On a déjà montré que le schéma d’Euler explicite est consistant d’ordre 1. Par ailleurs, si  $f$  est  $C^1$  par rapport à  $x$ , alors  $x \mapsto \Phi(t, x) = f(t, x, \Delta t)$  est  $C^1$ . Donc d’après le théorème précédent, le schéma est convergent d’ordre 1.

## Erreurs d’arrondi et pas optimal

A chaque itération, lorsque la machine calcule  $x^{j+1}$ , elle commet en plus de l’erreur de troncature de l’intégrale des erreurs d’arrondi de l’ordre de la précision machine. La solution obtenue est donc en fait donnée par

$$\hat{x}^{j+1} = \hat{x}^j + \Delta t_j (\Phi(t_j, \hat{x}^j, \Delta t_j) + \rho^{j+1}) + \varepsilon^{j+1}$$

au lieu de

$$x^{j+1} = x^j + \Delta t_j \Phi(t_j, x^j, \Delta t_j) ,$$

où  $\rho$  modélise l’erreur commise sur le calcul de  $\Phi$  et  $\varepsilon$  l’erreur sur l’addition finale. La stabilité nous donne alors l’écart

$$\max_{0 \leq j \leq J} \|x^j - \hat{x}^j\| \leq S(T) \sum_{j=1}^J \Delta t_{j-1} \|\rho^j\| + \|\varepsilon^j\| .$$

En considérant une borne  $\varepsilon$  des  $\varepsilon^j$  et  $\rho$  des  $\rho^j$ , on obtient

$$\max_{0 \leq j \leq J} \|x^j - \hat{x}^j\| \leq S(T)(T\rho + J\varepsilon) \leq S(T)T \left( \rho + \frac{\varepsilon}{\min \Delta t_j} \right) ,$$

et donc finalement, en supposant l’algorithme convergent d’ordre  $p$ ,

$$\begin{aligned} \max_{0 \leq j \leq J} \|x(t_j) - \hat{x}^j\| &\leq \max_{0 \leq j \leq J} \|x(t_j) - x^j\| + \|x^j - \hat{x}^j\| \\ &\leq c_v (\max_j \Delta t_j)^p + S(T)T \left( \rho + \frac{\varepsilon}{\min_j \Delta t_j} \right) . \end{aligned}$$

Les paramètres  $\varepsilon$  et  $\rho$  sont typiquement petits de l’ordre d’un facteur de la précision machine. Cependant, on voit que plus le pas de temps décroît, plus il y a d’itérations et plus les erreurs d’arrondi se propagent. D’un autre côté, plus il augmente, plus les erreurs de quadrature augmentent. En supposant le pas constant, il y a donc un pas “optimal” donné par

$$\Delta t_{opt} = \left( \frac{S(T)T\varepsilon}{c_v p} \right)^{\frac{1}{p+1}} .$$

## Annexe – Choix du pas de temps

Jusqu’à présent, on a présenté des schémas dépendant de pas de temps  $\Delta t_j$ , sans jamais dire comment les choisir. Le plus simple est de choisir un pas  $\Delta t$  fixe mais il est difficile de savoir à l’avance quel pas est nécessaire. En particulier, comment savoir si la solution obtenue est suffisamment précise, sans connaître la vraie ?

## Pas fixe

Une voie empirique est de fixer un pas, lancer la simulation, puis fixer un pas plus petit, relancer la simulation, jusqu'à ce que les résultats *ne semble plus changer* (au sens de ce qui nous intéresse d'observer). Notons que la connaissance des constantes de temps présentes dans le système peut aider à fixer un premier ordre de grandeur du pas. On pourrait aussi directement choisir le pas  $\Delta t_{opt}$  obtenu plus haut en prenant en compte les erreurs d'arrondis. Mais les constantes  $c_v$  et  $S(T)$  sont souvent mal connues et conservatives.

## Adaptation du pas de temps

Les méthodes à pas fixe exploitent la convergence des schémas, mais

- on ne peut pas prendre un pas de temps arbitrairement petit car on est contraint par le temps de simulation.
- on n'a aucune idée de l'erreur commise et on n'est jamais sûr d'avoir la bonne solution.
- l'utilisation d'un pas très petit peut n'être nécessaire qu'autour de certains points *sensibles* (proches de singularités par exemple) et consomme des ressources inutiles ailleurs.

L'idée serait donc plutôt d'adapter la valeur du pas  $\Delta t_j$  à chaque itération. En d'autres termes, on se fixe une tolérance d'erreur que l'on juge acceptable et on modifie le pas de temps en ligne, selon si l'on estime être au-dessus ou en-dessous du seuil d'erreur. Mais cela suppose d'avoir une idée de l'erreur commise... Il existe justement des moyens de l'estimer.

Tout d'abord, de quelle erreur parle-t-on ?

- erreur *globale* ? L'idéal serait de contrôler  $\max_{0 \leq j \leq N} \|x^j - x(t_j)\|$ . Or la stabilité nous dit que

$$\max_{0 \leq j \leq N} \|x^j - x(t_j)\| \leq S(T) \sum_{j=1}^J \Delta t_{j-1} \|\eta^j\|$$

avec  $\eta^j$  les erreurs de consistances locales. Donc si on se fixe une tolérance sur l'erreur globale  $\text{To1}_g$ , on a

$$\|\eta^j\| \leq \frac{\text{To1}_g}{TS(T)} \implies \max_{0 \leq j \leq N} \|x^j - x(t_j)\| \leq \text{To1}_g .$$

En d'autre termes,  $\text{To1}_g$  nous fixe une erreur maximale *locale* sur  $\eta^j$ , à chaque itération. Notons cependant que cette borne ne prend pas en compte la propagation des erreurs d'arrondis : plus  $\Delta t$  diminue, plus l'erreur globale risque d'augmenter. Ce phénomène devrait donc en toute rigueur aussi nous donner un pas de temps minimal  $\Delta t_{\min}$ . Notons que tous ces calculs dépendent des constantes  $c_v$  et  $S(T)$  qui sont souvent mal connues ou très conservatives.

- erreur (absolue) *locale* ? A chaque itération, une erreur locale est commise due à l'approximation de l'intégrale. Cette erreur est donnée par

$$e^{j+1} = \tilde{x}(t_{j+1}) - x^{j+1} = \left( x^j + \int_{t_j}^{t_{j+1}} f(s, \tilde{x}(s)) ds \right)$$

où  $\tilde{x}$  est la solution de  $\dot{x} = f(t, x)$  qui serait initialisée à  $x^j$  au temps  $t_j$ . Notons que si on avait  $x^j = x(t_j)$ , on aurait exactement  $e^{j+1} = \Delta t_j \eta^{j+1}$ , où  $\eta^{j+1}$  est l'erreur de consistance. On se donne donc une tolérance d'erreur locale

$$\|e^{j+1}\| \leq \text{Tol}_{abs} .$$

— erreur *relative* ? Fixer une erreur absolue est parfois trop contraignant et n'a de sens que si les solutions gardent un certain ordre de grandeur. En effet, l'erreur acceptable quand la solution vaut 1000 n'est peut-être pas la même que lorsqu'elle vaut 1. On peut donc plutôt exiger une certaine erreur relative  $\text{Tol}_{rel}$ , i.e.,

$$\frac{\|e^{j+1}\|}{\|x^j\|} \leq \text{Tol}_{rel} .$$

En général, les solvers assurent (approximativement)

$$\|e^{j+1}\| \leq \text{Tol}_{abs} + \text{Tol}_{rel} \|x^j\| .$$

Par défaut, dans les solvers de Numpy,  $\text{Tol}_{abs} = 10^{-6}$  et  $\text{Tol}_{rel} = 10^{-3}$ .

Mais pour cela nous devons trouver un moyen d'estimer l'erreur locale. C'est souvent fait en utilisant une même méthode à deux pas différents (par exemple  $\Delta t_j$  et  $\Delta t_j/2$ ), ou bien en imbriquant des schémas de Runge-Kutta d'ordres différents.

En fait, il est possible de montrer (exercice) que si  $f$  est  $C^1$ , on a pour un schéma d'Euler explicite

$$\|e^{j+1}\| = \Delta t_j \frac{\|f(t_{j+1}, x^{j+1}) - f(t_j, x^j)\|}{2} + o(\Delta t_j^2)$$

On peut donc estimer à chaque itération l'erreur commise  $e^{j+1}$  et adapter le pas selon si celle-ci est inférieure ou supérieure au seuil de tolérance. En effet, puisque l'on sait par ailleurs que  $e^{j+1} = O(\Delta t_j^2)$ , une possible stratégie d'adaptation est de prendre

$$\Delta t_{new} = \Delta t_j \sqrt{\frac{\text{Tol}_{abs}}{\|e^{j+1}\|}}$$

(éventuellement avec une marge de sécurité)

La fonction correspondante

```
def solve_euler_explicit_variable_step(f, x0, t0, tf, dtmin, dtmax, atol):
    ...
    return t, x
```

est fournie dans le notebook Equations Differentielles II.ipynb.

## Exercices

### Consistance et ordre de schémas

Supposons  $f$  de classe  $C^2$ . Montrer que :



**Question 1** le schéma de Heun est consistant d'ordre  $\geq 2$  et égal à 2 si  $f$  est  $C^3$ . (Solution p. 19.)

**Question 2** le schéma d'Euler implicite est consistant d'ordre  $\geq 1$  (Solution p. 19.)

**Question 3** la méthode des trapèzes est consistante d'ordre  $\geq 2$ . (Solution p. 20.)

**Question 4** le schéma du point milieu est consistant d'ordre  $\geq 2$ . (Solution p. 21.)

**Question 5** le schéma de Runge-Kutta d'ordre 4 est bien consistant d'ordre 4 si  $f$  est  $C^5$ . (Solution p. 21.)

On supposera le pas suffisamment petit pour que les schémas implicites soient définis.

## Convergence de schémas

Sous l'hypothèse que  $f$  est  $C^1$ , montrer que les schémas de Heun et d'Euler implicite sont convergents. (Solution p. 21.)

## Explicite ou implicite ?

**Question 1** Comparer les performances des schémas d'Euler implicites et explicites à pas fixe dans le cas de  $\dot{x} = -\lambda x$ ,  $x(0) = 1$ , et  $\dot{x} = \lambda x$ ,  $x(0) = 1$ , sur un horizon de temps  $T$  donné. (Solution p. 22.)

**Question 2** Lorsqu'on modélise des systèmes chimiques ou biologiques, on obtient souvent des réactions aux constantes de temps très différentes. Vaut-il mieux utiliser un schéma d'Euler implicite ou explicite pour simuler

$$\dot{x} = \begin{pmatrix} -1 & 0 \\ 0 & -\mu \end{pmatrix} x$$

avec  $\mu \gg 1$  ? (Solution p. 23.)

## Euler symplectique

Pour  $\omega > 0$  donné, considérons le système

$$\dot{x}_1 = x_2 \quad , \quad \dot{x}_2(t) = -\omega^2 x_1$$

de condition initiale  $x(0) = (1, 0)$ . On rappelle que pour une suite de la forme  $x^{j+1} = Ax^j$  converge vers 0 si les valeurs propres de  $A$  sont à l'intérieur du cercle unité et diverge si au moins une valeur propre est à l'extérieur.

**Question 1** Montrer que pour n'importe quel pas  $\Delta t$  fixé, un schéma d'Euler explicite donne une solution divergente, et un schéma d'Euler implicite donne une solution qui converge vers 0. Lequel a raison ? (Solution p. 23.)

On définit maintenant le schéma suivant qui “mélange” les schémas d'Euler implicites et explicites :

$$\begin{aligned}x_1^{j+1} &= x_1^j + \Delta t x_2^j \\x_2^{j+1} &= x_2^j - \Delta t \omega^2 x_1^{j+1}\end{aligned}$$

**Question 2** Montrer que la quantité  $\omega^2 x_1^2 + x_2^2 + \Delta t \omega^2 x_1 x_2$  est conservée. Quelle est alors la forme des solutions obtenues dans le plan de phase si  $\omega \Delta t < 2$  ? En déduire la pertinence de ce schéma. On parle de schéma *symplectique*, car il conserve les volumes. (Solution p. 23.)

**Question 3** En écrivant le schéma sous la forme  $x^{j+1} = Ax^j$ , montrer qu'il diverge par contre si  $\Delta t \omega > 2$ . (Solution p. 23.)

**Question 4** Plus généralement, proposer un schéma pour simuler un système Hamiltonien du type

$$\begin{aligned}\dot{q} &= \nabla_p H(q, p) \\ \dot{p} &= -\nabla_q H(q, p)\end{aligned}$$

où  $(q, p) \in \mathbb{R}^N \times \mathbb{R}^N$  sont les positions généralisées et quantités de mouvement,  $H$  est le Hamiltonien que l'on pourra vérifier être conservé le long des trajectoires. (Solution p. 24.)

A noter que les conclusions de cet exercice sont les mêmes si l'on utilise un schéma d'Euler implicite sur la première composante et un schéma d'Euler explicite sur la deuxième. Ces deux schémas s'appellent respectivement Euler symplectique A et B.

## Corrections

### Consistance et ordre de schémas

Vu que  $f$  est  $C^2$ , la dérivée seconde (en temps) des solutions s'écrit

$$\begin{aligned}\ddot{x}(t) &= \frac{d}{dt} \left( f(t, x(t)) \right) = \partial_t f(t, x(t)) + \partial_x f(t, x(t)) \dot{x}(t) \\ &= \partial_t f(t, x(t)) + \partial_x f(t, x(t)) f(t, x(t)).\end{aligned}$$

On a donc

$$f^{[1]}(t, x) = \partial_t f(t, x) + \partial_x f(t, x) f(t, x)$$

et la formule de Taylor le long des solutions donne

$$\begin{aligned}x(t_{j+1}) &= x(t_j) + \Delta t f(t_j, x(t_j)) \\ &\quad + \frac{\Delta t^2}{2} \left( \partial_t f(t_j, x(t_j)) + \partial_x f(t_j, x(t_j)) f(t_j, x(t_j)) \right) + O(\Delta t^3).\end{aligned}\tag{1}$$

Pour les calculs de consistance, deux options possibles:

- soit on compare à la main (1) aux développements en puissances de  $\Delta t$  de la solution numérique  $x^{j+1}$ , partant de  $x(t_j)$ .
- soit on utilise la condition nécessaire et suffisante de consistance (p. 11).

**Question 1** A la main, pour la méthode de Heun,

$$\eta^{j+1} = \frac{1}{\Delta t} \left( x(t_{j+1}) - x(t_j) - \frac{\Delta t}{2} \left[ f(t_j, x(t_j)) + f(t_{j+1}, x(t_j) + \Delta t f(t_j, x(t_j))) \right] \right).$$

Or,

$$\begin{aligned} f(t_{j+1}, x(t_j) + \Delta t f(t_j, x(t_j))) &= f(t_j, x(t_j)) + \partial_t f(t_j, x(t_j))(t_{j+1} - t_j) + \partial_x f(t_j, x(t_j))\Delta t f(t_j, x(t_j)) + O(\Delta t^2) \\ &= f(t_j, x(t_j)) + \Delta t \left( \partial_t f(t_j, x(t_j)) + \partial_x f(t_j, x(t_j))f(t_j, x(t_j)) \right) + O(\Delta t^2). \end{aligned}$$

On en déduit que  $\eta^{j+1} = O(\Delta t^2)$  en utilisant (1). Donc on a une consistance d'ordre  $\geq 2$ .

Sinon il suffit de constater que

$$\Phi(t, x, 0) = f(t, x) \quad , \quad \frac{\partial \Phi}{\partial \Delta t}(t, x, 0) = \frac{1}{2} f^{[1]}(t, x) .$$

Par contre, si  $f$  est  $C^3$ ,  $\frac{\partial^2 \Phi}{\partial \Delta t^2}(t, x, 0) \neq \frac{1}{3} f^{[2]}(t, x)$  donc le schéma n'est pas d'ordre 3 et donc il est d'ordre égal exactement à 2.

**Question 2** Pour le schéma d'Euler implicite

$$\eta^{j+1} = \frac{x(t_{j+1}) - x(t_j) - \Delta t f(t_{j+1}, x^{j+1})}{\Delta t},$$

où  $x^{j+1}$  est solution de

$$x^{j+1} = x(t_j) + \Delta t f(t_{j+1}, x^{j+1}).$$

Comme vu dans la section Définition implicite de  $\Phi$  (p. 7), cette définition implicite de  $x^{j+1}$  admet une unique solution pour  $\Delta t$  suffisamment petit si  $f$  est Lipschitzienne par rapport à  $x$ . On aimerait dire ici que  $x^{j+1} = x(t_j) + O(\Delta t)$  pour  $\Delta t$  suffisamment petit. Pour faire ça proprement, fixons  $(t_j, x(t_j))$  et posons

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

qui est de classe  $C^1$ . On a alors

$$F(x(t_j), 0) = 0 \quad , \quad F(x^{j+1}, \Delta t) = 0 .$$

Puisque  $\partial_x F(x(t_j), 0) = \text{Id}$  est inversible, le théorème des fonctions implicites nous dit que pour  $\Delta t$  suffisamment petit, il existe une fonction  $\psi$  de classe  $C^1$  telle que

$$F(x^{j+1}, \Delta t) = 0 \quad , \quad x^{j+1} = \psi(\Delta t)$$

au voisinage de  $(x(t_j), 0)$ . Puisque  $\psi$  est continue, il s'ensuit donc bien que  $x^{j+1} = x(t_j) + O(\Delta t)$ . On a donc

$$\begin{aligned} f(t_{j+1}, x^{j+1}) &= f(t_j, x(t_j)) + \partial_t f(t_j, x(t_j))(t_{j+1} - t_j) + \partial_x f(t_j, x(t_j))(x^{j+1} - x(t_j)) + O(\|h\|^2) \\ &= f(t_j, x(t_j)) + O(\Delta t) \end{aligned}$$

avec l'incrément  $h = (t_{j+1} - t_j, x^{j+1} - x(t_j))$  qui vérifie  $\|h\| = O(\Delta t)$ .

Ainsi, toujours au vu de (1),

$$\eta^{j+1} = \frac{1}{\Delta t} \left[ x(t_{j+1}) - \left( x(t_j) + \Delta t f(t_j, x(t_j)) + O(\Delta t^2) \right) \right] = O(\Delta t).$$

**Question 3** Pour la méthode des trapèzes, on a l'erreur

$$\eta^{j+1} = \frac{x(t_{j+1}) - x(t_j) - \frac{\Delta t}{2} (f(t_j, x(t_j)) + f(t_{j+1}, x^{j+1}))}{\Delta t},$$

où  $x^{j+1}$  est solution de

$$x^{j+1} = x(t_j) + \frac{\Delta t}{2} (f(t_j, x(t_j)) + f(t_{j+1}, x^{j+1})).$$

Cette fois-ci on voudrait montrer que  $x^{j+1} = x(t_j) + \Delta t f(t_j, x(t_j)) + O(\Delta t^2)$ . Pour cela, on redéfinit

$$F(x, \Delta t) = x - \left( x(t_j) + \frac{\Delta t}{2} (f(t_j, x(t_j)) + f(t_j + \Delta t, x^{j+1})) \right)$$

qui est de classe  $C^1$ . On a alors

$$F(x(t_j), 0) = 0 \quad , \quad F(x^{j+1}, \Delta t) = 0.$$

Puisque  $\partial_x F(x(t_j), 0) = \text{Id}$  est inversible, le théorème des fonctions implicites nous dit que pour  $\Delta t$  suffisamment petit, il existe une fonction  $\psi$  de classe  $C^1$  telle que

$$F(x^{j+1}, \Delta t) = 0 \quad , \quad x^{j+1} = \psi(\Delta t)$$

au voisinage de  $(x(t_j), 0)$  et de plus,

$$\psi'(0) = \text{Id}^{-1} \cdot \partial_{\Delta t} F(x(t_j), 0) = f(t_j, x(t_j)).$$

Puisque  $\psi$  est de classe  $C^1$ , on a donc bien

$$x^{j+1} = x(t_j) + \Delta t f(t_j, x(t_j)) + O(\Delta t^2).$$

Il s'ensuit que

$$f(t_{j+1}, x^{j+1}) = f(t_j, x(t_j)) + \Delta t (\partial_t f(t_j, x(t_j)) + \partial_x f(t_j, x(t_j)) f(t_j, x(t_j))) + O(\Delta t^2)$$

soit

$$x^{j+1} = x(t_j) + \Delta t f(t_j, x(t_j)) + \frac{\Delta t^2}{2} (\partial_t f(t_j, x(t_j)) + \partial_x f(t_j, x(t_j)) f(t_j, x(t_j))) + O(\Delta t^3)$$

On en déduit donc  $\eta^{j+1} = O(\Delta t^2)$ .

**Question 4** Enfin, pour la méthode du point milieu, on a l'erreur

$$\eta^{j+1} = \frac{x(t_{j+1}) - x(t_j) - \Delta t f\left(t_j + \frac{\Delta t}{2}, \frac{x(t_j) + x^{j+1}}{2}\right)}{\Delta t},$$

où  $x^{j+1}$  est solution de

$$x^{j+1} = x(t_j) + \Delta t f\left(t_j + \frac{\Delta t}{2}, \frac{x(t_j) + x^{j+1}}{2}\right).$$

On montre de la même façon qu'à la question précédente que pour  $\Delta t$  suffisamment petit,  $x^{j+1} = x(t_j) + \Delta t f(t_j, x(t_j)) + O(\Delta t^2)$  et donc

$$\begin{aligned} f\left(t_j + \frac{\Delta t}{2}, \frac{x(t_j) + x^{j+1}}{2}\right) &= f(t_j, x(t_j)) + \partial_t f(t_j, x(t_j)) \frac{\Delta t}{2} + \partial_x f(t_j, x(t_j)) \left(\frac{x(t_j) + x^{j+1}}{2} - x(t_j)\right) + O(\|h\|^2) \\ &= f(t_j, x(t_j)) + \frac{\Delta t}{2} (\partial_t f(t_j, x(t_j)) + \partial_x f(t_j, x(t_j)) f(t_j, x(t_j))) + O(\Delta t^2) \end{aligned}$$

avec l'incrément  $h = \left(\frac{\Delta t}{2}, \frac{x(t_j) + x^{j+1}}{2} - x(t_j)\right)$ , soit

$$x^{j+1} = x(t_j) + \Delta t f(t_j, x(t_j)) + \frac{\Delta t^2}{2} (\partial_t f(t_j, x(t_j)) + \partial_x f(t_j, x(t_j)) f(t_j, x(t_j))) + O(\Delta t^3)$$

On en déduit donc  $\eta^{j+1} = O(\Delta t^2)$ .

**Question 5** On vérifie que le critère est vérifié pour  $0 \leq k \leq 3$  mais pas pour  $k = 4$ !

## Convergence de schémas

Tout d'abord, dans l'exercice précédent, nous avons montré que les schémas de Heun et d'Euler implicite étaient consistants d'ordre 2 et 1 respectivement. Il ne nous reste donc plus qu'à montrer que  $\Phi$  est localement lipschitzienne (ou  $C^1$ ) par rapport à  $x$  pour  $\Delta t$  suffisamment petit, pour en déduire la convergence à l'ordre 2 et 1 respectivement.

Pour le schéma de Heun,

$$\Phi(t, x, \Delta t) = \frac{1}{2} \left( f(t, x) + f\left(t + \Delta t, x + \Delta t f(t, x)\right) \right)$$

donc  $\Phi$  est  $C^1$  par rapport à  $x$  si  $f$  l'est.

Prenons maintenant le schéma d'Euler implicite. Pour  $\Delta t \leq \Delta t_m$ ,  $\Phi$  est définie par

$$\Phi(t, x, \Delta t) = f\left(t + \Delta t, x + \Delta t \Phi(t, x, \Delta t)\right).$$

Soit  $B$  un compact de  $\mathbb{R}^n$ . Soit  $B'$  un compact tel que  $x + \Delta t \Phi(t, x, \Delta t) \in B'$  pour tout  $x \in B$ , tout  $t \in [0, T]$  et tout  $\Delta t \in [0, \Delta t_m]$ . Puisque  $f$  est continue, et  $C^1$  par rapport à  $x$ , il existe  $L_f > 0$  tel que

$$\|f(t, x_a) - f(t, x_b)\| \leq L_f \|x_a - x_b\| \quad \forall (x_a, x_b, t) \in B' \times B' \times [0, T + \Delta t_m].$$

Ceci est vrai par le théorème des accroissements finis appliqué à  $x \mapsto f(t, x)$  et pour  $t$  dans un intervalle fermé et borné. Prenons maintenant  $(x_a, x_b) \in B \times B$ ,  $t \in [0, T]$  et  $\Delta t \in [0, \Delta t_m]$ , alors

$$\begin{aligned} & \|\Phi(t, x_a, \Delta t) - \Phi(t, x_b, \Delta t)\| \\ &= \|f\left(t + \Delta t, x_a + \Delta t \Phi(t, x_a, \Delta t)\right) - f\left(t + \Delta t, x_b + \Delta t \Phi(t, x_b, \Delta t)\right)\| \\ &\leq L_f (\|x_a - x_b\| + \Delta t \|\Phi(t, x_a, \Delta t) - \Phi(t, x_b, \Delta t)\|) \end{aligned}$$

soit

$$\|\Phi(t, x_a, \Delta t) - \Phi(t, x_b, \Delta t)\| \leq \frac{L_f}{1 - L_f \Delta t} \|x_a - x_b\|$$

si  $\Delta t < 1/L_f$ . Donc  $\Phi$  est bien localement lipschitzienne par rapport à  $x$ .

## Explicite ou implicite ?

**Question 1** Prenons d'abord  $\dot{x} = -\lambda x$ ,  $x(0) = 1$ , dont la solution exacte est  $x(t) = e^{-\lambda t}$ .

Le schéma d'Euler explicite donne

$$x^{j+1} = x^j - \lambda \Delta t x^j = (1 - \lambda \Delta t)^j$$

soit

$$x^J = (1 - \lambda \Delta t)^J = (1 - \lambda \Delta t)^{\frac{T}{\Delta t}} = \left( (1 - \lambda \Delta t)^{\frac{T}{\lambda \Delta t}} \right)^\lambda.$$

On a bien

$$\lim_{\Delta t \rightarrow 0} x^J = e^{-\lambda T}.$$

Cependant, il faut  $|1 - \lambda \Delta t| < 1$  pour que la solution converge au moins vers 0. Sinon, pour  $\lambda \Delta t = 2$ ,  $x^J = (-1)^J$ , qui n'a rien à voir avec la solution. Pire, pour  $\lambda \Delta t = 2$ , l'algorithme diverge. Il faut donc adapter  $\Delta t$  à la constante de temps  $\lambda$  du système. Ceci peut poser problème lorsque l'on simule des systèmes sur des temps longs (par rapport à  $\lambda$ )

De l'autre côté, le schéma d'Euler implicite donne

$$x^{j+1} = x^j - \lambda \Delta t x^{j+1}$$

soit

$$x^J = \frac{1}{(1 + \lambda \Delta t)^J} = \frac{1}{(1 + \lambda \Delta t)^{\frac{T}{\Delta t}}}$$

qui tend vers 0 quelque soit le pas  $\Delta t$  ! On parle de stabilité inconditionnelle. Ceci est très pratique pour des simulations sur temps longs, où la condition  $\lambda \Delta t < 1$  est trop contraignante.

Prenons maintenant  $\dot{x} = \lambda x$ ,  $x(0) = 1$ , dont la solution exacte est  $x(t) = e^{\lambda t}$ . Cette fois-ci, Euler explicite donne

$$x^J = (1 + \lambda \Delta t)^{\frac{T}{\Delta t}}$$

qui fait maintenant sens même pour des pas grands. Par contre, Euler implicite donne

$$x^J = \frac{1}{(1 - \lambda \Delta t)^{\frac{T}{\Delta t}}}$$

qui n'est pas défini pour  $\lambda \Delta t = 1$  et qui explose pour des valeurs proche de 1.

**Question 2** Lorsque l'on a deux dynamiques asymptotiquement stables aux constantes de temps très différentes la condition de stabilité de Euler explicite exige de choisir un pas c  l   sur la plus petite constante de temps, i.e. il faut  $\Delta t < \frac{1}{\mu}$ . Ceci est tr  s exigeant car il faut attendre un nombre d'it  rations de l'ordre de  $\mu$  pour voir l'  volution du syst  me lent. Par contre, une m  thode implicite permet de choisir librement le pas de temps en fonction des performances souhait  es.

## Euler symplectique

**Question 1** Dans le cas d'Euler explicite,  $x^{j+1} = Ax^j$  avec

$$A = \begin{pmatrix} 1 & \Delta t \\ -\Delta t \omega^2 & 1 \end{pmatrix}$$

dont les valeurs propres sont  $1 \pm i\omega\Delta t$  de norme  $\sqrt{1 + \Delta t^2 \omega^2} > 1$ . Donc les solutions divergent.

Dans le cas d'Euler implicite,  $x^{j+1} = Ax^j$  avec

$$A = \frac{1}{1 + \Delta t^2 \omega^2} \begin{pmatrix} 1 & \Delta t \\ -\Delta t \omega^2 & 1 \end{pmatrix}$$

dont les valeurs propres sont  $1/(1 \pm i\omega\Delta t)$  de norme  $1/\sqrt{1 + \Delta t^2 \omega^2} < 1$ . Donc les solutions convergent vers 0.

Or on peut v  rifier que le long des vraies solutions, l'  nergie  $\omega^2 x_1^2 + x_2^2$  est constante donc les trajectoires sont born  es et ne peuvent pas converger vers z  ro. Aucun des deux sch  mas n'approxime les solutions correctement sur le long-terme.

**Question 2** Le sch  ma symplectique donne

$$\begin{pmatrix} \omega^2 & 0 \\ \Delta t \omega^2 & 1 \end{pmatrix} \begin{pmatrix} x_1^{j+1} \\ x_2^{j+1} \end{pmatrix} = \begin{pmatrix} \omega^2 & \Delta t \omega^2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1^j \\ x_2^j \end{pmatrix}$$

En multipliant    gauche alternativement par  $\begin{pmatrix} x_1^{j+1} & x_2^{j+1} \end{pmatrix}$  et  $\begin{pmatrix} x_1^j & x_2^j \end{pmatrix}$ , on obtient

$$\begin{aligned} \omega^2 (x_1^{j+1})^2 + (x_2^{j+1})^2 + \Delta t \omega^2 x_1^{j+1} x_2^{j+1} &= \omega^2 x_1^{j+1} x_1^j + x_2^{j+1} x_2^j + \Delta t \omega^2 x_1^{j+1} x_2^j \\ &= \omega^2 (x_1^j)^2 + (x_2^j)^2 + \Delta t \omega^2 x_1^j x_2^j \end{aligned}$$

et donc la quantit    $\omega^2 x_1^2 + x_2^2 + \Delta t \omega^2 x_1 x_2$  est bien constante. Pour  $\omega^2 - \frac{\Delta t^2 \omega^4}{4} > 0$ , soit  $\omega\Delta t < 2$ , cette matrice est d  finie positive, donc les solutions restent sur une ellipse. Cette ellipse se rapproche de la vraie solution lorsque  $\Delta t$  tend vers 0. Ce sch  ma est donc appropri   pour simuler les trajectoires sur un temps long.

**Question 3** L'algorithme symplectique est d  crit par  $x^{j+1} = Ax^j$  avec

$$A = \begin{pmatrix} 1 & \Delta t \\ -\Delta t \omega^2 & 1 - \Delta t^2 \omega^2 \end{pmatrix}$$

dont le polynôme caractéristique s'écrit

$$s^2 - (2 - \Delta t^2 \omega^2)s + 1$$

On a les cas suivants :

- si  $(2 - \Delta t^2 \omega^2)^2 - 4 < 0$ , i.e., si  $\omega \Delta t < 2$ , les valeurs propres sont imaginaires conjuguées et de module 1.
- si  $\omega \Delta t > 2$ , les valeurs propres sont réelles de produit 1, donc l'une est supérieure à 1 est le schéma diverge.
- dans le cas extrême où  $\omega \Delta t = 2$ , il y a une valeur propre double en -1.

**Question 4** Pour un système hamiltonien, on peut donc proposer

$$\begin{aligned} q^{j+1} &= q^j + \Delta t \nabla_p H(q^j, p^j) \\ p^{j+1} &= p^j - \Delta t \nabla_q H(q^{j+1}, p^{j+1}) \end{aligned}$$

ou bien

$$\begin{aligned} q^{j+1} &= q^j + \Delta t \nabla_p H(q^{j+1}, p^{j+1}) \\ p^{j+1} &= p^j - \Delta t \nabla_q H(q^j, p^j) \end{aligned}$$

pour  $\Delta t$  suffisamment petit.

## Références

- Demailly, J.-P. 2006. *Analyse Numérique et équations Différentielles*. EDP Sciences. Grenoble Sciences.
- Hairer, E., C. Lubich, and G. Wanner. 2010. *Geometric Numerical Integration : Structure-Preserving Algorithms for Ordinary Differential Equations*. Edited by Springer Series in Computational Mathematics. 2nd ed. Springer-Verlag, Berlin.
- Hairer, E., and G. Wanner. 1996. *Solving Ordinary Differential Equations II. Stiff and Differential-Algebraic Problems*. Edited by Springer Series in Computational Mathematics. 2nd ed. Springer-Verlag, Berlin.