# Résolution numérique des équations différentielles

Benjamin BOUVIER

sous la direction de Mr Thierry DUMONT

Semestre de printemps 2010

# Table des matières

# Chapitre 1

# Introduction

Tous les phénomènes autour de nous sont des systèmes dynamiques, c'est-à-dire des systèmes interagissant constamment les uns avec les autres ou encore avec l'environnement. Certains de ces systèmes sont régis par des lois physiques et mathématiques que les scientifiques ont tenté ou tentent de comprendre et d'analyser. Pour cela, ils ont recours à des modélisations de ces systèmes, c'est-à-dire de la mise en place de modèles qui permettent à partir des comportements passés des systèmes, d'en prévoir les comportements futurs.

Ainsi, Albert Einstein découvre en 1905, dans le cadre de sa théorie sur la relativité restreinte, une des formules physiques les plus connues jusque maintenant qui à toute masse, associe une énergie : $E = mc^2$. Mais toutes les lois régissant les systèmes dynamiques ne sont pas aussi simples à énoncer et recourent à des modélisations plus compliquées, faisant intervenir des équations différentielles.

Au 18$^{e}$ siècle, Isaac Newton, pour énoncer les lois de la mécanique classique, énonce que la force exercée sur un corps vaut la masse de cet objet multipliée par l'accélération de ce corps : $f = m.a$. Pour le cas d'un mouvement rectiligne horizontal d'un corps se déplaçant en ligne droite, on a alors la relation suivante, si $x$ est l'abscisse du corps :

$$f(x) = m.x''$$

Équation qui décrira entièrement le mouvement du corps, pourvu qu'on sache la résoudre.

La dynamique des populations utilise également beaucoup de modèles avec des équations différentielles. Au cours du même siècle, Thomas Robert Malthus établit la conjecture que les populations augmentent de manière géométrique si les ressources sont illimitées ; ce qui se modélise, si $N(t)$ désigne le nombre d'habitants à un temps $t$ donné, par $N'(t) = \lambda N(t)$, où $N'(t)$ correspond à la variation de population, et $\lambda$ est un paramètre positif. En 1840, Pierre François Verhulst établit un modèle plus réaliste, connu maintenant sous le nom de modèle logistique :

$$N'(t) = \lambda N(t)(1 - \frac{N(t)}{K})$$

où $\lambda$ et $K$ sont des paramètres définissant respectivement la croissance de la population et la "capacité d'accueil" du système (qui est la capacité maximale de population). Ce modèle admet une solution explicite, c'est-à-dire qu'on peut trouver une solution avec une formule de la forme $N(t) = ...$ ; si $y_0$ est la population au temps 0, alors la solution est ici

$$N(t) = K\frac{1}{1 + \left(\frac{K}{y_0} - 1\right)e^{\lambda t}}$$

Beaucoup de modèles ont recours à ces équations différentielles, dans beaucoup de domaines d'application : économie, mécanique des corps ou mécanique céleste, dynamique des populations ou épidémiologie, météorologie,...

Cependant, on ne sait pas toujours résoudre explicitement ces équations et il faut parfois des solutions numériques pour avoir des estimations probantes des perturbations futures du système. Par exemple, une fonction comme $x \mapsto \frac{e^x}{x}$ n'admet pas de primitive *simple*, c'est-à-dire exprimée à l'aide de fonctions *élémentaires* ($cos, sin, log, exp...$). Cette primitive existe bien, mais elle ne peut pas être énoncée *explicitement* : il s'agit d'une conséquence du théorème de Liouville-Rosenlicht, dont les détails ne seront pas vus ici.

Il faut donc des méthodes qui nous permettent de trouver des résultats numériques résolvant ces équations différentielles. Les mathématiciens ont donc trouvé ces méthodes, appelées *schémas* et ont développé toute

une théorie qui montrait que ces schémas calculaient bien des solutions efficientes et avec une précision donnée.

Nous allons donc voir les idées de la théorie générale des équations différentielles, puis la théorie des schémas, avec une présentation de quelques schémas encore utilisés de nos jours.

# Chapitre 2

# Résultats généraux sur les équations différentielles

## 2.1 Théorème d'existence et d'unicité des solutions

### 2.1.1 Théorème de Cauchy-Lipschitz

**Problème de Cauchy**

Augustin-Louis Cauchy, au 19e siècle, souhaite ne pas séparer la recherche des solutions générales d'une équation différentielle de la recherche des solutions particulières. Dans ce cadre, il pose un problème de système différentiel, connu sous le nom de *problème de Cauchy*.

**Définition 1.** On appelle **problème de Cauchy** un système différentiel le problème consistant à étudier les solutions du système :

$$\begin{cases} u'(t) = f(t, u(t)) \\ u(t_0) = u_0 \end{cases} \tag{2.1}$$

où $u(t)$ est l'inconnue à valeurs dans un ouvert de $\mathbb{R}^d$ (la solution recherchée), $t$ est la variable représentant le temps (par exemple), les données sont *l'instant initial* $t_0$, *la condition initiale* $u_0$, et la fonction $f$ de $\mathbb{R} \times \mathbb{R}^d$ dans $\mathbb{R}^d$ définie par $(t, u) \mapsto f(t, u)$.

**Exemple.** Soit le système défini par :

$$\begin{cases} u'(t) = 2u(t) \\ u(0) = 3 \end{cases}$$

Ici, $t_0 = 0, u_0 = 3, f(t, u) = 2u$. On sait résoudre ce système explicitement et la solution unique est $u(t) = 3\exp(2t)$.

**Exemple.** Soit le système défini par :

$$\begin{cases} u'(t) = t + u(t) \\ u(0) = 1 \end{cases}$$

Ici, $t_0 = 0, u_0 = 1, f(t, u) = t + u$. On peut résoudre ce système explicitement assez facilement et montrer que l'unique solution est $u(t) = \exp(\frac{t^2}{2})$. Cependant, si on retire la condition initiale, la solution n'est plus unique : $u(t) = u_0.\exp(\frac{t^2}{2})$.

On constate que dans ces exemples, la solution, si elle existe, est unique. Il s'agit exactement de l'énoncé du théorème d'existence et d'unicité des solutions, portant le nom de *théorème de Cauchy-Lipschitz* dans certains pays et le nom de *théorème de Picard-Lindelöf* dans d'autres. Il s'avère que Leonhard Euler est le premier à avoir trouvé l'existence et l'unicité d'une solution à ce problème, grâce à sa méthode (*méthode d'Euler*) que nous verrons plus en détail par la suite ; cependant Euler ne vérifia pas que sa méthode convergeait, il faudra attendre la première moitié du 19e siècle pour que Cauchy vérifie la convergence de cette méthode.

**Théorème de Cauchy-Lipschitz**

On considère le cas où $d = 1$, c'est-à-dire où les inconnues sont à valeurs dans $\mathbb{R}$. Le système de Cauchy s'écrit alors :

$$\begin{cases} u'(t) = f(t, u(t)) \\ u(t_0) = u_0 \end{cases}$$

avec $u : \mathbb{R} \to \mathbb{R}, I := [t_0, T], f : I \times \mathbb{R} \to \mathbb{R}$.

On cherche à montrer qu'il existe une unique solution, dont la forme générale est donnée par la méthode d'Euler que nous allons détailler ici.

Soit une subdivision $S$ de $I$ formée des points $t_0 \leq t_1 \leq \cdots \leq t_n := T. \forall\, i \in [\![0, n]\!]$, on notera $u_i := u(t_i)$.

On fait l'approximation de $u$, sur chaque sous-intervalle de cette subdivision, par sa série de Taylor au premier ordre :

$$\begin{aligned} u_1 - u_0 &= (t_1 - t_0) f(t_0, u_0) \\ u_2 - u_1 &= (t_2 - t_1) f(t_1, u_1) \\ &\vdots \\ u_n - u_{n-1} &= (t_n - t_{n-1}) f(t_{n-1}, u_{n-1}) \end{aligned}$$

$\forall\, i \in [\![0, n-1]\!]$, on note $h_i := t_{i+1} - t_i$ ; la subdivision est alors notée $h = (h_0, h_1, \ldots, h_{n-1})$.

On peut relier chacun des segments formés les uns avec les autres, on obtient alors le *polygône d'Euler*, d'équation : $u_h(t) = u_i + (t - t_i) f(t_i, u_i)$ si $t \in [t_i, t_{i+1}]$.

On va recourir à deux lemmes avant de pouvoir prouver le théorème :

**Lemme 1.** *Soit $D := \{(t, u) \in \mathbb{R}^2 \,/\, t_0 \leq t \leq T, |u - u_0| \leq b\}$. On suppose que $|f|$ est bornée par $A > 0$ sur $D$. Si $T - t_0 \leq \frac{b}{A}$, alors la solution numérique $(t_i, u_i)$ calculée avec la méthode d'Euler est aussi contenue dans $D$ pour toute subdivision $S$ et on a :*

$$|u_h(t) - u_0| \leq A|t - t_0| \tag{2.2}$$

*De plus, si $|f(t, u) - f(t_0, u_0)| \leq \epsilon$, alors*

$$|u_h(t) - (u_0 + (t - t_0) f(t_0, u_0))| \leq \epsilon |t - t_0| \tag{2.3}$$

*Preuve.* Soient $(t, u) \in D$, soit $i \in [\![0, n]\!] \,/\, t_i \leq t \leq t_{i+1}$. On suppose que $\forall\, j, (t_j, u_j) \in D$. On a

$$\begin{aligned} |u_h(t) - u_0| &= |u_i + (t - t_i) f(t_i, u_i) - u_0| \\ &\leq |u_i - u_0| + |t - t_i| A \end{aligned}$$

par inégalité triangulaire. On écrit alors $u_i - u_0 = u_i - u_{i-1} + u_{i-1} - u_{i-2} \cdots - u_0$, alors en réutilisant l'inégalité triangulaire

$$\begin{aligned} &\leq |u_i - u_{i-1}| + |u_{i-1} - u_{i-2}| + \cdots + |u_1 - u_0| + |t - t_i| A \\ &\leq |t_i - t_{i-1}| A + |t_{i-1} - t_{i-2}| A + \cdots + |t_1 - t_0| A + |t - t_i| A \end{aligned}$$

d'où en factorisant par A et en remarquant que $t_j \leq t_{j+1} \forall\, j$,

$$\begin{aligned} &\leq (t_i - t_{i-1} + t_{i-1} - t_{i-2} + \cdots + t_1 - t_0 + t - t_i) A \\ &= A|t - t_0| \end{aligned}$$

La deuxième inégalité se trouve également avec l'inégalité triangulaire. On suppose

$$\forall\, j, |f(t_j, u_j) - f(t_0, u_0)| \leq \epsilon$$

On écrit d'abord

$$t - t_0 = t - t_i + t_i - t_{i-1} + \cdots - t_0$$

On note $E := |u_h(t) - (u_0 + (t-t_0)f(t_0,u_0))|$. Alors on obtient :

$$\begin{aligned}
E &= |u_i - u_0 + (t-t_i)f(t_i,u_i) - (t-t_i)f(t_0,u_0) - \cdots - (t_1-t_0)f(t_0,u_0)| \\
&= |(t-t_i)f(t_i,u_i) + (t_i-t_{i-1})f(t_{i-1},u_{i-1}) + \cdots + (t_1-t_0)f(t_0,u_0) - (t-t_i)f(t_0,u_0) - \cdots - (t_1-t_0)f(t_0,u_0)| \\
&= |(t-t_i)(f(t_i,u_i) - f(t_0,u_0) + (t_i-t_{i-1})(f(t_{i-1},u_{i-1}) - f(t_0,u_0)) + \cdots + (t_1-t_0)(f(t_1,u_1) - f(t_0,u_0))| \\
&\leq |t_i - t_{i-1}|\epsilon + \cdots + |t_1 - t_0|\epsilon \\
&= (t-t_0)\epsilon
\end{aligned}$$

L'équation (2.2) nous montre que pour $A(t-t_0) \leq b$, on a que le polygône d'Euler reste dans D. □

Le problème suivant consiste à donner une estimation de la variation de $u_h(t)$ quand la condition initiale $u_0$ est modifiée. Soit $v_0$ une autre condition initiale, on considère alors $v_1 - v_0 = (t_1 - t_0)f(t_0, v_0)$ de la même manière que pour la condition initiale $u_0$.

Il s'agit d'estimer la différence entre les deux nouvelles valeurs calculées : $|v_1 - u_1|$. Au vu des définitions de $v_1 - v_0$ et de $u_1 - u_0$, on obtient alors :

$$v_1 - u_1 = v_0 - u_0 + (t_1 - t_0)(f(t_0, v_0) - f(t_0, u_0))$$

Si on rajoute la condition suivante :

$$\exists\, L \geq 0 \,/\, \forall\, (x,y,z) \in \mathbb{R}^3, |f(x,y) - f(x,z)| \leq L|y-z| \tag{2.4}$$

Alors on obtient que $|v_1 - u_1| \leq |v_0 - u_0| + L|t_1 - t_0||v_0 - u_0|$, et comme $\forall\, x \in \mathbb{R}, 1 + x \leq \exp x$, on a par suite :

$$|v_1 - u_1| \leq |v_0 - u_0| \exp(L|t_1 - t_0|)$$

Ce qui nous amène au lemme suivant :

**Lemme 2.** *Soit $S$ une subdivision fixée, soient $u_h(t)$ et $v_h(t)$ les polygônes d'Euler obtenus pour les conditions initiales $u_0$ et $v_0$. On suppose $f$ différentiable par rapport à $y$ et $\exists\, L \geq 0 \,/\, \forall\, (x,y) \in \mathbb{R}^2, |\dfrac{\partial f}{\partial y}(x,y)| \leq L$ sur une partie convexe qui contient $(t, u_h(t))$ et $(t, v_h(t))$ $\forall\, t \in [t_0, T]$, alors*

$$|v_h(t) - u_h(t)| \leq \exp(L(t-t_0))|v_0 - u_0| \tag{2.5}$$

*Remarque.* Les lecteurs avisés auront reconnu une condition de Lipschitz dans l'équation (2.4). La condition de la dérivée bornée est historique dans la preuve de Cauchy et impliquera cette condition de Lipschitz grâce à l'inégalité des accroissements finis, corollaire du théorème des accroissements finis, découverte de Cauchy dont ce dernier est très fier. Lipschitz avait, de son coté, fait presque la même preuve du théorème d'existence, sans connaître les travaux de Cauchy, mais en affaiblissant l'hypothèse du lemme 2 en une simple condition de Lipschitz.

*Preuve.* Pour une fonction dérivable, la dérivée est bornée par $L$ ssi la fonction est $L$-lipschitzienne. Ainsi l'hypothèse implique (2.4) ainsi que la majoration par l'exponentielle pour $t_1$. On réitère le procédé par récurrence, en utilisant la propriété de morphisme de l'exponentielle et en écrivant $t - t_0 = t - t_i + t_i - t_{i-1} \cdots - t_0$. On détaille ici l'hérédité de $i = 1$ à $i = 2$ :

$$|v_2 - u_2| \leq \exp(L(t_2 - t_1))|v_1 - u_1| \leq \exp(L(t_2 - t_1)) \exp(L(t_1 - t_0))|v_0 - u_0| = \exp(L(t_2 - t_0))|v_0 - u_0|$$

□

On voudrait maintenant rendre la subdivision de plus en plus *fine*, i.e que l'écart entre deux pas de discrétisation devienne de plus en plus petit, ou plus formellement :

$$|h| = \max_{i \in [\![0, n-1]\!]} h_i \to 0$$

On s'attend dans ce cas que les polygônes d'Euler convergent vers une solution du système différentiel. C'est l'énoncé exact du théorème suivant. Ce théorème existant en deux versions, on a mis en avant celle qui affaiblissait le plus les hypothèses, en mettant entre parenthèses les hypothèses émises par Cauchy.

**Théorème 1** (Cauchy-Lipschitz). *Soit $f$ continue, $f$ bornée par $A$ et $L$-lipschitzienne (resp. différentiable par rapport à la seconde variable, et cette différentielle est bornée par $L$) sur $D = \{(t,u) \,/\, t_0 \leq t \leq T, |u-u_0| \leq b\}$. Si $T - t_0 \leq \dfrac{b}{A}$, alors on a :*

*i) Les polygones d'Euler $u_h(t)$ convergent uniformément vers une fonction continue $\phi(t)$ quand $|h| \to 0$.*

*ii) $\phi$ est de classe $C^1$ et solution de (2.1) sur $[t_0, T]$.*

*iii) Cette solution est unique sur $[t_0, T]$.*

*Preuve.* i) Soit $\epsilon > 0$. Comme $f$ est continue sur $D$ compact, elle est uniformément continue sur $D$, donc $\exists\, \delta > 0 \,/\, |s_1 - s_2| \leq \delta$ et $|v_1 - v_2| \leq A\delta$ impliquent

$$|f(s_1, v_1) - f(s_2, v_2)| \leq \epsilon \tag{2.6}$$

Comme la subdivision $h$ devient de plus en plus petite, on peut supposer $\forall\, i, |t_{i+1} - t_i| \leq \delta$, i.e $|h_i| \leq \delta$.

En premier lieu, nous allons voir ce qu'il se passe lorsqu'on rajoute des points de discrétisation. On considère ainsi une première subdivision $h(1)$ obtenue en ajoutant des points de discrétisation uniquement dans le premier sous-intervalle $[t_0, t_1]$. Le premier lemme appliqué sur $[t_0, t_1]$ donne que la solution plus fine $u_{h(1)}$ vérifie l'estimation en $t_1$ suivante :

$$|u_{h(1)}(t_1) - u_h(t_1)| \leq \epsilon |t_1 - t_0|$$

De plus, les subdivisions $h$ et $h(1)$ étant égales sur $[t_1, T]$, $u_h$ et $_{h(1)}$ définissent des polygones d'Euler de valeurs initiales $u_h(t_1)$ et $u_{h(1)}(t_1)$, on obtient en appliquant le second lemme sur cet intervalle que

$$\forall\, t \in [t_1, T], |u_{h(1)}(t) - u_h(t)| \leq \exp(L|t - t_1|)|u_h(t_1) - u_{h(1)}(t_1)| \leq \epsilon |t_1 - t_0| \exp(L(t - t_1))$$

On réitère alors le procédé en ajoutant de nouveaux points sur le sous-intervalle $[t_1, t_2]$, ce qui nous donnera une nouvelle subdivision que l'on notera $h(2)$. On peut de nouveau appliquer les deux lemmes et obtenir que

$$\forall\, t \in [t_2, T], |u_{h(2)}(t) - u_{h(1)}(t)| \leq \epsilon |t_2 - t_0| \exp(L(t - t_2))$$

On continue alors et on note $\hat{h}$ la subdivision finale obtenue en continuant le procédé jusque $T$. Soit $t_i < t \leq t_{i+1}$, on obtient alors par inégalité triangulaire que

$$\begin{aligned}
|u_{\hat{h}}(t) - u_h(t)| &= |u_{\hat{h}}(t_i) + (t - t_i) f(t_i, u_{\hat{h}}(t_i)) - u_h(t_i) - (t - t_i) f(t_i, u_h(t_i))| \\
&\leq |u_{\hat{h}}(t_i) - u_h(t_i)| + |t - t_i||f(t_i, u_{\hat{h}}(t_i)) - f(t_i, u_h(t_i))| \\
&\leq \epsilon(\exp(L(t - t_1))(t_1 - t_0) + \cdots + \exp(L(t - t_i))(t_i - t_{i-1})) + \epsilon(t - t_i) \\
&\leq \epsilon \int_{t_0}^{t} \exp(L(t - s)) \mathrm{d}s \\
&= \frac{\epsilon}{L}(\exp(L(t - t_0)) - 1)
\end{aligned}$$

Supposons maintenant que l'on dispose de deux subdivisions différentes $h$ et $\hat{h}$, qui sont toutes les deux de pas inférieur à $\delta$ ; on introduit une troisième subdivision $\tilde{h}$ qui est plus fine que les deux, et on applique l'inégalité ci-dessus deux fois, puis on écrit $u_h(t) - u_{\hat{h}}(t) = u_h(t) - u_{\tilde{h}}(t) + u_{\tilde{h}}(t) - u_{\hat{h}}(t)$, alors par inégalité triangulaire on a :

$$|u_{\hat{h}}(t) - u_h(t)| \leq 2\frac{\epsilon}{L}(\exp(L(t - t_0)) - 1)$$

Pour $\epsilon$ assez petit, ceci devient arbitrairement petit, ce qui prouve que le critère de Cauchy uniforme est vérifié par $u_h$ ($\delta$ étant indépendant de $t$), donc comme l'espace vectoriel normé des fonctions continues est complet pour la norme de la convergence uniforme, on a convergence des polygones d'Euler vers une fonction continue $\phi$.

ii) On considère le module de continuité (dont la présentation est traitée en annexe)

$$\epsilon(\delta) := sup\{|f(s_1, v_1) - f(s_2, v_2)|, |s_1 - s_2| \leq \delta, |v_1 - v_2| \leq A\delta\, (s_i, v_i) \in D\}$$

Si $t$ appartient à la subdivision $S$, alors on obtient par (2.3) (lemme 2) en prenant $t := t + \delta$ et $(t_0, u_0) := (t, u_h(t))$ que

$$|u_h(t + \delta) - u_h(t) - \delta f(t, u_h(t))| \leq \epsilon(\delta)\delta$$

En prenant la limite quand $|h| \to 0$, on obtient que $|\phi(t+\delta) - \phi(t) - \delta f(t,\phi(t))| \leq \epsilon(\delta)\delta$. Or $\lim_{\delta\to 0}\epsilon(\delta) = 0$ donc on a bien que $\phi$ est dérivable et sa dérivée est $\forall t \in [t_0, T], \phi'(t) = f(t,\phi(t))$.

iii) Soit $\psi$ une autre solution du système (2.1). On note alors $u_h^i$ le polygone d'Euler correspondant à la valeur initiale $(t_i, \psi(t_i))$, défini pour $t_i \leq t \leq T$. Comme $\psi$ est solution de (2.1) on peut intégrer à vue pour obtenir que $\psi(t) = \psi(t_i) + \int_{t_i}^t f(s,\psi(s))\mathrm{d}s$. Or on a (2.6), donc

$$\begin{aligned}
|\psi(t) - u_h^i(t)| &= |\psi(t_i) + \int_{t_i}^t f(s,\psi(s))\mathrm{d}s - (\psi(t_i) + (t-t_i)f(t_i,\psi(t_i)))| \\
&= |\int_{t_i}^t f(s,\psi(s))\mathrm{d}s - (t-t_i)f(t_i,\psi(t_i))| \\
&= |\int_{t_i}^t (f(s,\psi(s)) - f(t_i,\psi(t_i)))\mathrm{d}s| \\
&\leq \epsilon|t-t_i|
\end{aligned}$$

En utilisant le lemme 2, on peut obtenir de la même manière que dans la preuve de i) que

$$|\psi(t) - u_h(t)| \leq \frac{\epsilon}{L}(\exp(L(t-t_0)) - 1)$$

Et en prenant la limite quand $|h| \to 0$ et $\epsilon \to 0$, on obtient que $|\psi(t) - \phi(t)| \leq 0$ et donc

$$\psi(t) = \phi(t)$$

d'où l'unicité.

□

Ce théorème est un théorème d'existence (et d'unicité) *locale*. Cependant, on peut considérer que le point final $T$ de la subdivision $S$ est le point de départ d'une autre subdivision et obtenir ainsi une nouvelle valeur initiale, sur laquelle on peut réappliquer l'algorithme d'Euler et appliquer le théorème de Cauchy-Lipschitz. On obtient alors :

**Théorème 2.** *Soit $U$ un ouvert de $\mathrm{R}^2$ et soit $f$ une fonction continue, différentiable par rapport à la première variable, et telle que $\dfrac{\partial f}{\partial y}$ soit continue sur $U$. Alors $\forall (t_0, u_0) \in U$, il existe une unique solution de (2.1) qui peut être prolongée par continuité sur la frontière de $U$.*

*Preuve.* Le théorème précédent nous permet d'avoir l'existence et l'unicité d'une solution sur un intervalle de la forme $[\hat{t}, t_0], \hat{t} \leq t_0$. Pour chaque point de $U$, il existe ensuite un voisinage qui vérifie les hypothèses du théorème précédent (ceci provenant principalement du fait qu'une fonction continue atteint ses bornes sur un compact), donc on peut appliquer de nouveau le théorème précédent à droite de $t_0$. □

A la fin du 19e siècle, Paul Painlevé et Émile Picard s'intéressent au cas des équations différentielles de premier et de second ordre. Par ailleurs, Henri Poincaré étudie la stabilité du système solaire et commence sa théorie des systèmes dynamiques. Les phénomènes étudiés dépendent de plusieurs variables, donc de valeurs vectorielles, et sont dans un espace à plusieurs dimensions de la forme $\mathbb{R}^d$. Une généralisation du théorème de Cauchy-Lipschitz pour des fonctions vectorielles sera donc formulée d'après le théorème du point fixe (attribué à Picard) appliqué par Lindelöf dans un espace fonctionnel.

### 2.1.2 Généralisation : théorème de Picard-Lindelöf

**Rappels de topologie**

**Définition 2.** Un **espace de Banach** est un espace vectoriel normé complet pour la distance induite de sa norme. Le produit d'espaces de Banach est un espace de Banach.

**Théorème 3** (Picard / Banach)**.** *Soit $f$ une application contractante dans $(E, ||.||)$ espace de Banach, alors alors $f$ admet un unique point fixe, i.e $\exists !\ x \in E\ /\ f(x) = x$.*

**Généralisation du théorème de Cauchy-Lipschitz**

**Théorème 4.** *Soient $[t_1, t_2]$ un intervalle compact de $\mathbb{R}$, $t_1 < t_2$, $f$ une application continue de $[t_1, t_2] \times \mathbb{R}^d$ dans $\mathbb{R}^d$ lipschitzienne par rapport à la seconde variable, i.e*

$$\exists\, L \in \mathbb{R} \,/\, \forall\, t \in [t_1, t_2], \forall\, (v, w) \in (\mathbb{R}^d)^2, |f(t, v) - f(t, w)| \leq L|v - w|$$

*où $|.|$ désigne une norme quelconque sur $\mathbb{R}^d$.*

*Alors $\forall\, t_0 \in [t_1, t_2], \forall\, u_0 \in \mathbb{R}^d, \exists!\ u \in C^1([t_1, t_2], \mathbb{R}^d)$ solution de (2.1).*

*Preuve.* On prendra tout le long de la preuve la convention de l'intégrale orientée (on rajoutera des valeurs absolues pour assurer les majorations).

*Analyse :* On considère l'application $T$ définie par $\forall\, v : [t_1, t_2] \to \mathbb{R}^d, \forall\, t \in [t_1, t_2]$,

$$(Tv)(t) := u_0 + \int_{t_0}^{t} f(s, v(s))\mathrm{d}s$$

$C^0([t_1, t_2], \mathbb{R})$ est l'espace de Banach des fonctions réelles continues sur $[t_1, t_2]$, il est muni de la norme $||v|| = \max_{t \in [t_1, t_2]} |v(t)|$, où $|.|$ désigne la valeur absolue. La notation $C^0([t_1, t_2], \mathbb{R}^d)$ désigne l'espace des fonctions continues de $[t_1, t_2]$ dans $\mathbb{R}^d$ ; il est isomorphe au produit de $d$ fois l'espace $C^0([t_1, t_2], \mathbb{R})$ et est donc un espace de Banach pour la norme $||v|| = \max_{t \in [t_1, t_2]} |v(t)|$, où $|.|$ désigne une norme de $\mathbb{R}^d$ (peu importe laquelle, toutes les normes sur $\mathbb{R}^d$ étant équivalentes, elles conservent les propriétés topologiques).

$\forall\, t \in [t_1, t_2], (Tv)(t)$ est un vecteur de $\mathbb{R}^d$, donc $Tv$ est une fonction à valeurs dans $\mathbb{R}^d$ ; ainsi $T$ est une application d'un espace de fonctions dans un espace de fonctions. Si $v$ est continue sur $[t_1, t_2], s \mapsto f(s, v(s))$ l'est aussi de $[t_1, t_2]$ dans $\mathbb{R}^d$ par composée de fonctions continues. Or la primitive d'une fonction continue est continue, donc $\forall\, v \in C^0([t_1, t_2], \mathbb{R}^d), Tv \in C^0([t_1, t_2], \mathbb{R}^d)$, ainsi $T$ envoie $C^0([t_1, t_2], \mathbb{R}^d)$ dans lui-même, et même dans $C^1([t_1, t_2], \mathbb{R}^d)$ vue que la dérivée de $Tv$ est $(Tv)'(t) = f(t, v(t))$ qui est continue sur $\mathbb{R}^d$.

Montrons que $T$ est une contraction stricte dans $C^0([t_1, t_2], \mathbb{R}^d)$ pourvu que $L$ soit assez petit. Soient $(v, w) \in (C^0([t_1, t_2], \mathbb{R}^d))^2, t \in [t_1, t_2]$.

$$\begin{aligned} |Tv(t) - Tw(t)| &= |\int_{t_0}^{t} f(s, v(s)) - f(s, w(s))\mathrm{d}s| \\ &\leq \Big| \int_{t_0}^{t} |f(s, v(s)) - f(s, w(s))|\mathrm{d}s \Big| \end{aligned}$$

et par l'hypothèse sur $f$,

$$\begin{aligned} &\leq \Big| \int_{t_0}^{t} L|v(s) - w(s)|\mathrm{d}s \Big| \\ &\leq \Big| L \int_{t_0}^{t} ||v - w||\mathrm{d}s \Big| \\ &\leq \Big| L||v - w||(t - t_0) \Big| \\ &\leq L||v - w|| \max(|t_1 - t_0|, |t_2 - t_0|) \end{aligned}$$

On a alors le résultat si $L \max(|t_1 - t_0|, |t_2 - t_0|) < 1$, $T$ est une contraction stricte et possède un unique point fixe. Cependant, l'énoncé du théorème ne mentionne aucune hypothèse de ce type. On va se débarrasser de cette condition à l'aide des *itérations de Picard*, en montrant $\exists p \in \mathbb{N}$ tel que $T^p$ est une contraction ; cela impliquera que $T$ est une contraction, comme nous allons le voir.

On retourne à l'estimation $|Tv(t) - Tw(t)| \leq L||v - w|||t - t_0|$. Montrons que la propriété $P(n)$ définie par

$$P(n) \iff |T^n(v(t)) - T^n(w(t))| \leq \frac{L^n |t - t_0|^n}{n!} ||v - w||$$

où la notation $T^n$ désigne $T$ composée $n$ fois par elle-même, est vraie pour $n \geq 1$. L'initialisation a déjà été faite, montrons l'hérédité. Supposons $\exists\, n \in \mathbb{N} \,/\, P(n)$ est vraie, montrons qu'alors $P(n+1)$ est vraie. Tout d'abord, on a

$$|Tv_1(t) - Tw_1(t)| \leq L \Big| \int_{t_0}^{t} |v_1(s) - w_1(s)|\mathrm{d}s \Big|$$

Alors en réappliquant cette à hypothèse à $v_1 := T^n(v)$ et $w_1 := T^n(w)$, on obtient

$$\begin{aligned}
|T^{n+1}v(t) - T^{n+1}w(t)| &\leq L\Big|\int_{t_0}^{t} |T^n v_1(s) - T^n w_1(s)|\mathrm{d}s\Big| \\
&\leq L\Big|\int_{t_0}^{t} \frac{L^n|s-t_0|^n}{n!}||v-w||\mathrm{d}s\Big| \\
&= \frac{L^{n+1}}{n!}||v-w||\Big|\int_{t_0}^{t}|s-t_0|\mathrm{d}s\Big| \\
&= \frac{L^{n+1}}{n!}||v-w||\Big|\frac{|t-t_0|}{n+1}\Big| \\
&= \frac{L^{n+1}}{(n+1)!}||v-w||.|t-t_0|^{n+1}
\end{aligned}$$

Or $\sum_{n=0}^{\infty}\frac{L^n|t-t_0|^n}{n!} = \exp(L|t-t_0|)$, donc en particulier $\frac{L^n|t-t_0|^n}{n!} \xrightarrow[n\to\infty]{} 0$. Donc

$$\forall\, L \geq 0, \forall\, t_0 \in [t_1,t_2], \exists\, p \in \mathbb{N}\,/\,\frac{L^p|t_1-t_0|^p}{p!} < 1, \frac{L^p|t_2-t_0|^p}{p!} < 1$$

ou encore

$$\frac{L^p max(|t_1-t_0|^p, |t_2-t_0|^p)}{p!} < 1$$

La propriété ci-dessus nous montre qu'en passant au maximum sur le membre de gauche de l'inégalité prouvée par récurrence, on a :

$$||T^p v - T^p w|| \leq \frac{L^p}{p!}||v-w||.|t-t_0|^p < ||v-w||$$

ce qui permet d'affirmer que $T^p$ est une contraction stricte de $C^0([t_1,t_2],\mathbb{R}^d)$, espace de Banach. On peut appliquer le théorème de Picard et on a donc que $T^p$ admet un unique point fixe $u \in C^0([t_1,t_2],\mathbb{R}^d)$.

Par ailleurs, les points fixes de $T^p$ sont exactement les points fixes de $T$. Si $u$ est point fixe de $T$, en composant $p-1$ fois l'égalité $Tu = u$, on obtient que $T^p u = u$, ce qui prouve que $u$ est un point fixe de $T^p$. Réciproquement, si $u$ est point fixe de $T^p$, on a

$$T^p u = u$$

d'où en composant par $T$,

$$\begin{aligned}
T^{p+1}u &= Tu \\
&= T^p(Tu)
\end{aligned}$$

ce qui prouve que $Tu$ est un point fixe de $T^p$. Or par unicité du point fixe, $Tu = u$, d'où $u$ point fixe de $T$.

On a ainsi montré l'existence et l'unicité d'un point fixe de $T$, et comme $T$ est à valeurs dans $C^1([t_1,t_2],\mathbb{R}^d)$, $u$ est de classe $C^1$ et $\forall\, t \in [t_1,t_2]$,

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

ce qui donne en dérivant cette égalité que $u$ est solution de (2.1).

*Synthèse :* Réciproquement, si $u$ est de classe $C^1$ solution de (2.1), alors en intégrant par rapport au temps, on obtient que

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

□

*Remarque.* Si $f$ est de classe $C^p$, alors la solution $u$ est de classe $C^{p+1}$.

**Exemple.** On considère le système

$$\begin{cases} u'(t) = e^{t^2} \\ u(t_0) = u_0 \end{cases}$$

Ici, $f(t,u) = e^{t^2}$. Alors comme $f$ est continue, lipschitzienne par rapport à $u$ ($|f(t,v) - f(t,w)| = 0$), ce système admet une unique solution par le théorème de Picard-Lindelöf. Cependant, comme indiqué dans l'introduction, on ne peut pas déterminer cette solution *explicitement* à l'aide de fonctions "simples". La même preuve montre que tout système différentiel avec une fonction continue $f$ pour laquelle $u$ n'apparait pas dans l'expression de $f$ admet une unique solution pour des paramètres donnés.

**Exemple.** On considère l'équation $u'(t) = \lambda u(t)$. Ici $f(t,u) = \lambda u$, qui est clairement lipschitzienne. On sait résoudre cette équation explicitement, $u(t) = Ke^t, K \in \mathbb{R}$. Si on impose une condition initiale $u(t_0) = u_0$, la solution devient unique par le théorème de Picard-Lindelöf. Soit $(x,y) \in \mathbb{R}^2$, alors en imposant la condition initiale $u(x) = y$, on trouve qu'il n'existe qu'une seule fonction solution de l'équation ci-dessus et qui passe par $(x,y)$. On peut faire ce raisonnement pour chaque point du plan, ce qui nous amène à la conclusion que les solutions de cette équation différentielle forment une partition de $\mathbb{R}^2$. On peut généraliser ce raisonnement à une fonction $f$ à valeurs dans $\mathbb{R}^d$ pour montrer que les solutions de cette équation forment une partition de $\mathbb{R}^{d+1}$.

**Exemple.** L'application $x \mapsto x^2$ n'étant pas lipschitzienne, il semble qu'on ne peut pas utiliser le théorème de Picard-Lindelhöf pour montrer que $u'(t) = u(t)^2$ admet une solution. Pourtant, on sait déterminer explicitement une solution de cette équation, qui est $u(t) = -\frac{1}{t+c}, c \in \mathbb{R}$. Le théorème de Picard-Lindelhöf donne une condition *suffisante* pour l'existence et l'unicité d'une solution, mais pas *nécessaire*.

Comme pour le cas du théorème en dimension 1 de Cauchy-Lipschitz, ce théorème admet une généralisation plus forte.

**Définition 3.** On dit qu'une application $g$ est localement lipschitzienne sur $J$ (ou lipschitzienne sur les bornés de $J$) ssi $g$ est lipschitzienne sur tout intervalle compact inclus dans $J$.

**Exemple.** L'application $g : x \mapsto x^2$ n'est pas lipschitzienne sur $\mathbb{R}$ car sa dérivée $x \mapsto 2x$ n'est pas bornée sur $\mathbb{R}$, mais elle est localement lipschitzienne sur $\mathbb{R}$. Si $|a| < |b|$, alors $\forall\, x \in [a,b], |g'(x)| \leq 2|b|$, ce qui montre que $g$ est lipschitzienne sur $[a,b]$, intervalle compact de $\mathbb{R}$.

**Théorème 5** (Corollaire)**.** *Supposons que $]t_1,t_2[$ est un intervalle ouvert non vide de $\mathbb{R}$, fini ou infini, et que $f$ est continue sur $]t_1,t_2[\times\mathbb{R}^d$, localement lipschitzienne sur $]t_1,t_2[$ par rapport à sa première variable. Alors on a la même conclusion que le théorème précédent, à savoir $\forall\, t_0 \in [t_1,t_2], \forall\, u_0 \in \mathbb{R}^d, \exists!\ u \in C^1([t_1,t_2],\mathbb{R}^d)$ solution de* (2.1).

*Preuve.* Soient $I \subset J$ des intervalles compacts, $t_0 \in I$. Par le théorème de Picard-Lindelhöf sur $I$ (resp. $J$), il existe $u_I$ (resp. $u_J$) solutions du système de Cauchy sur $I$ (resp. $J$). La restriction de $u_J$ à $I$ est clairement solution du système de Cauchy sur $I$, donc par unicité de la solution, la restriction de $u_J$ vaut $u_I$. On peut donc prolonger une solution sur un intervalle compact $I$ en une solution sur une réunion croissante d'intervalles compacts $I_K$. On conclut en évoquant que chaque ouvert $]t_1,t_2[$ s'écrit comme une réunion croissante de compacts. □

**Exemple.** On retourne à l'exemple de l'équation différentielle $u'(t) = u(t)^2$. $x \mapsto x^2$ étant localement lipschitzienne, le corollaire nous permet d'affirmer qu'il existe une unique solution à cette équation différentielle, pour une condition initiale donnée. On rajoute la condition initiale $u(0) = 1$, alors la solution explicite devient $u(t) = -\frac{1}{t-1}$. Quand $t \to +1$, $|u(t)| \to +\infty$ ; donc le domaine de définition de $u$ le plus grand possible est $]-\infty;1[$, on dit que la fonction $u$ définie sur cet intervalle est la *solution maximale*.

*Remarque.* On peut également étudier des équations différentielles de fonctions complexes, en considérant $\mathbb{C}$ comme $\mathbb{R}^2$ muni de la norme euclidienne.

## 2.2 Systèmes autonomes, d'ordre supérieur, lemme de Gronwall

### 2.2.1 Systèmes autonomes, non autonomes

**Définition 4.** Un système différentiel est dit **autonome** si la variable temporelle n'apparaît pas explicitement dans la fonction $f$. Dans le cas contraire, le système est dit **non autonome**.

**Exemple.** Le système

$$\begin{cases} u'(t) = 3\exp(u(t)) - \dfrac{1}{u(t)} \\ u(\pi) = e \end{cases}$$

est autonome ($f(t,u) = 3\exp(u) - \frac{1}{u}$), mais le système

$$\begin{cases} u'(t) = t.u(t) \\ u(\pi) = e \end{cases}$$

ne l'est pas ($f(t,u) = t.u$).

**Définition 5.** Un système différentiel est dit **linéaire** quand la fonction $f$ est linéaire en les dérivées successives de l'inconnue $u$ ; les coefficients devant ces dérivées sont des fonctions.

*Remarque.* Dans ce cas, on peut appliquer le *principe de superposition* : la somme de deux solutions est encore solution du système, le produit par une constante réelle d'une solution est encore solution. Ainsi, l'ensemble des solutions forme un $\mathbb{R}$-espace vectoriel.

**Exemple.** L'équation différentielle $\lambda u(t) + tu'(t) + u''(t) = 0$ est linéaire.

**Proposition 1.** *L'étude des systèmes non autonomes peut être ramenée à l'étude des systèmes autonomes ; cela augmente la dimension de l'espace des états d'une unité (d est incrémenté) et cela peut faire perdre l'éventuel caractère linéaire du système.*

*Preuve.* Soit le système du premier ordre $u'(t) = f(t,u(t))$, avec une condition initiale donnée. On va transformer ce système en un système autonome par la transformation suivante : soit $s(t) := t$. Pour $Y := \begin{pmatrix} y \\ z \end{pmatrix}, y \in \mathbb{R}^d, z \in \mathbb{R}$, on pose $F(Y) := \begin{pmatrix} f(z,y) \\ 1 \end{pmatrix}$, alors $U(t) := \begin{pmatrix} u(t) \\ s(t) \end{pmatrix}$ vérifie

$$U'(t) = \begin{pmatrix} u'(t) \\ 1 \end{pmatrix}, F(U(t)) = \begin{pmatrix} f(t,u(t)) \\ 1 \end{pmatrix}$$

Ainsi $U$ est solution de $U' = F(U)$, qui est bien un système autonome.

Par ailleurs, en considérant le système linéaire $u'(t) = a(t)u(t)$, cette transformation donne $U(t) = \begin{pmatrix} u(t) \\ t \end{pmatrix}, F(Y) = \begin{pmatrix} a(Y_2)u(Y_1) \\ 1 \end{pmatrix}$. Le système équivaut donc à

$$\begin{cases} U_1' = a(U_2).U_1 \\ U_2' = 1 \end{cases}$$

qui n'est plus linéaire. □

**Exemple.** On considère le système

$$\begin{cases} u'(t) = t.u(t) + \exp(t) \\ u(t_0) = u_0 \end{cases}$$

En appliquant la transformation, on a que ce système équivaut à

$$\begin{cases} U' = F(U) \\ U(t_0) = \begin{pmatrix} u_0 \\ t_0 \end{pmatrix} \end{cases}$$

avec $F(Y) := \begin{pmatrix} Y_1 Y_2 + \exp(Y_2) \\ 1 \end{pmatrix}$.

### 2.2.2 Systèmes d'ordre supérieur

**Définition 6.** Un système différentiel est dit **d'ordre** $p$ ssi l'ordre de la plus grande dérivée apparaissant dans la définition de $f$ est $p$.

**Exemple.** L'équation $u''(t) + 2u'(t) + u(t) = 0$ est d'ordre 2.
L'équation $u^{(p)}(t) + 3t - u(t) = 4$ est d'ordre $p$ ($u^{(p)}$ désigne la $p$-ième dérivée de $u$ par rapport au temps).

Comme pour les systèmes non autonomes qu'on peut transformer en système autonomes, on a la

**Proposition 2.** *L'étude des systèmes d'ordre $p > 1$ peut être ramenée à l'étude d'un système du premier ordre.*

*Preuve.* Formalisons le problème : $f$ est une fonction de $[0,T] \times (\mathbb{R}^d)^p$ dans $\mathbb{R}^d$ et le système de Cauchy s'écrit $u^{(p)}(t) = f(t, u(t), u'(t), \dots, u^{(p-1)}(t))$. On pose alors

$$\forall Y = \begin{pmatrix} y_0 \\ \vdots \\ y_{p-1} \end{pmatrix} \in \mathbb{R}^p, \forall t \in [0,T], F(t,Y) := \begin{pmatrix} y_1 \\ \vdots \\ y_{p-1} \\ f(t, y_0, \dots, y_{p-1}) \end{pmatrix}, U := \begin{pmatrix} u \\ u' \\ \vdots \\ u^{(p-1)} \end{pmatrix}$$

Alors on a que $U'(t) = F(t, U(t))$, système du premier ordre. □

**Exemple.** Soit l'équation $u''(t) = 2u'(t) + u(t)a(t)$. On a que $f(t,u,u') = 2u' + ua(t)$, donc $f(t,v,w) = 2w + va(t)$. On pose alors $F(t,Y) := \begin{pmatrix} y_1 \\ a(t)y_0 + 2y_1 \end{pmatrix}, U(t) := \begin{pmatrix} u(t) \\ u'(t) \end{pmatrix}$. Alors $U'(t) = \begin{pmatrix} u'(t) \\ u''(t) \end{pmatrix} = \begin{pmatrix} u'(t) \\ 2u'(t) + u(t)a(t) \end{pmatrix}$ et $F(t, U(t)) = \begin{pmatrix} u'(t) \\ a(t)u(t) + 2u'(t) \end{pmatrix}$, donc on a bien $U'(t) = F(t, U(t))$.

### 2.2.3 Lemme de Gronwall

**Lemme 3** (Gronwall). *Soit $u \in C^1([0,T], \mathbb{R}^d)$, soient $\phi, \psi$ des fonctions intégrables sur $[0,T]$, positives ou nulles. On suppose $\forall t \in [0,T], |u'(t)| \leq \phi(t) + \psi(t)|u(t)|$. Alors si on note $\Psi(t) = \int_0^t \psi(s)\mathrm{d}s$, $u$ vérifie l'estimation*

$$\forall t \in [0,T], |u(t)| \leq |u_0| e^{\Psi(t)} + \int_0^t \phi(s) e^{\Psi(t) - \Psi(s)} \mathrm{d}s$$

*Preuve.* Soit $\epsilon > 0$, soit $h(t,\epsilon) = e^{\Psi(t)}(|u_0| + \epsilon) + \int_0^t \phi(s) e^{\Psi(t)-\Psi(s)} \mathrm{d}s$. Alors $t \mapsto h(t,\epsilon)$ est continue et $h(0,\epsilon) = |u_0| + \epsilon > |u_0|$.

Au vu de la définition de $h$ et des formules de résolution des équations différentielles linéaires à coefficients variables (cf. section suivante), on a que

$$h'(t,\epsilon) = \phi(t) + h(t,\epsilon)\psi(t) \tag{2.7}$$

Par continuité de $t \mapsto h(t,\epsilon), \exists \tau > 0$ maximal $/ \forall t [0,\tau], |u(t)| \leq h(t,\epsilon)$. Montrons que $\tau = T$ par l'absurde : si $\tau < T$, alors comme $[0,\tau]$ est l'intervalle maximal sur lequel on a l'estimation précédente, on a nécessairement que

$$|u(\tau)| = h(\tau,\epsilon) \tag{2.8}$$

On écrit alors $u(t) = \int_0^t u'(s)\mathrm{d}s + u_0$, et on a pour $t \leq \tau$,

$$\begin{aligned} |u(t)| &\leq |u_0| + \int_0^t |u'(s)| \mathrm{d}s \\ &\leq |u_0| + \int_0^t (\phi(s) + \psi(s)|u(s)|) \mathrm{d}s \\ &\leq |u_0| + \int_0^t (\phi(s) + \psi(s)h(s,\epsilon)) \mathrm{d}s \end{aligned}$$

En revenant à (2.7),

$$\begin{aligned}\int_0^t \psi(s)h(s,\epsilon)\mathrm{d}s &= \int_0^t (h'(s,\epsilon)-\phi(s))\mathrm{d}s \\ &= h(t,\epsilon)-h(0,\epsilon)-\int_0^t \phi(s)\mathrm{d}s \\ &= h(t,\epsilon)-|u_0|-\epsilon-\int_0^t \phi(s)\mathrm{d}s\end{aligned}$$

Et en revenant à la majoration de l'intégrale,

$$|u(t)| \leq h(t,\epsilon)-\epsilon$$

Pour $t=\tau$, cela contredit (2.8), d'où l'absurdité. On a donc $\forall\, t\in[0,T], |u(t)|\leq h(t,\epsilon)$, d'où le résultat quand $\epsilon\to 0$. □

Ainsi, on peut transformer un système d'ordre supérieur à 1, non autonome en un système autonome d'ordre 1. On peut également écrire un système linéaire à plusieurs variables en un système différentiel matriciel, on va maintenant s'intéresser à la résolution des systèmes matriciels.

**Lemme 4.** *Tout système linéaire à plusieurs fonctions inconnues peut s'écrire comme un système matriciel différentiel.*

*Preuve.* Soient $x_1,\dots,x_n$ les inconnues du système différentiel. On a $\forall\, i\in[\![1,n]\!], x_i' = f_i(x_1,\dots,x_n) = \sum_{j=0}^n a_{ij}x_j$, où $a_{ij}$ est une fonction.

On pose alors $X := \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$, $A := (a_{ij})_{(i,j)\in([\![1,n]\!])^2}$. Alors $X' = AX$. □

**Exemple.** Soit le système différentiel suivant :

$$\begin{cases} x_1'(t) = tx_1(t) - x_2(t) \\ x_2'(t) = x_1(t) \\ x_1(t_0) = a \\ x_2(t_0) = b \end{cases}$$

En notant $X(t) := \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix}$, $A(t) := \begin{pmatrix} t & -1 \\ 1 & 0 \end{pmatrix}$, on a que ce système équivaut à

$$\begin{cases} X'(t) = AX(t) \\ X(t_0) = \begin{pmatrix} a \\ b \end{pmatrix} \end{cases}$$

## 2.3 Systèmes différentiels linéaires

Tout le long du problème, on considère comme corps de base $\mathbb{K}=\mathbb{R}$ ou $\mathbb{C}$.

### 2.3.1 Systèmes linéaires à coefficients constants

Soit $A\in M_d(\mathbb{K})$. On étudie le système différentiel

$$u'(t) = A.u(t) \tag{2.9}$$

où $u$ est l'inconnue, $t$ la variable temporelle, et sur lequel on a imposé une condition initiale $u(t_0)=u_0$.

Vérifions que l'on est bien dans les conditions d'application du théorème de Cauchy-Lipschitz : si $|.|$ est une norme sur $\mathbb{K}^d$, on a $f(t,u)=Au$. On considère la norme de l'ensemble des applications linéaires de $\mathbb{K}^d$ dans lui-même, subordonnée à la norme $|.|$ et notée $||.||$ ; par linéarité de A sur un espace vectoriel de

dimension finie, on a $\forall v \in \mathbb{K}^d, |Av| \leq ||A||.|v|$. Ainsi $\forall w \in \mathbb{K}^d, |Av - Aw| = |A(v-w)| \leq ||A||.|v-w|$, ce qui implique que l'on est bien dans les conditions d'application du théorème de Cauchy-Lipschitz.

Pour trouver la solution de ce système, on considère à présent l'équation différentielle à valeurs matricielles :

$$\begin{cases} M'(t) = AM(t) \\ M(0) = I_d \end{cases}$$

Si $||.||$ désigne une norme multiplicative sur les matrices, i.e

$$\forall (A,B) \in (M_d(\mathbb{K}))^2, ||A.B|| \leq ||A||.||B||$$

la fonction $g(t,M) := A.M(t)$ vérifie les conditions de Cauchy-Lipschitz. En effet, $||g(t,M)|| \leq ||A||.||M(t)||$, ce qui suffit car $g$ est linéaire par rapport à son deuxième argument. Ainsi, ce système a également une unique solution $M(t)$.

On s'intéresse alors à la fonction vectorielle $u(t) := M(t-t_0)u_0$. On remarque alors que

$$\begin{aligned} u'(t) &= M'(t-t_0)u_0 \\ &= AM(t-t_0)u_0 \\ &= Au(t) \end{aligned}$$

De plus $u(t_0) = M(0)u_0 = u_0$. Donc $u$ est solution de (2.9) ; par unicité de la solution, $u$ est l'unique solution de ce système.

Montrons quelques propriétés de la fonction matricielle $M(t)$. Tout d'abord, elle commute avec $A$. En effet, posons $\forall t \in [0,T], B(t) := AM(t) - M(t)A$. Alors on a que

$$\begin{aligned} B'(t) &= AM'(t) - M'(t)A \\ &= A(AM(t)) - (AM(t))A \\ &= A(AM(t) - M(t)A) \\ &= AB(t) \end{aligned}$$

Par ailleurs, $B(0) = AM(0) - M(0)A = A - A = 0$. Par unicité de la solution du système

$$\begin{cases} B'(t) = AB(t) \\ B(0) = 0 \end{cases}$$

dont la matrice identiquement nulle est solution, on a que $\forall t \in [0,T], B(t) = 0_d$, i.e $AM(t) = M(t)A$.

Par ailleurs, $t \mapsto M(t)$ est un morphisme des groupes $(\mathbb{R}, +)$ dans $(GL_d(\mathbb{K}), x)$. En effet, on considère $\forall s \in \mathbb{R}, C(t) := M(t+s) - M(t)M(s)$. Alors on a que

$$\begin{aligned} C'(t) &= M'(t+s) - M'(t)M(s) \\ &= AM(t+s) - AM(t)M(s) \\ &= A(M(t+s) - M(t)M(s)) \\ &= AC(t) \end{aligned}$$

et $C(0) = M(s) - I_dM(s) = 0$. De la même manière, par unicité de la solution du système matriciel ci-dessus, $C$ est identiquement nulle et donc $M(t+s) = M(t)M(s)$. De plus, $M(t)$ est inversible, en effet $M(t-t) = M(0) = I_d = M(t)M(-t)$, donc on a bien que $t \mapsto M(t)$ est un morphisme de groupes.

**Définition 7** (Exponentielle de matrice)**.** On appelle **exponentielle de matrice** et on note $e^{At}$ l'unique solution $M(t)$ du système matriciel ci-dessus.

**Lemme 5** (Propriétés de l'exponentielle de matrice)**.** *On a les propriétés suivantes :*

*i)* $\forall t \in \mathbb{R}, e^{At}A = Ae^{At}$

*ii)* $\forall (t,s) \in \mathbb{R}^2, e^{A(t+s)} = e^{At}e^{As}$

*iii)* $t \mapsto M(t)$ *est une fonction analytique de la variable* $t$*, qui admet le développement en série entière suivant, de rayon de convergence infini :*

$$e^{At} = \sum_{k=0}^{+\infty} \frac{A^k t^k}{k!}$$

*iv)* $\forall t \in [0,T], ||e^{At}|| \leq e^{||A||.t}$

*v) L'unique solution du système* (2.9) *est donnée par* $u(t) = e^{A(t-t_0)}u_0$.

*Preuve.* On a déjà prouvé i), ii), v). Il reste à prouver la majoration et le développement en série entière. On remarque d'abord que $\frac{d^pM(0)}{dt^p} = \frac{d^{p-1}M'(0)}{dt^{p-1}} = A\frac{d^{p-1}M(0)}{dt^{p-1}} = \cdots = A^pM(0) = A^p$. On a alors le développement en série entière iii) par une formule de Taylor sur chacun des coefficients de la matrice $M(t)$. Il reste à déterminer son rayon de convergence.

Or, on a par la propriété multiplicative de la norme $||.||$, que $||A^p|| \leq ||A||^p$. On a donc que

$$\begin{aligned}||e^{At}|| = ||\sum_{k=0}^{\infty}\frac{A^kt^k}{k!}|| &\leq \sum_{k=0}^{\infty}\frac{||A^kt^k||}{k!} = \sum_{k=0}^{\infty}\frac{||A^k||.|t^k|}{k!}\\ &\leq \sum_{k=0}^{\infty}\frac{||A||^k.|t|^k}{k!} = e^{||A||.|t|}\end{aligned}$$

ce qui prouve iv) et le rayon de convergence infini. □

On s'est intéressé à un système linéaire à coefficients constants sans second membre, considérons maintenant le cas du second membre.

**Proposition 3** (Formule de Duhamel)**.** *Soit $g$ une fonction continue de $[0,T]$ dans $\mathbb{R}^d$, soit $A \in M_d(\mathbb{R})$. Alors le système*

$$\begin{cases} u'(t) = Au(t) + g(t) \\ u(0) = u_0 \end{cases}$$

*possède une unique solution, donnée par la formule de Duhamel :*

$$u(t) = e^{At}u_0 + \int_0^t e^{A(t-s)}g(s)\mathrm{d}s$$

*Preuve.* Soit $f(t,v) := Av + g(t)$. Alors $|f(t,v) - f(t,w)| = |Av - Aw| \leq ||A||.|v-w|$, d'où le système vérifie bien les conditions de Cauchy-Lipschitz et il y a existence et unicité de la solution. On pose $\forall\, t \in [0,T], v(t) := e^{-At}u(t)$. Alors on a que

$$\begin{aligned}v'(t) &= -Ae^{-At}u(t) + e^{-At}u'(t)\\ &= -Ae^{-At}u(t) + e^{-At}(Au(t) + g(t))\\ &= e^{-At}g(t)\end{aligned}$$

En intégrant alors cette relation à vue, on a que

$$v(t) = \int_0^t e^{-As}g(s)\mathrm{d}s + v(0)$$

et donc en revenant à $u$ en multipliant par $e^{At}$ dans la définition de $v$,

$$\begin{aligned}u(t) &= e^{At}v(t)\\ &= v(0)e^{At} + \int_0^t e^{At-As}g(s)\mathrm{d}s\\ &= u_0e^{At} + \int_0^t e^{A(t-s)}g(s)\mathrm{d}s\end{aligned}$$

car $v(0) = e^0u(0) = u_0$. □

### 2.3.2 Systèmes linéaires à coefficients variables

On s'intéresse maintenant au problème où $A : [0,T] \to M_d(\mathbb{K})$ est une application matricielle continue, $g : [0,T] \to \mathbb{K}^d$ est une application vectorielle continue, et le système à étudier est $u'(t) = A(t)u(t) + g(t)$ en posant la condition initiale $u(0) = u_0$. La fonction $f$ est ici $f(t,v) := A(t)v + g(t)$, et vérifie

$$|f(t,v) - f(t,w)| = |A(t)v - A(t)w| \leq \max_{t\in[0,T]}||A(t)||.|v-w|$$

d'où existence et unicité des solutions pour une condition initiale donnée.

Pour le cas où $d = 1$ (cas scalaire), on sait trouver aisément les solutions. Le système se réécrit $u'(t) = a(t)u(t) + g(t)$, avec $a$ et $g$ des applications continues de $[0, T]$ dans $\mathbb{K}$. Pour trouver la solution générale, on résout d'abord *l'équation homogène*, i.e avec $g(t) = 0$. On a alors $u'(t) = a(t)u(t)$, donc

$$u(t) = u_0.\exp(\int_0^t a(s)\mathrm{d}s)$$

Ensuite, on applique la méthode dite de *la variation de la constante* en posant $v(t) := u(t)\exp(-\int_0^t a(s)\mathrm{d}s)$, ou en notant $a_1(t) := \int_0^t a(s)\mathrm{d}s, v(t) = u(t)\exp(-a_1(t))$.

Alors $v$ est dérivable et

$$\begin{aligned} v'(t) &= u'(t)\exp(-a_1(t)) + u(t)(-a(t))\exp(-a_1(t)) \\ &= (a(t)u(t) + g(t))\exp(-a_1(t)) - a(t)u(t)\exp(-a_1(t)) \\ &= g(t)\exp(-a_1(t)) \end{aligned}$$

Et donc $v(t) = v(0) + \int_0^t g(s)\exp(-a_1(s))\mathrm{d}s$. Or $v(0) = u(0)e^0 = u(0)$, donc

$$v(t) = u(t)\exp(-a_1(t)) = u_0 + \int_0^t g(s)\exp(-a_1(s))\mathrm{d}s$$

et en revenant à la définition de $u$,

$$u(t) = u_0\exp(a_1(t)) + \int_0^t g(s)\exp(a_1(t) - a_1(s))\mathrm{d}s$$

En dimension supérieure, on n'a pas d'expression directe de la solution. En effet, ce qui fait que la méthode ci-dessus marche est que les éléments de $\mathbb{K}$ commutent ; comme ce n'est pas nécessairement le cas en dimension supérieure, cela implique que la dérivée de l'application matricielle $M(t) \mapsto e^{M(t)}$ ne vaut pas nécessairement $M'(t)e^{M(t)}$. On va dériver l'application exponentielle de matrices ; si $M$ et $N$ sont des matrices de $M_d(\mathbb{K})$, si $s \in \mathbb{R}^*$, alors

$$\begin{aligned} \frac{\exp(M + sN) - exp(M)}{s} &= \frac{1}{s}(1 + (M + sN) + \frac{1}{2!}(sMN + M^2 + sNM + s^2N^2) + \cdots - (1 + M + \frac{1}{2!}M^2 \dots)) \\ &= \frac{1}{s}(sN + \frac{1}{2!}(s(MN + NM) + s^2N^2) + \dots) \end{aligned}$$

d'où, quand $s \to 0$,

$$= N + \frac{1}{2!}(MN + NM) + \frac{1}{3!}(M^2N + MNM + N^2M) + \dots$$

En prenant alors $M := A(t), N := A'(t)$, on aura que

$$\frac{d\exp(A(t))}{dt} = A'(t) + \frac{A(t)A'(t) + A'(t)A(t)}{2!} + \dots$$

Ce qui, en général, ne vaut pas $A'(t)\exp(A(t))$. Par contre si $A(t)$ et $A'(t)$ commutent $\forall t \in [0, T]$, alors on aura bien que

$$\frac{d\exp(A(t))}{dt} = A'(t)\exp(A(t))$$

mais cette condition n'est en général pas remplie. Cependant, on sait le résoudre en se ramenant à un autre système. Soit $s \in [0, T]$, on considère le système

$$\begin{cases} \dfrac{\partial G(t,s)}{\partial t} = A(t)G(t,s) \\ G(s,s) = I_d \end{cases}$$

où l'inconnue est $G(.,s)$. Avec la fonction $f$ valant ici $f(t,u) = A(t)u$, on a bien

$$|f(t,v) - f(t,w)| = |A(t)v - A(t)w| \leq \max_{t\in[0,T]} ||A(t)||.|v-w|$$

et donc le système vérifie les conditions de Cauchy-Lipschitz et possède donc une unique solution.

Or, l'équation $u'(t) = A(t)u(t)$ a pour unique solution $u(t) = G(t,t_0)u_0$. En effet,

$$u'(t) = \frac{\partial G(t,t_0)}{\partial t}u_0 = A(t)G(t,t_0)u_0 = A(t)u(t)$$

et $u(t_0) = G(t_0,t_0)u_0 = I_d u_0 = u_0$

**Définition 8.** La matrice $G(t,s)$ est appelée **matrice résolvante** du système; elle permet de trouver la valeur de $u(t) \forall\, t \in [0,T]$, connaissant la condition initiale $u_0$ obtenue en $s$.

**Lemme 6** ("Relation de Chasles"). $\forall\,(s,t,\tau) \in [0,T]^3, G(\tau,t)G(t,s) = G(\tau,s)$.

*Preuve.* On pose $B(\tau) := G(\tau,t)G(t,s) - G(\tau,s)$. Alors on a

$$\begin{aligned} B'(\tau) &= \frac{\partial G(\tau,t)}{\partial \tau}G(t,s) - \frac{\partial G(\tau,s)}{\partial \tau} \\ &= A(\tau)G(\tau,t)G(t,s) - A(\tau)G(\tau,s) \\ &= A(\tau)B(\tau) \end{aligned}$$

Or $B(t) = G(t,t)G(t,s) - G(t,s) = I_d G(t,s) - G(t,s) = 0$, donc par unicité de la solution d'un système différentiel dont la matrice nulle est aussi solution, on a que $B = 0$, d'où le résultat. □

*Remarque.* Physiquement, cela exprime que l'état du système à l'instant $\tau$ est entièrement déterminé par l'état du système à n'importe quel moment : on parle de système *déterministe*.

On a alors une nouvelle formule de Duhamel, pourvu qu'on sache calculer la matrice résolvante du système :

**Proposition 4** (Formule de Duhamel). *Soient $g : [0,T] \to \mathbb{K}^d, A : [0,T] \to M_d(\mathbb{K})$. Le système*

$$\begin{cases} u'(t) = A(t)u(t) + g(t) \\ u(t_0) = u_0 \end{cases}$$

*possède une unique solution donnée par la formule de Duhamel sur $[0,T]$ :*

$$u(t) = G(t,t_0)u_0 + \int_{t_0}^{t} G(t,s)g(s)\mathrm{d}s$$

*Preuve.* En dérivant la solution proposée $u$ ci-dessus par rapport au temps, on obtient que

$$u'(t) = \frac{\partial G(t,t_0)}{\partial t}u_0 + \frac{\partial}{\partial t}\int_{t_0}^{t} G(t,s)g(s)\mathrm{d}s$$

et la règle de Leibniz (traitée en annexe A), nous donne que

$$\begin{aligned} &= \frac{\partial G(t,t_0)}{\partial t}u_0 + G(t,t)g(t) + \int_{t_0}^{t} A(t)G(t,s)g(s)\mathrm{d}s \\ &= A(t)G(t,t_0)u_0 + g(t) + A(t)\int_{t_0}^{t} G(t,s)g(s)\mathrm{d}s \\ &= A(t)u(t) + g(t) \end{aligned}$$

ce qui prouve que $u$ est l'unique solution du système. □

# Chapitre 3

# Étude des schémas

## 3.1 Théorie des schémas à un pas

### 3.1.1 Convergence, consistance, stabilité, ordre

Trouver la solution explicite d'un système différentiel de Cauchy (2.1) n'étant pas toujours possible, on doit pouvoir trouver des méthodes de résolution numérique qui nous donnent de bonnes approximations numériques. La méthode générale consiste à découper l'intervalle sur lequel on souhaite résoudre le système de Cauchy en découpant l'intervalles en plusieurs *pas* : on appelle ce procédé la discrétisation de l'intervalle. Au lieu d'avoir la solution en toutes les valeurs de l'intervalle, on aura la solution sur certains points de l'intervalle. Cette discrétisation nous donne ensuite des valeurs numériques par l'intermédiaire d'un *schéma*, défini par une fonction. On dit qu'un schéma est à *un pas* si pour trouver la valeur actuelle, on ne s'intéresse qu'à la valeur prise en le point de discrétisation précédent ; on ne s'intéressera ici qu'à des schémas à un pas. De plus, on considérera le pas entre deux points *uniforme*, i.e prenant toujours la même valeur. On supposera toujours que la fonction $f$ définie dans le problème de Cauchy vérifie les conditions de Cauchy-Lipschitz.

**Définition 9.** Avec ces précisions, un **schéma à un pas** est une relation de récurrence de la forme

$$U_{k+1} = U_k + hF(t_k, U_k, h) \tag{3.1}$$

où $h$ désigne le pas de temps, compris dans l'intervalle $]0, h^*]$, $h^*$ étant le pas de temps maximum ; $(t_k)$ désigne le temps et est défini par

$$\begin{cases} t_0 \in [0, T] \\ t_{k+1} = t_k + h \end{cases}$$

On travaille donc dans l'intervalle de temps $[t_0, T]$ ; on aura d'autant plus de points que $h$ sera petit. Le nombre de points discrétisés vaut $J(h) := E(\frac{T-t_0}{h})$, où $E$ désigne la partie entière.

Si $u$ est une solution de l'équation différentielle, le vecteur $U_k$ doit approcher le mieux possible la valeur $u(t_k)$, pourvu que la condition initiale du schéma $U_0$ approche le mieux possible la condition initiale du système $u_0$.

La fonction $F$ est définie sur $[t_0, T] \times \mathbb{R}^d \times [0, h^*]$ à valeurs dans $\mathbb{R}^d$.

**Exemple** (Schéma d'Euler explicite)**.** En réécrivant la relation (3.1) $\dfrac{U_{k+1} - U_k}{h} = F(t_k, U_k, h)$, on reconnait quasiment un taux d'accroissement à gauche. En faisant l'approximation

$$\frac{u(t_{k+1}) - u(t_k)}{h} \simeq u'(t_k) = f(t_k, u(t_k))$$

on aboutit à un premier schéma, appelé le *schéma d'Euler explicite* donné par la relation $U_{k+1} = U_k + hf(t_k, U_k)$, avec $F(t, U, h) := f(t, U)$. On détaillera les propriétés de ce schéma un peu plus loin.

#### Convergence

**Définition 10.** i) L'approximation du système de Cauchy définie par (3.1), i.e le schéma est dit **convergente** ssi

$$\forall\, u_0 \in \mathbb{R}^d, \lim_{U_0 \to u_0; h \to 0} \max_{0 \le k \le J(h)} |u(t_k) - U_k| = 0$$

ii) Le schéma est dit **convergent d'ordre** $p$ ssi $\exists K \leq 0 \,/\, \max_{0\leq k\leq J(h)} |u(t_k) - U_k| \leq Kh^p$

*Remarque.* 1. La convergence d'ordre $p$ implique bien évidemment la convergence.

2. Le théorème de Cauchy-Lipschitz (dans sa version prouvée par Cauchy et Lipschitz, présentée dans le premier chapitre) montre que le schéma d'Euler explicite est convergent si $d = 1$.

Intuitivement, la convergence d'un schéma veut dire que mieux on réussit à approximer la condition initiale (on ne peut pas avoir la valeur exacte à cause des erreurs de mesures ou des erreurs d'arrondis) et plus on choisit un pas de petite taille, au mieux seront approximées les valeurs de la solution de (2.1). Cependant, prouver la convergence n'est pas aisé ; on introduit alors des notions qui impliqueront la convergence, et qui seront plus simples à montrer.

### Stabilités

**Définition 11** (Stabilité par rapport aux erreurs)**.** Un schéma (3.1) est dit stable par rapport aux erreurs ssi $\exists M \in \mathbb{R}^+ \,/\, \forall (U_0, V_0) \in (\mathbb{R}^d)^2, \forall h \leq h^*, \forall (\epsilon_j)_{j\in\mathbb{N}} \subset \mathbb{R}^d$, les suites $(U_j)$ et $(V_j)$ définies par

$$\begin{aligned} U_{j+1} &= U_j + hF(t_j, U_j, h) \\ V_{j+1} &= V_j + hF(t_j, V_j, h) + \epsilon_j \end{aligned}$$

vérifient $\forall j \leq J(h), |U_j - V_j| \leq M(|U_0 - V_0| + \sum_{k=0}^{j-1} |\epsilon_k|)$.

Cette propriété est propre au schéma et ne dépend pas du système à approximer. Essayons de comprendre ce que signifie cette propriété : si un schéma n'est pas stable, cela veut dire que $\forall M \geq 0$, on a des suites $(U_j), (V_j), (\epsilon_j)$ qui à partir d'un rang $j_0$ vérifient $|U_{j_0} - V_{j_0}| > M(|U_0 - V_0| + \sum_{k=0}^{j_0-1} |\epsilon_k|)$. En outre, si $\epsilon_k$ représente une erreur de calcul, cela veut dire que les solutions trouvées subiront un écart plus grand à celui des erreurs de calcul cumulées : les erreurs d'approximation prennent de plus en plus d'importance. Ainsi, dire qu'un schéma est stable revient à dire que l'erreur commise sur un pas de temps ne se propage pas *trop* d'un pas de temps au suivant.

**Définition 12** (Stabilité)**.** Un schéma (3.1) est dit stable ssi $\exists M \in \mathbb{R}^+ \,/\, \forall j \leq J(h), U_j \in B(0, M)$, où $B(0, M)$ désigne une boule centrée à l'origine de rayon $M$, par rapport à une norme de $\mathbb{R}^d$ fixée.

Cette propriété est plus simple à appréhender que la précédente. Elle signifie que les valeurs calculées restent confinées dans un domaine bien précis ; la solution calculée *n'explose pas*, i.e ne tend en norme dans aucune direction vers $\infty$.

### Consistance

**Définition 13.** Un schéma (3.1) est dit consistant avec le système de Cauchy (2.1) ssi pour toute solution $u$ de ce système, on a

$$\lim_{h\to 0} \sum_{0\leq j\leq J(h)-1} |u(t_{j+1}) - u(t_j) - hF(t_j, u(t_j), h)| = 0$$

*Remarque.* Cette propriété dépend du schéma ET du système.

Tâchons de comprendre ce que signifie cette propriété pour un schéma. Cela assure que si l'on prend une solution du système, si le pas de temps est petit, l'approximation calculée à partir de la solution restera très proche de la solution.

**Définition 14.** On appelle **erreur locale** la quantité $|u(t_{j+1}) - u(t_j) - hF(t_j, u(t_j), h)|$.

La consistance signifie donc que la somme des erreurs locales tend vers 0 quand le pas de temps devient petit.

### Ordre d'un schéma

Il est intéressant, une fois qu'on a trouvé une méthode numérique pour calculer une approximation d'une solution, de savoir à quel point cette approximation est proche de la solution réelle.

**Définition 15.** Le schéma (3.1) est dit d'ordre $p$ ssi pour toute solution $u$ de ce système, on a

$$\exists K \geq 0 \,/ \sum_{0 \leq j \leq J(h)-1} |u(t_{j+1}) - u(t_j) - hF(t_j, u(t_j), h)| \leq Kh^p$$

*Remarque.* Lorsque l'on prend la limite quand $h$ tend vers 0, on trouve que le schéma est consistant avec le système.

Plus $p$ est grand, plus l'approximation de la solution est bonne.

### Théorèmes de convergence

Voici un résultat, parfois connu sous le nom de *théorème de Lax.*

**Théorème 6.** *Si un schéma à un pas* (3.1) *est consistant avec un système* (2.1)*, et stable par rapport aux erreurs, alors il est convergent.*

*Preuve.* Dans la définition de la stabilité par rapport aux erreurs, on pose $V(j) := u(t_j)$. L'erreur locale associée est alors $\epsilon_j := |V_{j+1} - V_j - hF(t_j, V_j, h)|$. On définit ainsi une suite de vecteurs $(\epsilon_j)$ à valeurs dans $\mathbb{R}^d$.

Or le schéma est stable, donc

$$\exists M \geq 0 \,/ \, \forall j \leq J(h), |U_j - u(t_j)| \leq M(|U_0 - u_0) + \sum_{k=0}^{j-1} |\epsilon_k|)$$

ce qui est en particulier vrai en passant au maximum à gauche, i.e

$$\exists M \geq 0 \,/ \max_{0 \leq j \leq J(h)} |U_j - u(t_j)| \leq M(|U_0 - u_0) + \sum_{k=0}^{j-1} |\epsilon_k|)$$

Or quand $h \to 0, \sum_{k=0}^{j-1} |\epsilon_k| \to 0$ car le schéma est consistant avec le système. Donc

$$\lim_{h \to 0, U_0 \to u_0} \max_{0 \leq j \leq J(h)} |U_j - u(t_j)| \to 0$$

ce qui prouve la convergence. □

De plus, si on a des informations sur l'ordre du schéma, on a le résultat suivant, plus précis en terme d'erreur :

**Théorème 7.** *Si un schéma à un pas* (3.1) *est d'ordre $p$, stable par rapport aux erreurs, et si*

$$\exists C' \geq 0 \,/ \, |U_0 - u_0| \leq C'h^p$$

*alors le schéma est convergent d'ordre $p$.*

*Preuve.* On suit le même déroulement de preuve; on pose $V(j) := u(t_j)$. L'erreur locale associée est alors $\epsilon_j := |V_{j+1} - V_j - hF(t_j, V_j, h)|$. On définit ainsi une suite de vecteurs $(\epsilon_j)$ à valeurs dans $\mathbb{R}^d$.

Le schéma est d'ordre $p$ donc $\exists C \geq 0 \,/ \, \forall j \leq J(h), \sum_{k=0}^{j-1} |\epsilon_k| \leq Ch^p$.

Or le schéma est stable, donc

$$\begin{aligned} \exists M \geq 0 \,/ \, \forall j \leq J(h), |U_j - u(t_j)| &\leq M(|U_0 - u_0) + \sum_{k=0}^{j-1} |\epsilon_k|) \\ &\leq M(C'h^p + Ch^p) \\ &= M(C + C')h^p \end{aligned}$$

d'où la convergence à l'ordre $p$. □

Il existe encore une dernière version de ce théorème qui prouve en plus la stabilité du schéma, sous certaines conditions, mais dont nous ne détaillerons pas la preuve ici.

**Théorème 8.** *On suppose le schéma consistant d'ordre p, et de plus*

$$\exists\, h^* > 0 \,/\, \forall\, A > 0, \exists\, M_A > 0 \,/\, \forall\, (y,z) \in B(0,A)^2, \forall\, t \in [0,T], \forall\, h \in [0,h^*], |F(t,y,h) - F(t,z,h)| \leq M_A |y-z|$$

*i.e la fonction F est lipschitzienne par rapport à sa deuxième variable sur les bornés pour des petites valeurs de h.*

*Alors* $\exists\, h^{**} > 0, \epsilon > 0, K > 0$ *tels que si* $0 < h \leq h^{**}$ *et* $|U_0 - u_0| \leq \epsilon$*, alors :*

*i) le schéma est stable :* $\forall\, j, U_j \in B(0,2A)$*, avec* $A = max_{t\in[0,T]}|u(t)| < \infty$*, où u désigne la solution du système.*

*ii) le schéma est convergent d'ordre p :* $\forall\, j \leq J(h), |U_j - u(t_j)| \leq K(h^p + |U_0 - u0|)$.

Maintenant qu'on a montré que la consistance et la stabilité impliquent la convergence, on va montrer quelques résultants qui facilitent la consistance et la stabilité.

### 3.1.2 Condition nécessaire et suffisante de consistance

**Théorème 9.** *En reprenant les notations de* (2.1) *et* (3.1)*, une condition nécessaire et suffisante pour que le schéma soit consistant avec le système est que*

$$\forall\, t \in [t_0, T], \forall\, u \in \mathbb{R}^d, F(t,u,0) = f(t,u)$$

*Preuve.* On considère l'erreur locale $\epsilon_j = u(t_{j+1}) - u(t_j) - hF(t_j, U_j, h)$. On peut la réécrire

$$\epsilon_j = \int_{t_j}^{t_{j+1}} (f(s,u(s)) - f(t_j,u(t_j)))\mathrm{d}s + h((f(t_j,u(t_j)) - F(t_j,u(t_j),0)) + h(F(t_j,u(t_j),0) - F(t_j,u(t_j),h))$$

On note $\alpha_j := \int_{t_j}^{t_{j+1}} (f(s,u(s)) - f(t_j,u(t_j)))\mathrm{d}s$,
$\beta_j := h((f(t_j,u(t_j)) - F(t_j,u(t_j),0))$,
$\gamma_j := h(F(t_j,u(t_j),0) - F(t_j,u(t_j),h))$, alors $\epsilon_j = \alpha_j + \beta_j + \gamma_j$.

On considère les modules de continuité (présentés brièvement en annexe) $w$ de $t \mapsto f(t,u(t))$ et $w_1$ de $(t,h) \mapsto F(t,u(t),h)$. Pour $|s - t_j| \leq h$, on a

$$|f(s,u(s)) - f(t_j,u(t_j))| \leq w(h)$$

d'où en intégrant sur $[t_j, t_{j+1}]$,

$$\int_{t_j}^{t_{j+1}} |f(s,u(s)) - f(t_j,u(t_j))|\mathrm{d}s \leq \int_{t_j}^{t_{j+1}} w(h)\mathrm{d}s = (t_{j+1} - t_j)w(h) = hw(h)$$

i.e comme $\left| \int_{t_j}^{t_{j+1}} f(s,u(s)) - f(t_j,u(t_j))\mathrm{d}s \right| \leq \int_{t_j}^{t_{j+1}} |f(s,u(s)) - f(t_j,u(t_j))|\mathrm{d}s$

$$|\alpha_j| \leq hw(h)$$

De même, $|F(t_j,u(t_j),0) - F(t_j,u(t_j),h)| \leq w_1(|h-0|) = w_1(h)$ donc $|\gamma_j| \leq hw_1(h)$.
Alors $|\beta_j| = |\epsilon_j - \alpha_j - \gamma_j| \leq |\epsilon_j| + h(w(h) + w_1(h))$.

Supposons le schéma consistant, montrons qu'on a bien la condition de l'énoncé.

$$\begin{aligned}\sum_{j=0}^{J(h)-1} |\beta_j| &\leq \sum_{j=0}^{J(h)-1} \Big(|\epsilon_j| + h(w(h) + w_1(h))\Big) \\ &= \Big(\sum_{j=0}^{J(h)-1} |\epsilon_j|\Big) + J(h)h(w(h) + w_1(h)\end{aligned}$$

or $\sum_{j=0}^{J(h)-1} |\epsilon_j| \to 0$ quand $h \to 0$, car le schéma est consistant et $J(h)h(w(h) + w_1(h) \to 0$ également quand $h \to 0$ car

$$(T - t_0) \leq hJ(h) \leq (T - t_0) + h$$

et en multipliant par $w(h)+w_1(h)$ qui tend vers 0 quand $h \to 0$. Donc

$$\lim_{h\to 0} \sum_{j=0}^{J(h)-1} |\beta_j| = \lim_{h\to 0} \sum_{j=0}^{J(h)-1} |h((f(t_j,u(t_j)) - F(t_j,u(t_j),0))| = 0$$

Mais cette somme est une somme de Riemann et en particulier une formule des rectangles pour la fonction continue $t \mapsto |f(t,u(t)) - F(t,u(t),0)|$ sur l'intervalle $[t_0, t_0 + J(h)h]$. Donc

$$\lim_{h\to 0} \sum_{j=0}^{J(h)-1} |h((f(t_j,u(t_j)) - F(t_j,u(t_j),0))| = \int_{t_0}^{T} |f(s,u(s)) - F(s,u(s),0)| \mathrm{d}s = 0$$

Donc par continuité de la fonction intégrée, pour toute solution $u$ du système, $\forall\, t \in [t_0,T], f(t,u(t)) = F(t,u(t),0)$.

Or $\forall\, (t,u) \in [t_0,T] \times \mathbb{R}^d$, il existe une unique solution du système (2.1) $v$ (par le théorème de Cauchy-Lipschitz) telle que $v(t) = u$, et donc par le résultat ci-dessus,

$$f(t,v(t)) = f(t,u) = F(t,v(t),0) = F(t,u,0)$$

On a donc montré que la condition était nécessaire, montrons maintenant qu'elle est aussi suffisante : on suppose donc que $\forall\, (t,u) \in [t_0,T] \times \mathbb{R}^d$,

$$f(t,u) = F(t,u,0)$$

alors en reprenant les notations ci-dessus, $|\beta_j| = 0$, or $|\epsilon_j| = |\alpha_j + \beta_j + \gamma_j| \leq h(w(h) + w_1(h))$, donc

$$\begin{aligned} \sum_{j=0}^{J(h)-1} |\epsilon_j| &\leq hJ(h)(w(h)+w_1(h)) \\ &\leq (T-t_0)(w(h)+w_1(h)) \end{aligned}$$

qui tend vers 0 quand $h \to 0$, d'où la consistance. □

**Exemple.** Le schéma d'Euler explicite est consistant : on avait que $F(t,U,h) := f(t,U)$ par définition, d'où $F(t,U,0) = f(t,U)$ et la consistance.

### 3.1.3 Lemme de Gronwall discret, condition suffisante de stabilité

Commençons par le lemme de Gronwall discret, qui nous permettra de montrer une condition suffisante de stabilité.

**Lemme 7** (Lemme de Gronwall discret)**.** *Soient* $M \geq 0, h \geq 0, (a_j) \subset \mathbb{R}^+, (b_j) \subset \mathbb{R}^+ / \forall\, j \in \mathbb{N}, a_{j+1} \leq (1+Mh)a_j + b_j$. *Alors*

$$\forall\, j \in \mathbb{N}, a_j \leq e^{Mjh} a_0 + \sum_{k=0}^{j-1} b_k e^{M(j-k-1)h}$$

.

*Preuve.* L'idée est de faire apparaitre une exponentielle dans l'hypothèse, en remarquant $\forall\, x \in \mathbb{R}, 1+x \leq e^x$. Soit $j \in \mathbb{N}$. Alors

$$a_{j+1} \leq e^{Mh} a_j + b_j$$

On pose ensuite $\alpha_j := a_j e^{-Mhj}$, i.e $a_j = \alpha_j e^{Mhj}$. Alors

$$\alpha_{j+1} e^{Mh(j+1)} \leq e^{Mh} e^{Mhj} \alpha_j + b_j$$

et donc $\alpha_{j+1} \leq \alpha - j + b_j e^{-Mh(j+1)}$. On continue de majorer $\alpha_j$ de cette manière à droite, ce qui nous donne par récurrence sur $j$ que

$$\begin{aligned} \alpha_{j+1} &\leq (\alpha_{j-1} + b_{j-1} e^{-Mhj}) + b_j e^{-Mh(j+1)} \\ &\leq \alpha_0 + \sum_{k=0}^{j-1} b_k e^{-Mh(k+1)} \end{aligned}$$

et en revenant à $a_j$,

$$\alpha_j e^{Mhj} \leq \alpha_0 e^{Mhj} + \sum_{k=0}^{j-1} b_k e^{-Mh(k+1)} e^{Mhj}$$

i.e comme $\alpha_0 = a_0$,

$$a_j \leq a_0 e^{Mhj} + \sum_{k=0}^{j-1} b_k e^{Mh(j-k-1)}$$

□

**Théorème 10** (Condition suffisante de stabilité). *Si* $\exists M \geq 0 \,/\, \forall t \in [t_0, T], \forall (u, v) \in (\mathbb{R}^d)^2, \forall h \in [0, h^*]$,

$$|F(t, u, h) - F(t, v, h)| \leq M|u - v|$$

*i.e la fonction $F$ est lipschitzienne par rapport à la seconde variable, alors le schéma défini par $F$ est stable, et on peut prendre pour "constante de stabilité" la valeur $e^{M(T-t_0)}$.*

*Preuve.* Soient $(U_0, V_0) \in (\mathbb{R}^d)^2, h \leq h^*, (\epsilon_j) \subset \mathbb{R}^d, j \in [\![0, J(h)]\!]$, et les suites définies de la manière suivante :

$$\begin{aligned} U_{j+1} &= U_j + hF(t_j, U_j, h) \\ V_{j+1} &= V_j + hF(t_j, V_j, h) + \epsilon_j \end{aligned}$$

Alors

$$\begin{aligned} |V_{j+1} - U_{j+1}| &= |V_j - U_j + h(F(t_j, V_j, h) - F(t_j, U_j, h)) + \epsilon_j| \\ &\leq |V_j - U_j| + h|F(t_j, V_j, h) - F(t_j, U_j, h)| + |\epsilon_j| \\ &\leq (1 + Mh)|V_j - U_j| + |\epsilon_j| \end{aligned}$$

donc par le lemme de Gronwall discret ($a_j := |V_j - U_j|, b_j := |\epsilon_j|$),

$$|V_j - U_j| \leq e^{Mjh}|V_0 - U_0| + \sum_{k=0}^{j-1} |\epsilon_j| e^{M(j-k-1)h}$$

or $jh \leq hJ(h) \leq T - t_0$ donc $e^{Mjh} \leq e^{(T-t_0)M}$ et $\forall k \in [\![0, j-1]\!], e^{M(j-k-1)h} \leq e^{Mjh} \leq e^{M(T-t_0)}$. D'où

$$|V_j - U_j| \leq K(|U_0 + V_0| + \sum_{k=0}^{j-1} |\epsilon_j|)$$

avec $K := e^{M(T-t_0)}$. □

## 3.2 Présentation de quelques schémas

### 3.2.1 Schémas d'Euler, Théta-schéma, schéma de Crank-Nicholson

On s'occupera ici de présenter quelques schémas numériques, tout en s'efforçant le plus possible de comprendre la motivation des mathématiciens qui les ont mis en place.

Le problème (2.1) se réécrit de manière équivalente $u(t) = u_0 + \int_{t_0}^{t} f(s, u(s))\mathrm{d}s$. On peut notamment écrire cette relation entre deux points de discrétisation :

$$u(t_{k+1}) = u(t_k) + \int_{t_k}^{t_{k+1}} f(s, u(s))\mathrm{d}s$$

Une bonne méthode pour trouver des schémas consiste alors à estimer la valeur de l'intégrale ci-dessus.

**Schéma d'Euler explicite**

On propose d'abord d'évaluer la valeur de cette intégrale par la méthode des rectangles à gauche, i.e en faisant l'approximation

$$\int_a^b g(s)\mathrm{d}s \simeq (b-a)g(a)$$

Ce qui nous donne ici $u(t_{k+1}) \simeq u(t_k) + (t_{k+1} - t_k)f(t_k, u(t_k)) = u(t_k) + hf(t_k, u(t_k))$, ce qui nous mène à poser $F(t_k, U_k, h) := f(t_k, u_k)$ : cette fonction $F$ définit le **schéma d'Euler explicite**. On dit qu'il est explicite, car contrairement à d'autres schémas que nous étudierons plus tard, on n'a pas besoin de résoudre un système linéaire pour trouver l'expression de $U_{k+1}$ :

$$U_{k+1} = U_k + hf(t_k, U_k)$$

Comme $F(t, u, 0) = f(t, u)$ par définition, et $f$ est lipschitzienne (car on a toujours supposé que $f$ vérifiait les conditions du théorème de Cauchy-Lipschitz) donc $F$ est lipschitzienne par rapport à sa seconde variable, donc $F$ définit un schéma stable. Ainsi $F$ définit un schéma convergent.

*Remarque.* La preuve originelle du théorème de Cauchy-Lipschitz montre également que le schéma est convergent, puisqu'il permet de construire une solution du système de Cauchy.

Si $f$ est de classe $C^1$, alors ce schéma est d'ordre 1. En faisant un développement de Taylor à l'ordre 1 de $u$ en $t_k$, on trouve que

$$\exists\, \epsilon \,/\, \lim_{h\to 0} \epsilon(h) = 0, |\epsilon_j| = |u(t_{k+1}) - u(t_k) - hf(t_k, u(t_k))| = h|\epsilon(h)| = O(h^2)$$

et comme le nombre de pas total $J(h)$ est en $O(\frac{1}{h})$, cela nous assure que l'erreur globale

$$\sum_{k=0}^{J(h)-1} |\epsilon_k| = O(h^2).O(\frac{1}{h}) = O(h)$$

et donc que le système est d'ordre 1.

**Schéma d'Euler implicite**

La méthode des rectangles peut aussi être prise sur le point à droite de l'intervalle d'intégration, i.e on fait l'approximation

$$\int_a^b g(s)\mathrm{d}s \simeq (b-a)g(b)$$

Ce qui nous donne ici $u(t_{k+1}) \simeq u(t_k) + (t_{k+1} - t_k)f(t_{k+1}, u(t_{k+1})) = u(t_k) + hf(t_{k+1}, u(t_{k+1}))$ ; on a donc $U_{k+1} = U_k + f(t_{k+1}, U_{k+1})$ : c'est le schéma d'Euler implicite (ou rétrograde) ; il est implicite car on doit résoudre un système pour trouver l'expression de $U_{k+1}$, généralement non linéaire. Montrons d'abord que ce schéma peut bien s'écrire sous la forme des schémas à un pas (3.1).

On considère le problème de point fixe

$$v = u + hf(s, v) \tag{3.2}$$

où $v$ est l'inconnue. On pose alors $\forall\, (s, u, v, h) \in [0, T] \times \mathbb{R}^d \times \mathbb{R}^d \times [0, h^*], g(s, u, v, h) := u + hf(s, v)$. Alors on a

$$|g(s, u, v, h) - g(s, u, v', h)| = h|f(s, v) - f(s, v')| \leq Lh|v - v'|$$

donc $g$ est une contraction stricte par rapport à $v$ si $hL < 1$. On fixe $h^* < \frac{1}{L}$, alors pour $h \leq h^*$, on aura aussi que $g$ est contractante dans $\mathbb{R}^d$, complet pour toutes les normes : donc il y a existence et unicité d'une solution à (3.2), qu'on note $G(s, u, h)$.

Alors $G$ est continue par rapport à tous ses arguments, lipschitzienne par rapport à $u$. En effet, en considérant

$$v_1 = u_1 + h_1 f(s_1, v_1)$$
$$v_2 = u_2 + h_2 f(s_2, v_2)$$

On aura que

$$\begin{aligned}|v_1 - v_2| &= |u_1 - u_2 + h_1 f(s_1, v_1) - h_2 f(s_2, v_2)| \\ &= |u_1 - u_2 + h_1 f(s_1, v_1) - h_1 f(s_1, v_2) + h_1 f(s_1, v_2) - h_1 f(s_2, v_2) + h_1 f(s_2, v_2) - h_2 f(s_2, v_2)| \\ &\leq |u_1 - u_2| + h_1 |f(s_1, v_1) - f(s_1, v_2)| + h_1 |f(s_1, v_2) - f(s_2, v_2)| + |h_1 - h_2||f(s_2, v_2)|\end{aligned}$$

et en fixant $v_2, s_2, h_2$,

$$\leq |u_1 - u_2| + h_1.L.|v_1 - v_2| + h_1 |f(s_1, v_2) - f(s_2, v_2)| + |h_1 - h_2|.|f(s_2, v_2)|$$

et donc

$$(1 - h_1 L)|v_1 - v_2| \leq |u_1 - u_2| + h_1 |f(s_1, v_2) - f(s_2, v_2)| + |h_1 - h_2|.|f(s_2, v_2)|$$

et $h_1 L < 1$ donc

$$|v_1 - v_2| \leq \frac{1}{1 - h^* L}(|u_1 - u_2| + h_1 |f(s_1, v_2) - f(s_2, v_2)| + |h_1 - h_2|.|f(s_2, v_2)|)$$

ce qui montre que quand $(s_1, u_1, h_1) \to (s_2, u_2, h_2), v_1 \to v_2$ et donc la continuité de $G$.

De plus, si $s_1 = s_2 =: s, h_1 = h_2 =: h$ alors

$$|v_2 - v_1| = |G(s, u_2, h) - G(s, u_1, h)| \leq \frac{1}{1 - hL}|u_2 - u_1|$$

ce qui montre que $G$ est lipschitzienne par rapport à son deuxième argument.

Ainsi, $U_{k+1} = U_k + f(t_{k+1}, U_{k+1})$ se réécrit $U_{k+1} = G(t_{k+1}, U_k, h) = G(t_k + h, U_k, h)$. On a donc que

$$U_{k+1} = U_k + h f(t_k + h, G(t_k + h, U_k, h))$$

donc $F(t, u, h) := f(t + h, G(t + h, u, h))$.

Par continuité de $G$, $(t, h) \mapsto t + h$ et $f$, $F$ est continue par rapport à ses arguments. Le schéma ainsi défini est consistant : $F(t, u, 0) = f(t, G(t, u, 0))$, où $G(t, u, 0)$ est l'unique solution de $v = u + 0 f(t, v) = u$, donc $G(t, u, 0) = u$ et $F(t, u, 0) = f(t, u)$ ce qui prouve la consistance. On a donc convergence du schéma d'Euler implicite.

Par ailleurs le schéma est stable :

$$\begin{aligned}|F(t, u_1, h) - F(t, u_2, h)| &= |f(t + h, G(t + h, u_1, h)) - f(t + h, G(t + h, u_2, h))| \\ &\leq L|G(t + h, u_1, h) - G(t + h, u_2, h)| \\ &\leq \frac{L}{1 - Lh}|u_1 - u_2| \\ &\leq \frac{L}{1 - Lh^*}|u_1 - u_2|\end{aligned}$$

donc $F$ est lipschitzienne par rapport à sa deuxième variable, donc $F$ définit un schéma stable par rapport aux erreurs.

On peut prouver que ce schéma est d'ordre 1 si $f$ est de classe $C^1$. De plus il est également stable (cf deuxième stabilité des schémas), i.e que pour des petites valeurs de $h$, la solution reste incluse dans un domaine bien précis.

**Théta schéma, schéma de Crank-Nicholson**

Pour mieux estimer l'intégrale avec la méthode des rectangles, on peut écrire que l'intégrale vaut une combinaison convexe des valeurs obtenus avec la méthode des rectangles à gauche et à droite, i.e on choisit $\theta \in [0, 1]$ et on estime

$$\int_a^b g(s)\mathrm{d}s \simeq (b - a)\Big(\theta g(a) + (1 - \theta) g(b)\Big)$$

On est amenés à poser $U_{k+1} = U_k + h\Big(\theta f(t_k, U_k) + (1 - \theta) f(t_{k+1}, U_{k+1})\Big)$. On peut montrer, comme pour le schéma d'Euler implicite par un problème de point fixe que ce schéma peut s'écrire comme sous la forme générale des schémas à un pas, qu'il est consistant et stable et donc convergent.

Si $\theta = \frac{1}{2}$, le schéma est d'ordre 2 pourvu que $f$ soit de classe $C^2$, s'inspire directement de la méthode des trapèzes (détaillée plus bas) et est appelé schéma de Crank-Nicholson. Sinon, pour $\theta \neq \frac{1}{2}$, le schéma est d'ordre 1 et appelé théta-schéma. On note que pour $\theta < 1$, ce schéma est implicite.

*Remarque.* Quand $\theta \to 0$, on retrouve l'expression du schéma d'Euler implicite ; quand $\theta \to 1$, on retrouve l'expression du schéma d'Euler explicite.

**Schéma d'Euler amélioré**

On aimerait un schéma d'ordre 2 comme celui de Crank-Nicholson, mais sans le caractère implicite. On peut estimer une intégrale par la méthode du point-milieu, i.e

$$\int_a^b g(s)\mathrm{d}s \simeq (b-a)g(\frac{a+b}{2})$$

Ce qui nous donne ici $U_{k+1} = U_k + hf(t_k + \frac{h}{2}, U_{k+\frac{1}{2}})$, où $U_{k+\frac{1}{2}}$ est une estimation de la valeur en $u(t_k + \frac{h}{2})$, évaluée avec la méthode d'Euler selon

$$U_{k+\frac{1}{2}} \simeq U_k + \frac{h}{2} f(t_k, U_k)$$

Ainsi, on a posé

$$U_{k+1} = U_k + hf(t_k + \frac{h}{2}, U_k + \frac{h}{2} f(t_k, U_k))$$

et la fonction $F$ qui définit le schéma est alors $F(t,u,h) = f(t + \frac{h}{2}, u + \frac{h}{2} f(t,u))$.

Alors $F(t,u,0) = f(t,u)$ ce qui prouve que le schéma est consistant. De plus,

$$\begin{aligned} |F(t,u,h) - F(t,v,h)| &\leq |u + \frac{h}{2} f(t,u) - v - \frac{h}{2} f(t,v)| \\ &\leq |u-v| + \frac{h}{2}|u-v| \\ &\leq (1 + \frac{h^*}{2})|u-v| \end{aligned}$$

et donc la stabilité du schéma. Ainsi ce schéma est convergent.

On peut prouver que ce schéma est d'ordre 2 si $f$ est de classe $C^2$. On note que ce schéma est explicite.

### 3.2.2 Schéma de Heun

Pour approximer notre intégrale, on peut aussi utiliser la méthode du trapèze, i.e

$$\int_a^b g(s)\mathrm{d}s \simeq \frac{(b-a)}{2}(g(a) + g(b))$$

Le schéma qui en est déduit est dû à Heun et utilise l'idée du prédicteur-correcteur : on va *prédire* la valeur de $U_{k+1}$ à l'aide de la méthode d'Euler qu'on *corrigera* par la méthode des trapèzes :

$$\begin{aligned} Y_1 &:= U_k \\ Y_2 &:= U_k + h(t_k, U_k) \\ U_{k+1} &:= U_k + \frac{h}{2}\Big(f(t_k, Y_1) + f(t_k + h, Y_2)\Big) \end{aligned}$$

On définit donc $F(t,u,h) := \frac{1}{2}(f(t,u) + f(t+h, u+hf(t,u)))$. On a directement que $F$ définit un schéma consistant ($F(t,u,0) = f(t,u)$) et on peut montrer aisément que $F$ est lipschitzienne par rapport à sa seconde variable, d'où la stabilité du schéma de Heun : on a donc convergence du schéma de Heun. Ce schéma est explicite, on peut montrer qu'il est d'ordre 2 si $f$ est de classe $C^2$.

### 3.2.3 "Le" schéma de Runge-Kutta, formalisation des schémas

Les mathématiciens Karl Runge et Martin Wilhelm Kutta ont l'idée au début du 20e siècle d'utiliser les formules de quadratures numériques des intégrales pour trouver des schémas de n'importe quel ordre. "Le" schéma de Runge-Kutta le plus utilisé est d'ordre 4 et inspiré de la méthode de Simpsons pour le calcul des intégrales.

**Formule de Simpsons**

Si $f$ est une fonction de $I$ dans $\mathbb{R}^d$, où $I$ est l'intervalle $[a,b], a<b$, alors on pose $m := \dfrac{a+b}{2}$ et on donne une approximation de $f$ par le polynôme de Lagrange de degré 2 :

$$\forall\, x \in I, P(x) = f(a)\frac{(x-b)(x-m)}{(a-b)(a-m)} + f(b)\frac{(x-a)(x-m)}{(b-a)(b-m)} + f(m)\frac{(x-a)(x-b)}{(m-a)(m-b)}$$

On sait alors plus "facilement" calculer l'aire sous la courbe de $P$. Après un calcul (douloureux...), on trouve que

$$\int_a^b f(s)\mathrm{d}s \simeq \int_a^b P(s)\mathrm{d}s = \frac{b-a}{6}(f(a) + 4f(\frac{a+b}{2}) + f(b))$$

**Schéma de Runge-Kutta d'ordre 4 : RK4**

On est amenés à poser le schéma défini par la récurrence

$$U_{k+1} = U_k + \frac{h}{6}(f(t_k, U_k) + 4f(t_k + \frac{h}{2}, U_{k+\frac{1}{2}}) + f(t_{k+1}, U_{k+1}))$$

On pose alors

$$\begin{aligned} p_1 &:= f(t_k, U_k) \\ p_2 &:= f(t_k + \frac{h}{2}, U_k + \frac{h}{2}p_1) \\ p_3 &:= f(t_k + \frac{h}{2}, U_k + \frac{h}{2}p_2) \\ p_4 &:= f(t_k + h, U_k + hp_3) \end{aligned}$$

et le schéma devient

$$U_{k+1} = U_k + \frac{h}{6}(p_1 + 2p_2 + 2p_3 + p_4)$$

$p_1$ désignant la pente au début de l'intervalle, $p_2$ la pente au milieu de l'intervalle calculée à partir de $p_1$, $p_3$ la pente au milieu de l'intervalle calculée à partir de $p_2$, $p_4$ la pente à la fin de l'intervalle calculée à l'aide de $p_3$.

Cette méthode est explicite (la valeur explicite de $F$ fait apparaitre 4 fois la fonction $f$ !), on peut montrer (de la même manière que pour le schéma d'Euler amélioré) qu'elle est stable et consistante et donc convergente. Si $f$ est de classe $C^4$, alors ce schéma est d'ordre 4.

**Formalisme**

**Définition 16.** Une méthode de Runge-Kutta est définie par la donnée d'un tableau $(a_{ij})$ de $q$ lignes et $q$ colonnes, d'un vecteur colonne $c$ à $q$ lignes, d'un vecteur ligne $b$ à $q$ colonnes, le tout disposé de la manière suivante :

| | | | |
|---|---|---|---|
| $c_1$ | $a_{11}$ | $\dots$ | $a_{1q}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $c_q$ | $a_{q1}$ | $\dots$ | $a_{qq}$ |
| | $b_1$ | $\dots$ | $b_q$ |

La méthode associée à ce schéma est alors

$$\forall\, i \in [\![1,q]\!], U_{n,i} := U_n + h\sum_{j=1}^{q} a_{ij} f(t_n + c_j h, U_{n,j})$$

$$U_{n+1} := U_n + h\sum_{j=1}^{q} b_j f(t_n + c_j h, U_{n,j})$$

*Remarque.* Si $\forall\,(i,j) \in [\![1,q]\!]^2, i \leq j \implies a_{ij} = 0$, alors le schéma associé est explicite. En effet, sinon on a $\exists\, i \leq j$ et $a_{ij} \neq 0$, donc les calcul de $U_{n,i}$ et de $U_{n,j}$ sont liés l'un à l'autre : le schéma est implicite.

**Formalisme de Runge-Kutta pour les schémas étudiés**

Tous les schémas étudiés jusque maintenant peuvent se réécrire sous la forme d'un tableau de Runge-Kutta :

**Schéma d'Euler explicite**

| | |
|---|---|
| 0 | 0 |
| | 1 |

**Schéma d'Euler implicite**

| | |
|---|---|
| 1 | 1 |
| | 1 |

**Théta schéma**

| | | |
|---|---|---|
| 0 | 0 | 0 |
| 1 | $\theta$ | $(1-\theta)$ |
| | $\theta$ | $(1-\theta)$ |

**Euler amélioré**

| | | |
|---|---|---|
| 0 | 0 | 0 |
| 1/2 | 1/2 | 0 |
| | 0 | 1 |

**Heun**

| | | |
|---|---|---|
| 0 | 0 | 0 |
| 1 | 1 | 0 |
| | 1/2 | 1/2 |

**RK4**

| | | | | |
|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 |
| 1/2 | 1/2 | 0 | 0 | 0 |
| 1/2 | 0 | 1/2 | 0 | 0 |
| 1 | 0 | 0 | 1 | 0 |
| | 1/6 | 2/3 | 2/3 | 1/6 |

On reconnait facilement les schémas explicites, pour lesquels tous les nombres au dessus de la diagonale sont des 0.

## 3.3 Étude des approximations d'une EDO simple

### 3.3.1 Présentation

On considère maintenant l'étude d'un système différentiel, pour essayer de trouver quel est le meilleur schéma pour trouver l'approximation la plus proche de la solution. On s'intéresse à un problème qu'on sait résoudre explicitement : pour $\lambda \in \mathbb{R}$,

$$\begin{cases} u'(t) = \lambda u(t) \\ u(0) = 1 \end{cases}$$

dont la solution explicite est $u(t) = e^{\lambda t}$. On a ici $f(t,u) = \lambda u$, l'équation est autonome du premier ordre.

### 3.3.2 Euler explicite ?

Si on essaie de résoudre cette équation avec le schéma d'Euler explicite, on va avoir pour récurrence :

$$U_{k+1} = U_k + h\lambda U_k$$

i.e $U_{k+1} = (1+h\lambda)U_k$, ce qui est la relation de récurrence d'une suite géométrique, dont on connaît le terme général :

$$U_k = (1+h\lambda)^k U_0$$

En supposant qu'on ait $U_0 = 1, \lambda = -1$, cela nous donne $U_k = (1-h)^k$. Comme $\lambda < 0$, la solution réelle est $e^{-t}$, qui est bornée ; or la valeur calculée ici est bornée ssi la raison de la suite géométrique est inférieure à

1 ssi $|1-h| \leq 1$, i.e $h \leq 2$. Mais si on veut calculer l'approximation sur une grande échelle de temps, on a besoin d'un pas de temps le plus grand possible ; or, si $h = 2$, la solution donnée est $U_k = (-1)^k$, qui n'a plus rien à voir avec la solution réelle, et qui de plus change de signe alors que la solution réelle est strictement positive. Il faut donc faire tendre le pas vers 0 pour avoir une approximation satisfaisante de la solution du système, ce qui augmente très vite le nombre de calculs pour avoir des approximations sur de longues durées. De plus, la solution calculée ne donne qu'une approximation assez inexacte de la solution réelle, qui elle tend très vite vers 0 : on dit que l'équation est *raide*.

Par contre, si on essaie le schéma d'Euler implicite, on se ramène à étudier

$$U_{k+1} = U_k + h\lambda U_{k+1}$$

ie $U_{k+1} = \frac{U_k}{1-h\lambda}$, ce qui donne la relation de récurrence $U_k = \frac{U_0}{(1-h\lambda)^k}$. Pour les mêmes paramètres ($\lambda = -1, U_0 = 1$), cela nous donne l'équation $U_k = \frac{1}{(1+h)^k}$, qui reste bornée et positive quelle que soit la valeur du pas de temps $h$, et qui approxime mieux la solution réelle. On pourrait alors en conclure que le schéma d'Euler implicite donne de meilleurs résultats, mais...

### 3.3.3 Euler implicite ?

On reprend la même équation mais en prenant comme paramètre $\lambda := 1$ cette fois. Alors avec le schéma d'Euler implicite, on aboutit à la récurrence $U_{k+1} = \frac{1}{(1-h)^k}$. Si le pas de temps vaut $h = 1$, cette solution n'est même pas définie ! De plus, si $h > 1$, la solution calculée change de signe et on perd le caractère positif de la solution réelle une fois de plus (c'est notamment le cas pour $h = 2$, où $U_k = (-1)^k$). Enfin, si $h$ est très proche de 1, la solution va "exploser" très rapidement, i.e tendre en valeur absolue vers $+\infty$.

Cette fois, il semble que le schéma d'Euler explicite soit mieux adapté : en effet, la solution calculée est alors $U_k = (1+h)^k$, qui tend bien vers $+\infty$ quand $k \to +\infty$. La solution est positive, même pour des grands pas de temps, définie partout. Ainsi le schéma d'Euler semble mieux adapté pour ce cas-là.

### 3.3.4 Conclusion ?

La conclusion est qu'on ne peut pas dire qu'il y ait un schéma meilleur que tous les autres pour résoudre une équation différentielle donnée ; par contre, il y a bien des schémas meilleurs que d'autres pour résoudre des problèmes donnés, le problème étant que pour trouver quel schéma correspondra le mieux à une équation donnée, il faut déjà connaître pour quelles équations ledit schéma donne de bons résultats.

# Chapitre 4

# Conclusion

Bien que les équations différentielles régissant certains systèmes dynamiques peuvent être résolues au moins numériquement, cela ne peut pas permettre de prévoir tous les comportements des systèmes. En effet, ces systèmes sont très sensibles aux conditions initiales et une légère modification des conditions initiales entraînera des modifications très importantes des comportements attendus : c'est ce qu'on appelle *l'effet papillon*, ou encore la *théorie du chaos*, le point de vue extrême pour expliquer ce principe étant que le battement d'ailes d'un papillon au Brésil pourrait provoquer une tornade au Texas.

En météorologie par exemple, il suffit qu'une petite perturbation surgisse et tous les calculs effectués pour prévoir le temps des jours suivants s'en trouveront faussés ; aussi il ne faut pas en vouloir à Météo France de prédire une tempête les jours de soleil radieux ou un temps magnifique les jours noirs de nuages, les erreurs sont simplement dûes à la théorie du chaos.

# Annexe A

# Règle de Leibniz - Dérivation sous le signe somme

Pour calculer une dérivée par rapport au temps de la forme $\int_{a(t)}^{b(t)} f(t,s)\mathrm{d}s$ on a besoin de la règle de Leibniz de dérivation sous le signe somme.

Soient $p, q, s$ des entiers strictement positifs, $A \subset \mathbb{R}^q, B \subset \mathbb{R}^s, f : \mathbb{R}^q \times \mathbb{R}^s \to \mathbb{R}^p$. On considère l'intégrale paramétrique $F$ définie par $\forall\, y \in A, F(y) := \int_B f(y,z)\mathrm{d}z$.

**Lemme 8** (Règle de Leibniz Locale)**.** *On suppose que $A$ est d'intérieur non vide et que :*

*i)* $\forall\, y \in A, f(y,.)$ *est intégrable sur* $B$.

*ii)* $\exists\, i \in [\![1,q]\!], \exists\, a \in \mathring{A}, \exists\, r > 0 \,/\, B_\infty(a,r) \subset A$ *et* $\forall\, z \in B, f(.,z)$ *possède une dérivée partielle par rapport à* $y_i$ $\forall\, y \in B_\infty(a,r)$.

*iii)* $\exists\, g, h$ *fonctions intégrables sur* $B$ *telles que*

$$\forall\, y \in B_\infty(a,r), \forall\, z \in B, g(z) \leq \frac{\partial}{\partial y_i} f(y,z) \leq h(z)$$

.

*Alors l'intégrale paramétrique $F$ possède en $a$ une dérivée partielle par rapport à $y_i$ et celle-ci vaut :*

$$\frac{\partial F(a)}{\partial y_i} = \frac{\partial}{\partial y_i} \int_B f(a,z)\mathrm{d}z = \int_B \frac{\partial}{\partial y_i} f(a,z)\mathrm{d}z$$

*Preuve.* Soit $\phi$ définie par $\forall\, h \in [-r,r]\backslash\{0\}, \forall\, z \in B$,

$$\phi(h,z) = \frac{f(a + h\hat{e_i}, z) - f(a,z)}{h}$$

où $\hat{e_i}$ désigne le vecteur de base correspondant à la variable $y_i$. Alors $\phi(h,.)$ est intégrable sur $B$ (par l'hypothèse i) $\forall\, h \in [-r,r]\backslash\{0\}$ et par l'hypothèse ii,

$$\forall\, z \in B, \lim_{h \to 0} \phi(h,z) = \frac{\partial}{\partial y_i} f(a,z)$$

Par le théorème des accroissements finis, $\forall\, h \in [-r,r]\backslash\{0\}, \exists\, h' \,/\, 0 < |h'| < |h|$ et

$$f(a + h\hat{e_i}, z) - f(a,z) = h \frac{\partial}{\partial y_i} f(a + h'\hat{e_i}, z)$$

Ainsi, on a $\phi(h,z) = \dfrac{\partial}{\partial y_i} f(a + h'\hat{e_i}, z)$.

En utilisant la troisième hypothèse, comme $\forall\, z \in B, g(z) \leq \phi(h,z) \leq h(z)$, par le théorème de convergence dominée de Lebesgue, on a :

$$\lim_{h \to 0} \int_B \phi(h,z)\mathrm{d}z = \int_B \frac{\partial}{\partial y_i} f(a,z)\mathrm{d}z$$

or

$$L := \lim_{h\to 0}\int_B \phi(h,z)\mathrm{d}z = \lim_{h\to 0}\frac{1}{h}\int_B (f(a+h\hat{e_i},z)-f(a,z))\mathrm{d}z$$

et par définition de $F$,

$$L = \lim_{h\to 0}\frac{F(a+h\hat{e_i})-F(a)}{h} = \frac{\partial}{\partial y_i}F(a)$$

□

**Proposition 5** (Règle de Leibniz Globale)**.** *On suppose maintenant $A$ ouvert, $B$ fermé borné, et $\exists\, i \in [\![1,q]\!] \,/\, \forall\,(y,z)\in A\times B, f$ possède une dérivée partielle par rapport à $y_i$ et $\dfrac{\partial}{\partial y_i}f$ est continue sur $A\times B$.*

*Alors $F$ possède en chaque point $y\in A$ une dérivée partielle par rapport à $y_i$, $\dfrac{\partial}{\partial y_i}F$ est continue sur $A$ et*

$$\forall\, y\in A, \frac{\partial}{\partial y_i}F(y) = \int_B \frac{\partial}{\partial y_i}f(y,z)\mathrm{d}z$$

*Preuve.* Soit $a\in A, \exists\, r>0\,/\, B_\infty(a,r)\subset A; \forall\, y\in B_\infty(a,r), \dfrac{\partial}{\partial y_i}f(y,.)$ est continue (par hypothèse) donc intégrable sur $B$. Par continuité, $\exists\, M>0\,/\,\forall\,(y,z)\in B_\infty(a,r)\times B, -M\leq \dfrac{\partial}{\partial y_i}f(y,z)\leq M$. Comme $B$ est fermé borné, $z\mapsto \pm M$ est intégrable sur $B$, on peut alors appliquer la règle de Leibniz locale partout. La continuité de $\dfrac{\partial}{\partial y_i}F$ provient du théorème de continuité des intégrales à paramètres. □

On en déduit le cas particulier des intégrales des fonctions de deux variables à valeurs dans $\mathbb{R}$, utilisé ici pour la formule de Duhamel :

**Proposition 6.** *Soit $f$ continue de $\mathbb{R}^2$ dans $\mathbb{R}$, possédant une dérivée partielle continue sur $\mathbb{R}$ par rapport à la première variable. Soient $a,b$ des fonctions dérivables de $\mathbb{R}$ dans $\mathbb{R}$. On considère l'intégrale paramétrique définie par*

$$\forall\, y\in\mathbb{R}, F(y) := \int_{a(y)}^{b(y)} f(y,z)\mathrm{d}z$$

*Alors $F$ est dérivable sur $\mathbb{R}$ et on a*

$$\forall\, y\in\mathbb{R}, F'(y) = f(y,b(y)).b'(y) - f(y,a(y)).a'(y) + \int_{a(y)}^{b(y)}\frac{\partial f(y,z)}{\partial y}\mathrm{d}z$$

*Preuve.* On considère $H$ définie de $\mathbb{R}^3$ dans $\mathbb{R}$ par

$$\forall\,(u,v,y)\in\mathbb{R}^3, H(u,v,y) := \int_u^v f(y,z)\mathrm{d}z$$

Alors $\forall\, y\in\mathbb{R}, H(a(y),b(y),y) = F(y)$. Par le théorème de dérivation des fonctions composées, on a

$$\forall\, y\in\mathbb{R}, F'(y) = \frac{\partial H}{\partial a}\frac{\partial a}{\partial y} + \frac{\partial H}{\partial b}\frac{\partial b}{\partial y} + \frac{\partial H}{\partial y}$$

Or, d'après le théorème fondamental de l'analyse,

$$\frac{\partial H}{\partial a} = \frac{\partial}{\partial a}\int_{a(y)}^{b(y)} f(y,z)\mathrm{d}z = -f(y,a(y))$$

$$\frac{\partial H}{\partial b} = \frac{\partial}{\partial b}\int_{a(y)}^{b(y)} f(y,z)\mathrm{d}z = f(y,b(y))$$

Et la règle de Leibniz globale (valable en prenant pour $B$ l'intervalle fermé de bornes $a(y)$ et $b(y)$) nous permet de calculer le dernier terme :

$$\frac{\partial H}{\partial y} = \int_{a(y)}^{b(y)}\frac{\partial f(y,z)}{\partial y}\mathrm{d}z$$

□

**Exemple.** Si on prend $a(y) := a \in \mathbb{R}$, $b(y) := b \in \mathbb{R}$, on retrouve un théorème de dérivabilité des intégrales à paramètres :

$$F'(y) = \int_a^b \frac{\partial}{\partial y} f(y, z) \mathrm{d}z$$

**Exemple.** Dans la preuve de la formule de Duhamel pour des matrices à coefficients variables, on avait besoin de calculer la dérivée par rapport à $t$ de $\int_{t_0}^t G(t, s)g(s)\mathrm{d}s$. La dérivée est trouvée grâce à la règle de Leibniz ($G$ possédant une dérivée partielle par rapport à son premier argument), il s'agit de $G(t, t)g(t) + \int_{t_0}^t \frac{\partial G(t, s)}{\partial t} A(s)\mathrm{d}s$.

# Annexe B

# Module de continuité

**Définition 17.** Soit $E$ un espace vectoriel normé, $f$ une fonction continue de $K$ compact dans $E$. Le **module de continuité** $w$ associé est une fonction de $\mathbb{R}^+$ dans lui-même, définie pour $h \geq 0$ par

$$w(h) := \sup_{(x,y)\in K^2, |x-y|\leq h} |f(x) - f(y)|$$

### Quelques propriétés utilisées du module de continuité

On a $\forall\, (x,y) \in K^2, |f(x) - f(y)| \leq w(|x-y|)$.

$w(0) = \sup_{(x,y)\in K^2, x=y} |f(x) - f(y)| = 0$.

Soient $h \geq 0, h' \geq 0, w(h_1 + h_2) = \sup_{(x,y)\in K^2, |x-y|\leq h+h'} |f(x) - f(y)|$. Soient $(x,y) \in K^2 \,/\, |x-y| \leq h$, alors $|f(x) - f(y)| \leq w(h)$ mais comme $h \leq h + h'$, $|f(x) - f(y)| \leq w(h + h')$ et au vu de la définition de $w$ par un sup, $w(h) \leq w(h + h')$ et donc $w$ est croissante.

Par ailleurs, $f$ est continue sur un compact $K$, donc uniformément continue sur $K$ par le lemme de Heine, ce qui implique directement que $w$ est continue en 0. On peut montrer la continuité de $w$ sur les compacts connexes, mais ce n'est pas utile pour ce dossier.