Equations Différentielles Ordinaires

Tony Lelièvre

2009-2010

L'objectif de ce cours est d'introduire plusieurs outils nécessaires pour bâtir et utiliser des modèles basés sur des équations différentielles ordinaires. On s'attachera en particulier à étudier plusieurs aspects du comportement qualitatif des solutions, ainsi que de leur simulation numérique.

1 Quelques modèles basés sur des équations différentielles ordinaires

Les modèles basés sur des équations différentielles ordinaires sont extrêmement courants. Donnons quelques exemples tirés de divers champs scientifiques.

1.1 Mécanique

L'évolution d'un ensemble de points matériels de vecteur position x est régie par la loi de Newton :

$$m\ddot{x} = F(x, \dot{x}),$$

où $\dot{x}$ désigne la dérivée première par rapport au temps (la vitesse) et $\ddot{x}$ la dérivée seconde par rapport au temps (l'accélération). On a ici supposé la masse m constante. De manière générale, m peut être un tenseur. Le second membre $F(x, \dot{x})$ désigne une force dépendant de la position et de la vitesse.

Remarquer que l'on peut se ramener à une équation du premier ordre en posant $v = \dot{x}$ et en travaillant dans *l'espace des phases* (x, v) :

$$\left\{ \begin{array}{l} \dot{x} = v, \\ \dot{v} = \frac{1}{m} F(x, \dot{x}). \end{array} \right. \tag{1}$$

Quelques exemple de champs de force que l'on retrouvera par la suite :

- pour un pendule pesant, on a : $x \in \mathbb{T} = \mathbb{R}/(2\pi\mathbb{Z})$ et $F(x, \dot{x}) = -m\dfrac{g}{l}\sin x - k\dot{x}$, où g est l'accélération de la gravité, l la longueur du fil, et $k \geq 0$ un coefficient de frottement visqueux.
- la force peut dériver d'un potentiel : $F(x) = -\nabla V(x)$. Dans ce cas, on a *un système hamiltonien*. Si on pose $(q, p) = (x, mv)$ (p désigne la quantité de

mouvement) et $H(q,p) = \frac{1}{2}m\|v\|^2 + V(x)$, le système se réécrit :

$$\begin{cases} \dot{q} = \dfrac{\partial H}{\partial p}(q,p), \\ \dot{p} = -\dfrac{\partial H}{\partial q}(q,p). \end{cases} \tag{2}$$

Remarquer que H est conservée le long des trajectoires : $\frac{d}{dt}H(p(t),q(t)) = 0$. De manière générale, une quantité qui reste constante le long des trajectoires est appelée une *intégrale première du mouvement*. L'étude de ce type de système est très important en pratique (on le rencontre notamment en dynamique moléculaire [5, 8], ou en mécanique céleste).

1.2 Dynamique des populations

Les équations différentielles ordinaires sont très utilisées pour modéliser l'évolution d'une population. Un exemple est donné par les équations de Lotka-Volterra. Il s'agit d'un modèle de type prédateurs-proies. On note P la concentration de prédateurs et N la concentration de proies. L'évolution de P et N satisfait :

$$\begin{cases} \dot{N} = N(a - bP), \\ \dot{P} = P(cN - d). \end{cases}$$

avec a, b, c, d quatre coefficients strictement positifs qui s'interprètent de la façon suivante : a est un taux de naissance pour les proies, b est un taux de prédation, d est un taux de mortalité pour les prédateurs et c est un taux de reproduction pour les prédateurs.

Quelques solutions sont représentées sur la Figure 1. On observe que les solutions semblent périodiques, restent positives et confinées le long d'une courbe dans le plan (N, P).

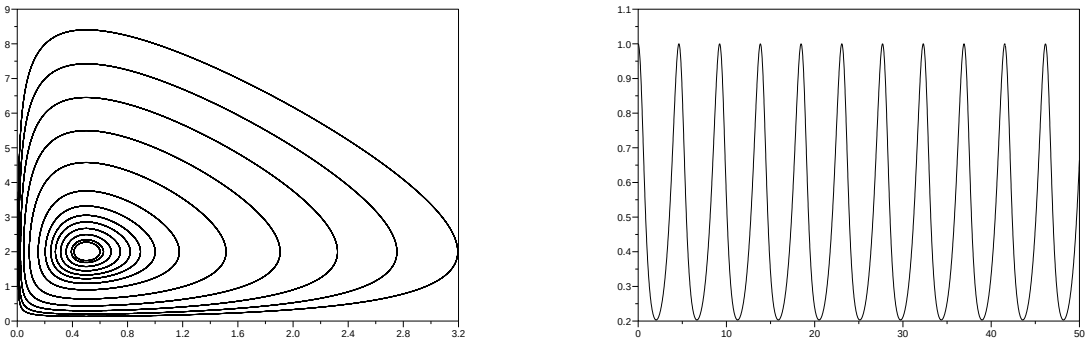

FIG. 1 – Solutions des équations de Lotka-Volterra. A gauche, quelques trajectoires $(N(t), P(t))_{t\geq 0}$ pour diverses conditions initiales (en abscisse N, et en ordonnée P). A droite, Une trajectoire $t \mapsto N(t)$ (en abscisse t, et en ordonnée N). Paramètres : $a = 2, b = 1, c = 2, d = 1$.

Exercice 1 – *Discuter le sens de chacun des termes des équations (signe, non-linéarité).*
– *Remarquer que si (N, P) est solution, alors $\dot{N}(c-d/N) = \dot{P}(a/P-b)$ (méthode de séparation des variables). En déduire que $H(N, P) = cN - d\ln N + bP - a\ln P$ est conservé le long des trajectoires.*

1.3 Météorologie

Les équations qui permettent de modéliser l'évolution des différents paramètres météorologiques (température, humidité, pression, etc...) sont très complexes. Un système très simple mais qui permet de retrouver plusieurs des caractéristiques de ces équations a été proposé par Lorenz. Il s'agit d'un système d'équations différentielles ordinaires en dimension 3 :

$$\begin{cases} \dot{x} = \sigma(y - x), \\ \dot{y} = rx - y - xz, \\ \dot{z} = xy - bz \end{cases}$$

Des valeurs classiques pour les paramètres sont $\sigma = 10$, $r = 28$ et $b = 8/3$.

Sur la Figure 2, on trace quelques solutions de ces équations. On observe en particulier que les trajectoires divergent très vite pour deux conditions initiales très proches. En particulier, il semble vain de vouloir étudier les trajectoires pour des conditions initiales précises, car une simple variation des conditions initiales induit rapidement des modifications importantes. Une meilleure approche serait peut-être d'étudier l'évolution d'un *ensemble de conditions initiales* (propagation d'une mesure de probabilité par la dynamique), ou bien le *comportement moyen le long d'une trajectoire* (moyenne trajectorielle).

Pourtant, il semble que la dynamique contienne des structures "simples" cachées derrière la complexité des trajectoires $t \mapsto (x, y, z)(t)$. Ainsi les trajectoires s'approchent de courbes limites dans la limite des temps longs. Sur la Figure 3, on vérife que la suite des maxima de la trajectoire $t \mapsto z(t)$ vérifie une relation simple du type

$$z_{m+1} = T(z_m),$$

où $T : I \to I$ (I un intervalle de $\mathbb{R}$) est une application de forme caractéristique (application tente). On étudiera dans la suite un tel système dynamique.

1.4 Cinétique chimique

Considérons une réaction chimique sur trois espèces : $Y_3 \overset{k_3}{\to} Y_2 \overset{k_2}{\to} Y_1 \overset{k_1}{\to} Y_2$.

L'évolution de la concentration $(y_i)_{1\leq i\leq 3}$ des trois espèces $(Y_i)_{1\leq i\leq 3}$ satisfait le système d'équations différentielles ordinaires :

$$\begin{cases} \dot{y_1} = -k_1 y_1 + k_2 y_2, \\ \dot{y_2} = k_1 y_1 - k_2 y_2 + k_3 y_3, \\ \dot{y_3} = -k_3 y_3 \end{cases}$$

Dans ce genre de problème, on observe typiquement de grandes disparités dans les coefficients k_i. Le système dynamique contient donc des échelles de temps très

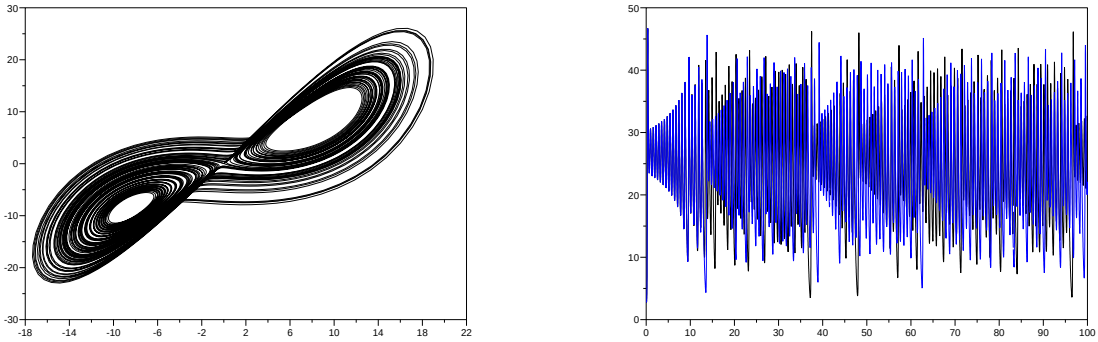

FIG. 2 – Solutions des équations de Lorenz. A gauche, une trajectoire typique $(x(t), y(t))_{t\geq 0}$ (en abscisse x, et en ordonnée y). A droite, Deux trajectoires $t \mapsto z(t)$ pour deux conditions initiales très proches mais différentes ($(x, y, z)(0) = (1, 2, 3)$ et $(x, y, z)(0) = (1.01, 2, 3)$) (en abscisse t, et en ordonnée z). Paramètres : $\sigma = 10$, $r = 28$ et $b = 8/3$.

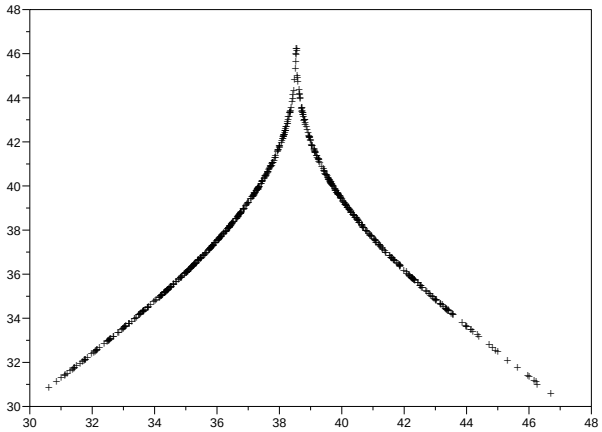

FIG. 3 – Solutions des équations de Lorenz. Chaque point a pour coordonnées (z_m, z_{m+1}) où les $(z_n)_{n\geq 0}$ sont les maxima successifs observés le long d'une trajectoire $t \mapsto z(t)$.

différentes. Il faut évidemment tenir compte de cette information pour discrétiser les équations de manière efficace. Dans l'exemple précédent, si k_3 est très grand par rapport à k_1 et k_2, on comprend que très rapidement, la concentration de Y_3 tend vers 0. On peut donc légitimement se demander si on ne peut pas construire analytiquement un modèle plus simple, seulement sur les concentrations des espèces Y_1 et Y_2.

1.5 Quelques questions soulevées par ces modèles

A partir de ces quelques exemples, on peut naturellement se poser les questions suivantes :

- Sous quelles conditions un modèle basé sur des équations différentielles ordinaires est bien posé ? Ceci recouvre plus précisément trois questions : (i) les équations admettent-elles une solution ? (ii) Cette solution est-elle unique ? (iii) Cette solution dépend-elle continûment des données ?
- Comment peut-on analyser le comportement qualitatif des solutions ? On peut par exemple s'interroger sur l'existence de solutions périodiques, le comportement en temps long des solutions, la dépendance des solutions en les conditions initiales.
- Comment discrétiser efficacement ces équations pour obtenir des solutions approchées ?
- Comment obtenir des modèles réduits, sur un nombre plus petit de variables ?

Nous allons essayer d'apporter quelques éléments de réponse à ces questions dans les sections suivantes.

Avertissement : Le lecteur ne cherchera pas de liens très formalisés entre les Sections 3, 4 et 5. L'objectif de cette première partie du cours est d'introduire qulques notions fondamentales liées à l'analyse théorique et numérique des systèmes dynamiques (à temps discret ou continu), en restant à un niveau élémentaire.

2 Résultats d'existence et d'unicité

On considère le problème suivant : pour une fonction $f(t, x)$ définie sur $I \times O$, où I est un ouvert de $\mathbb{R}$ et O un ouvert de $\mathbb{R}^d$, trouver une fonction $x(t)$ définie sur un ouvert $J \subset I$ dérivable en tout point et telle que

$$\dot{x} = f(t, x). \tag{3}$$

Tel quel, le problème admet une infinité de solutions : il faut ajouter typiquement une condition initiale : pour un temps $t_0 \in I$,

$$x(t_0) = x_0. \tag{4}$$

Le problème (3)–(4) est appelé un problème de Cauchy.

Remarque 1 *Si f est continue, étant donné qu'on cherche une solution dérivable, la solution sera en fait de classe $\mathcal{C}^1$.*

Remarque 2 *Ne pas confondre un problème de Cauchy, avec des problèmes où l'on se donne des conditions aux limites (aux extrémités de l'intervalle I).*

Remarque 3 *Dans le cas où f ne dépend pas du temps, on parle d'équation différentielle ordinaire* autonome *ou* homogène en temps.

2.1 Le cas linéaire

2.1.1 Problème non autonome

On suppose dans cette section que

$$f(t,x) = A(t)\,x, \tag{5}$$

où A est une fonction *continue* à valeur dans les matrices de taille $d \times d$, et définie sur $I = \mathbb{R}$.

Théorème 1 (Théorème de Cauchy) *On suppose $t \mapsto A(t)$ continue. Quelque soit la condition initiale $x(t_0) = x_0$, il existe une unique solution $x(t)$ de classe $\mathcal{C}^1$ définie sur $\mathbb{R}$.*

Preuve : Il suffit de montrer le théorème sur un intervalle compact $I = [0,1]$, et pour une condition initiale $x(0) = x_0$ (le vérifier en exercice).

On introduit l'application affine

$$T \left\{ \begin{array}{ccc} \mathcal{C}^0([0,1],\mathbb{R}) & \to & \mathcal{C}^0([0,1],\mathbb{R}), \\ (x(t))_{0\le t\le 1} & \mapsto & \left(x_0 + \int_0^t A(s)x(s)\,ds\right)_{0\le t\le 1}. \end{array} \right.$$

On munit l'espace $\mathcal{C}^0([0,1],\mathbb{R})$ de le norme L^∞ :

$$\|x\|_{L^\infty} = \sup_{0\le t\le 1} |x(t)|.$$

Un point fixe de l'application T (*i.e.* un point x tel que $Tx = x$) est une solution du problème de Cauchy. Le résultat d'existence et unicité est basé sur le théorème de point fixe de Picard (cf. Théorème 2 ci-dessous).

Montrons que $\exists m > 0$, T^m est une application contractante. On considère deux fonctions x, $y \in \mathcal{C}^0([0,1],\mathbb{R})$. On observe que $\forall t \in [0,1]$,

$$\begin{aligned} |(T(x) - T(y))(t)| &= \left| \int_0^t A(s)(x(s) - y(s)) \right|, \\ &\le \|A\|_{L^\infty} \|x - y\|_{L^\infty}\, t. \end{aligned}$$

Noter que $\|A\|_{L^\infty} < \infty$ car on a supposé A continue sur le compact $[0,1]$. En itérant cette relation, on a :

$$\begin{aligned} |(T^2(x) - T^2(y))(t)| &= \left| \int_0^t A(s)(T(x)(s) - T(y)(s)) \right|, \\ &\le \|A\|_{L^\infty} \|x - y\|_{L^\infty} \int_0^t s\,ds \\ &= \|A\|_{L^\infty} \|x - y\|_{L^\infty} \frac{t^2}{2}. \end{aligned}$$

Et par récurrence, on montre que pour tout $m \geq 1$, et pour tout $t \in [0,1]$,

$$|(T^m(x) - T^m(y))(t)| \leq \frac{\|A\|_{L^\infty}^m}{m!} \|x - y\|_{L^\infty} t^m.$$

En particulier, on a

$$\|T^m(x) - T^m(y)\|_{L^\infty} \leq \frac{\|A\|_{L^\infty}^m}{m!} \|x - y\|_{L^\infty}.$$

Ainsi, pour m assez grand, T^m est une application contractante, ce qui conclut la preuve, en utilisant le Théorème 2 et l'Exercice 2. ♢

On pourrait aussi démontrer le même résultat en supposant que A est intégrable sur les compacts (le vérifier en exercice).

Théorème 2 (Théorème du point fixe de Picard) *Soit (X, d) un espace métrique complet et $T : X \to X$ une application contractante : $\exists k \in [0,1)$, $\forall x, y \in X$,*

$$d(Tx, Ty) \leq k\, d(x, y).$$

Alors $\exists! x \in X$ tel que $Tx = x$.

Preuve : L'unicité est immédiate. Pour l'existence, on considère la suite récurrente :

$$\begin{cases} x_{n+1} = T(x_n), \\ x_0 \text{ donné.} \end{cases}$$

On va montrer que la suite (x_n) est de Cauchy. On remarque que, pour $x \in X$,

$$d(Tx, T^2x) \leq k\, d(x, Tx),$$

et, par une récurrence immédiate, pour $x \in X$ et $n \geq 0$,

$$d(T^n x, T^{n+1}x) \leq k^n\, d(x, Tx).$$

Ainsi, on a pour $x \in X$ et $p \geq 0$,

$$\begin{aligned} d(x, T^p x) &\leq d(x, Tx) + d(Tx, T^2x) + \ldots + d(T^{p-1}x, T^p x), \\ &\leq d(x, Tx)(1 + k + \ldots + k^{p-1}), \\ &\leq d(x, Tx)\frac{1 - k^p}{1 - k}. \end{aligned}$$

On en déduit que pour $n \geq 0$ et $p \geq 0$,

$$\begin{aligned} d(x_n, x_{n+p}) &= d(x_n, T^p(x_n)), \\ &\leq d(x_n, Tx_n)\frac{1 - k^p}{1 - k}, \\ &= d(T^n x_0, T^{n+1}x_0)\frac{1 - k^p}{1 - k}, \\ &\leq d(x_0, Tx_0)k^n \frac{1}{1 - k}. \end{aligned}$$

Ceci montre que la suite (x_n) est de Cauchy, donc converge vers un $x^* \in X$ (car X est supposé complet pour la métrique d). En passant à la limite dans la relation de récurrence $x_{n+1} = T(x_n)$ (en utilisant le fait que T est une application continue), on obtient $x^* = T(x^*)$, d'où l'existence d'un point fixe. ♢

Exercice 2 *Montrer que les conclusions du Théorème 2 restent valables sous l'hypothèse plus faible :* $\exists m > 0$ *tel que* T^m *soit une application contractante.*

Exercice 3 *On se place dans le cadre du Théorème 2, et on suppose que l'on doit calculer numériquement le point fixe de l'application* T*. On effectue donc les itérations de Picard* $x_{n+1} = T(x_n)$*. Donner un critère d'arrêt pour cet algorithme.* Solution : Il suffit de regarder si deux itérés successifs sont proches, puisqu'on a $d(x_n, x_\infty) \leq d(x_n, x_{n+1})/(1-k)$.

Le théorème précédent assure l'existence d'une solution définie pour tout temps, pour toute condition initiale. La solution est linéaire par rapport à la condition initiale : si on note $t \mapsto \phi(0, x_0; t)$ la solution de condition initiale $x(0) = x_0$ et $t \mapsto \phi(0, x_1; t)$ la solution de condition initiale $x(0) = x_1$, $t \mapsto \lambda\phi(0, x_0; t) + \phi(0, x_1; t)$ est la solution de condition initiale $x(0) = \lambda x_0 + x_1$.

Cette propriété de linéarité nous permet d'écrire la solution de condition initiale $x(t_0) = x_0$ sous la forme

$$\phi(t_0, x_0; t) = R(t_0, t)\, x_0$$

où $R(t_0, t)$ (qui est une matrice de $\mathbb{R}^{d\times d}$) est appelée *la résolvante*. La fonction $t \mapsto R(t_0, t)$ est la solution du problème de Cauchy linéaire (en dimension d^2)

$$\begin{cases} \dot{R} = A(t)R, \\ R(t_0) = \mathrm{Id}. \end{cases}$$

Dans le cas linéaire, on peut résoudre les problèmes avec second membre en connaissant la solution du problème sans second membre, plus une solution particulière.

Proposition 1 (Formule de Duhamel) *La solution de*

$$\begin{cases} \dot{x} = A(t)x + b(t), \\ x(t_0) = x_0, \end{cases}$$

(où b *est une fonction continue) est* $x(t) = R(t_0, t)x_0 + \int_{t_0}^{t} R(s, t)b(s)\, ds$*.*

La solution dépend donc de manière affine du second membre b.

Exercice 4 (Théorème de Liouville) *Soit* $R(t_0, t)$ *la résolvante associé à une équation différentielle ordinaire linéaire :*

$$\dot{x}(t) = A(t)x(t)$$

où A *est une fonction continue à valeur dans* $\mathbb{R}^{d\times d}$*. Pour* t_0 *fixé, soit* $u(t) = \det(R(t_0, t))$*. Pourquoi* u *ne s'annule pas ? Montrer que* u *vérifie*

$$\begin{cases} \dot{u} = \mathrm{tr}(A(t))u, \\ u(t_0) = 1. \end{cases}$$

Si $v(t, x) = A(t)x$*, vérifier que* $\mathrm{div}\, v = 0$ *est équivalent à* $\mathrm{tr}(A(t)) = 0$*. Dans ce cas, on a donc* $\det(R(t_0, t)) = 1$*. Interpréter ce résultat en terme de conservation de volume dans un champ de vitesse à divergence nulle. (Ce résultat peut se généraliser au cas non-linéaire : si* $\phi(t_0, x_0; t)$ *est le flot associé à une équation différentielle ordinaire* $\dot{x} = f(t, x)$ *(cf. Définition 1 ci-dessous) avec* $\mathrm{div}\, f(t, x) = 0$*, alors la différentielle du flot par rapport à* x_0 *est de déterminant* 1*.)*

2.1.2 Problème autonome

Dans cette section, on suppose que $A(t)$ ne dépend pas du temps dans (5). La solution s'écrit alors explicitement

$$x(t) = \exp(tA)x_0$$

où

$$\exp(tA) = \sum_{n\geq 0} \frac{(tA)^n}{n!}$$

est l'exponentielle matricielle de tA.

On rappelle que :

- Pour tout matrice A, la série $\sum_{n\geq 0} \frac{A^n}{n!}$ converge (car elle converge normalement, c'est-à-dire que $\sum_{n\geq 0} \frac{\|A\|^n}{n!} < \infty$).
- Si A et B commutent, $\exp(A+B) = \exp(A)\exp(B)$. Si A et B ne commutent pas, cette relation est en général fausse.

Attention : dans le cas où $A(t)$ dépend du temps, la solution du problème de Cauchy n'est pas en général $x(t) = \exp(\int_0^t A(s)\,ds)x_0$. Ceci n'est vrai que si $A(t)$ et $A(s)$ commutent (pour tout s et t). Cette propriété n'est vérifiée que pour des dépendances en temps très particulière (comme par exemple $A(t) = a(t)A$ où a est une fonction scalaire, et A une matrice fixée).

Exercice 5 *On considère le problème $\dot{x} = A(t)x$ avec*

$$A(t) = \begin{bmatrix} 1 & e^t \\ 0 & 0 \end{bmatrix}.$$

Pour une condition initiale $x(0)$ quelconque, calculer $x(t)$, et comparer à $\exp(\int_0^t A(s)\,ds)x(0)$. Commenter.

2.2 Le cas non linéaire

Théorème 3 (Théorème de Cauchy-Lipschitz) *On suppose que f est continue sur $I \times O$ et localement lipschitzienne par rapport à la variable x, c'est-à-dire : $\forall(\overline{t},\overline{x}) \in I \times O$, il existe un voisinage $(\overline{t}-\alpha,\overline{t}+\alpha) \times \mathcal{B}(\overline{x},\varepsilon) \subset I \times O$ de $(\overline{t},\overline{x})$ tel que f restreint à $(\overline{t}-\alpha,\overline{t}+\alpha) \times \mathcal{B}(\overline{x},\varepsilon)$ soit lipschitzienne par rapport à x.*

Alors, $\forall(\overline{t},\overline{x}) \in I \times O$, il existe un $\alpha > 0$ et un voisinage de $(\overline{t},\overline{x})$ (inclus dans $I \times O$) tel que, pour toute condition initiale (t_0,x_0) dans ce voisinage le problème de Cauchy

$$\begin{cases} \dot{x} = f(t,x), \\ x(t_0) = x_0, \end{cases}$$

admette une unique solution sur un intervalle de temps $(t_0-\alpha, t_0+\alpha)$.

Nous admettons ce théorème. La démonstration est très proche du cas linéaire (argument de point fixe), mais nécessite de se restreindre à des voisinages sur lesquels f est lipschitzienne, et tels que toute solution reste confinée dans ce voisinage.

Il est important de remarquer que la largeur 2α de l'intervalle de définition est indépendant du couple (t_0,x_0) définissant la condition initiale.

Noter également que toute fonction f de classe $\mathcal{C}^1$ est localement Lipschitzienne.

Les deux exercices suivant donnent des conditions plus faibles que le caractère localement lipschitzien pour assurer respectivement l'existence et l'unicité d'une solution.

Exercice 6 (Théorème de Cauchy Peano) *Soit une fonction f continue sur $[0,T]\times\mathbb{R}$. Montrer qu'il existe une solution x de classe $\mathcal{C}^1$ sur $[0,T]$ telle que*

$$\begin{cases} \dot{x}(t) = f(t, x(t)), \\ x(0) = x_0. \end{cases}$$

Sous ces hypothèses, on n'a pas unicité de la solution (considérer par exemple $\dot{x} = \sqrt{x}$ de condition initiale $x(0) = 0$, et montrer que ce problème admet une infinité de solutions). La démonstration est un peu délicate... S'aider de Google !

Exercice 7 (Théorème d'Osgood) *Soit f une fonction définie sur $\mathbb{R}$ et telle que, pour un $\epsilon > 0$, $\int_0^\epsilon \frac{1}{w(f,\delta)} d\delta = +\infty$ où w est le* module de continuité *de f : $w(f,\delta) = \sup_{|x-y|\leq\delta} |f(x) - f(y)|$.*

Montrer que si on dispose de deux solutions pour le problème de Cauchy

$$\begin{cases} \dot{x}(t) = f(x(t)), \\ x(0) = x_0, \end{cases}$$

alors, ces deux solutions coïncident.

En utilisant le théorème de Cauchy Lipschitz, on peut définir la notion de *solution maximale*, c'est-à-dire des solutions définies sur un intervalle ouvert $J \subset I$ qui ne sont pas prolongeables en des solutions sur un intervalle plus grand que J. Lorsque $J = I$, on parle de *solution globale*.

L'existence de solution maximale repose sur le lemme de recollement :

Lemme 1 *Si deux solutions u_1 et u_2 sont définies sur des intervalles J_1 et J_2 et sont telles que $\exists t_0 \in J_1 \cap J_2$, $u_1(t_0) = u_2(t_0)$, alors elles coïncident sur $J_1 \cap J_2$.*

Preuve : L'ensemble des points où les deux solutions coïncident est non vide (car $u_1(t_0) = u_2(t_0)$), ouvert (par l'unicité du théorème de Cauchy Lipschitz) et fermé (car u_1 et u_2 sont continues) dans $J_1 \cap J_2$, donc égal à $J_1 \cap J_2$, car $J_1 \cap J_2$ est connexe. $\diamondsuit$

Pour savoir si il existe une solution globale, le théorème suivant est fondamental :

Théorème 4 *Soit une solution maximale x, définie sur un intervalle J. Alors, si $\sup J < \sup I$, x sort définitivement de tout compact de $\mathbb{R}^d$ inclus dans O : $\forall K$ compact inclus dans O, $\exists t^* \in J$ tel que $\forall t \in J, t > t^*$, $x(t) \notin K$.*

On a évidemment un résultat analogue pour la borne inférieure de J.

Preuve : On raisonne par l'absurde : sinon, il existe une suite t_n croissante de temps convergeant vers $\sup J$ et telle que $x(t_n) \in K$, donc telle que $x(t_n)$ converge vers un x^∞ (quitte à extraire une sous-suite). D'après le théorème de Cauchy Lipschitz, il existe une solution du problème de Cauchy définie sur un intervalle de largeur 2α, pour tout couple (t_0', x_0') dans un voisinage de $(\sup J, x^\infty)$ et définissant

la condition initiale. Ceci vaut donc pour $(t_m, x(t_m))$ si m est assez grand, et permet alors de prolonger la solution sur $(t_m, t_m + \alpha)$, et $t_m + \alpha$ sera plus grand que $\sup J$ pour m assez grand, d'où la contradiction. $\diamondsuit$

Donc, si $O = \mathbb{R}^d$ et J est strictement inclus dans I, la solution tend vers $\pm\infty$ (on dit que la solution "explose") au borne de J. Réciproquement, si on montre que la solution ne peut pas exploser aux bornes de J, alors la solution est globale.

Pour prouver qu'une solution est globale, on a donc besoin de résultats du type : "si une solution existe, alors elle reste dans un compact". C'est ce qu'on appelle des estimations *a priori*. Un outil bien utile pour établir ces estimées est le lemme de Gronwall, dont il faut plutôt retenir la démonstration que l'énoncé.

Lemme 2 (Lemme de Gronwall) *Soit $u \in \mathcal{C}^0([0,T],\mathbb{R})$ et telle que $a, b \in \mathcal{C}^0([0,T],\mathbb{R})$ avec $a \geq 0$ et $u(t) \leq \int_0^t a(s)u(s)\,ds + b(t)$. Alors $u(t) \leq \int_0^t b(s)a(s)\exp\left(\int_s^t a(r)\,dr\right)ds + b(t)$.*

Preuve : L'astuce consiste à poser $v(t) = \int_0^t a(s)u(s)\,ds$. On a $\dot{v} = au \leq a(v+b)$. On en déduit que $\frac{d}{dt}\left(\exp\left(-\int_0^t a(r)\,dr\right)v(t)\right) \leq \exp\left(-\int_0^t a(r)\,dr\right)a(t)b(t)$, d'où le résultat après intégration. $\diamondsuit$

Remarque 4 *Le principe de la démonstration s'applique aussi pour donner des bornes sur une fonction u telle que $\dot{u} \leq au + b$, cette fois sans hypothèses sur le signe de a.*

Exercice 8 (Autour de Gronwall)

1. *Une version discrète de Gronwall : on suppose qu'une suite u_n vérifie $u_n \leq \sum_{k=0}^{n-1} a_k u_k + b_n$, avec $a_k \geq 0$. En déduire une majoration des u_n.*
2. *Soit $u \in \mathcal{C}^0([0,T],\mathbb{R})$ et telle que $a, b \in \mathcal{C}^0([0,T],\mathbb{R})$ avec $a \geq 0$. Que devient le lemme de Gronwall si on suppose que $0 \leq u(t) \leq \int_0^t a(s)\sqrt{u}(s)\,ds + b(t)$? Et si $u(t) \leq \int_0^t a(s)u^2(s)\,ds + b(t)$?*

Voici une application du lemme de Gronwall. Si $f(t,x)$ définie sur $\mathbb{R} \times \mathbb{R}^d$ croît au plus linéairement à l'infini, c'est-à-dire si il existe des constantes α et β telles que $|f|(t,x) \leq \alpha|x| + \beta$, alors la solution du problème de Cauchy est globale. Par exemple, une fonction $f(t,x)$ globalement lipschitzienne par rapport à x croît au plus linéairement à l'infini.

Exercice 9 *Faire la preuve de ce résultat.*

Si le second membre croît à l'infini plus rapidement que linéairement, les solutions maximales peuvent ne pas être globales. Considérons par exemple le problème :

$$\begin{cases} \dot{x} = x^2, \\ x(t_0) = x_0. \end{cases}$$

Si $x_0 = 0$, la solution maximale est globale : $x = 0$. En revanche, si $x_0 \neq 0$, la solution maximale $x(t) = (x_0^{-1} + t_0 - t)^{-1}$ est définie sur $(-\infty, t_0 + x_0^{-1})$ si $x_0 > 0$ et sur $(t_0 + x_0^{-1}, +\infty)$ si $x_0 < 0$. Les intervalles de définition de la solution sont définies de manière à ce qu'ils contiennent t_0.

Le lemme de Gronwall permet de montrer le

Lemme 3 (Lemme de comparaison) *soit x et y définis pour $t \geq 0$ et tels que $\dot{x} = f(t,x)$, $\dot{y} \geq f(t,y)$ et $x(0) = y(0)$. On suppose que f est localement lipschitzienne par rapport à x. On a alors, pour tout $t \geq 0$, $x(t) \leq y(t)$.*

Preuve : Faisons la preuve en supposant que f est globalement k-lipschitzienne (l'adaptation au cas localement lipschitzien est laissée au lecteur). Soit $z = x - y$. On a $z(0) = 0$ et $\dot{z} \leq k|z| = kz\,\mathrm{sgn}(z)$, où $\mathrm{sgn}(z) = 1_{z\geq 0} - 1_{z<0}$ désigne le signe de z. On pose alors $u(t) = \exp\left(-k\int_0^t \mathrm{sgn}(z(s))\,ds\right) z(t)$ et on a $\dot{u} \leq 0$, et donc $u \leq 0$. $\diamondsuit$

Exercice 10 *Dans le lemme de comparaison, que peut-on dire pour $t \leq 0$?*

Exercice 11 *Monter que si le potentiel V est de classe $\mathcal{C}^1$ et borné inférieurement, alors il existe une solution globale au problème (2).*

Solution *: on peut supposer $V \geq 0$ (quitte à considérer $V + \inf V$). Remarquer que $\frac{1}{2}m\|v\|^2 + V(x)$ est conservé sur la trajectoire, donc, en particulier, v reste borné en temps fini. Donc x reste borné en temps fini.*

Exercice 12 *Montrer qu'il existe une unique solution définie pour tout temps $t \geq t_0$ du problème de Cauchy*

$$\begin{cases} \dot{x} = \frac{1}{x} + g(t,x), \\ x(t_0) = x_0 > 0, \end{cases}$$

où g est une fonction k lipschitzienne de la variable x (avec k indépendant de t).

Définition 1 *On peut définir* le flot $\phi(t_0, x_0; t)$ *qui à une donnée initiale $x(t_0) = x_0$ associe $x(t)$, où x est la solution du problème de Cauchy. C'est une fonction définie sur un ouvert de $I \times O \times I$, qui est continue par rapport à ses trois variables, et qui vérifie* la relation de flot *: $\forall t_0 < s < t$, $\forall x_0$,*

$$\phi(s, \phi(t_0, x_0; s); t) = \phi(t_0, x_0; t).$$

Dans le cas linéaire (cf. section 2.1), $\phi(t_0, x_0; t) = R(t_0, t)x_0$, et la relation de flot s'écrit donc :

$$R(s,t)R(t_0,s) = R(t_0,t).$$

Remarque 5 *Si $v(t,x)$ désigne un champ de vitesse d'un fluide dans un domaine de $\mathbb{R}^3$, les particules de fluides satisfont l'équation différentielle ordinaire $\dot{x} = v(t,x)$. Le flot associé à cette équation donne donc les trajectoires des particules du fluide. On rappelle qu'il ne faut pas confondre trajectoires et lignes de courant (les lignes de courant correspondent aux orbites du problème $\dot{x} = v(t_0, x)$, pour un instant t_0 fixé).*

Remarque 6 *On peut en fait montrer que si f est de classe $\mathcal{C}^r$ ($r \geq 1$) sur $I \times O$, alors pour t_0 fixé, la fonction $(x_0, t) \mapsto \phi(t_0, x_0; t)$ est une application de classe $\mathcal{C}^r$ sur son domaine de définition.*

Nous terminons cette section en analysant la sensibilité des solutions à des variations des données.

Lemme 4 (Sensibilité à des perturbations) *Soit $f(t,x)$ une fonction continue en (t,x) et globalement L-lipschitzienne en x, et $\varepsilon(t)$ une fonction continue. Soit x solution de*

$$\begin{cases} \dot{x} = f(t,x), \\ x(t_0) = x_0, \end{cases}$$

et y solution de

$$\begin{cases} \dot{y} = f(t,y) + \varepsilon(t), \\ y(t_0) = y_0. \end{cases}$$

Alors,

$$\forall t \in [0,T],\ |x-y|(t) \leq \exp(LT)\left(|x_0 - y_0| + \int_0^T |\varepsilon|(t)\,dt\right). \tag{6}$$

Exercice 13 *En utilisant le lemme de Gronwall, et en s'inspirant de la preuve du Lemme de comparaison 3, prouver le Lemme 4.*

Ce lemme démontre une propriété de *stabilité* par rapport aux perturbations, qui repose essentiellement sur le caractère lipschitzien de f. Ce lemme (et le théorème de Cauchy Lipschitz) montre que le problème est *bien posé au sens de Hadamard* : il admet une unique solution qui dépend continûment des données du problème. De manière générale, on dit d'une solution à un problème qu'elle est stable si elle est peu sensible à des variations des données. La sensibilité est mesurée par *une constante de stabilité.*

Remarquer que la constante de stabilité $\exp(LT)$ dans (6) devient rapidement très grande quand T augmente, ou quand L est grand, et que donc ce résultat est sans grande utilité quand il s'agit de comprendre la sensibilité par rapport à des perturbations en temps long, ou pour des problèmes pour lesquels L est grand. Ceci sera l'objet de la Section 3.

Exercice 14 (Une pierre qui tombe)

On s'intéresse au mouvement d'une pierre qui tombe d'une hauteur h. On note r la distance de la pierre à la surface de la Terre. Les conditions initiales sont $r(0) = h$ et $\dot{r}(0) = 0$. Les équations du mouvement sont, pour le modèle exact :

$$\ddot{r} = -\frac{\gamma M}{(R+r)^2}$$

où γ est la constante de gravitation, R et M sont le rayon et la masse de la Terre. Un modèle approché et donné par :

$$\ddot{r} = -g,$$

où $g = \gamma M/R^2$ est l'accélération de la gravité sur Terre.

1) Résoudre analytiquement l'équation du mouvement pour le modèle approché. En combien de temps la pierre touche le sol ?

2) Montrer que le modèle exact admet une unique solution. Comparer les temps de chute pour les deux modèles. Indication : on pourra majorer et minorer la force dans le modèle exact.

3) On pose $\epsilon = 1/R$. Le modèle exact s'écrit

$$\ddot{r} = -\frac{\gamma M \epsilon^2}{(1+\epsilon r)^2}.$$

On suppose que la solution $r(t,\epsilon)$ est régulière par rapport à ses deux arguments. Montrer que

$$r(t) = h - g(1-h/R)t^2/2 + O(1/R^4).$$

Indication : on pourra utiliser l'Ansatz (*i.e.* chercher une solution sous la forme) $r(t,\epsilon) = r_0(t) + r_1(t)\epsilon + r_2(t)\epsilon^2 + r_3(t)\epsilon^3 + O(\epsilon^4)$.

2.3 Bilan

Nous avons répondu aux questions concernant le caractère bien posé des équations différentielles ordinaires. Mais nous avons noté que les constantes de stabilité ne sont pas informatives pour comprendre ce qui se passe "sur des temps longs".

Dans la suite, nous nous proposons d'étudier le comportement qualitatif des solutions selon trois points de vue :

- Section 3 : La constante de stabilité exponentielle dans (6) est parfois très pessimiste (et heureusement !). Ceci nous amènera à étudier la stabilité autour de solutions d'équilibre.
- Section 4 : Pour deux conditions initiales proches, les trajectoires peuvent vite diverger (cf. par exemple la Section 1.3). Mais la vision déterministe (conditions initiales fixées et déterministes) peut être "un mauvais point de vue". Nous introduirons la notion d'ergodicité,
- Section 5 : Une question naturelle est de préciser dans quelle unité on mesure "un temps long". Il faut pour cela comprendre les échelles de temps du système. Ce sera l'objet d'une section sur les problèmes raides. Au passage, nous introduirons les principales difficultés liées à la discrétisation des équations différentielles ordinaires.

3 Etude de la stabilité des points d'équilibre

3.1 Equations différentielles ordinaires autonomes, portrait de phase et notions de stabilité

On considère dans cette section des équations différentielles autonomes :

$$\begin{cases} \dot{x} = f(x), \\ x(0) = x_0, \end{cases} \tag{7}$$

où f est de classe $\mathcal{C}^1$ sur un ouvert $O \subset \mathbb{R}^d$.

Définition 2 *On appelle orbite de x_0 l'ensemble des $x(t)$ pour t appartenant à l'intervalle de définition J de la solution globale x :*

$$\mathcal{O}(x_0) = \{\phi(0, x_0; t),\ t \in J\}.$$

Noter que par le théorème de Cauchy Lipschitz, comme f est de classe $\mathcal{C}^1$ sur O, les orbites ne peuvent pas se recouper. On appelle *portrait de phase* la partition de O en orbites. En pratique, on trace sur un portrait de phase quelques orbites remarquables, et l'orientation du champ de vecteur dans quelques zones. On peut tracer le portrait de phase sans résoudre explicitement l'équation, et ceci permet d'avoir des informations importantes sur le comportement qualitatif des solutions (cf. par exemple la Section 3.5 ci-dessous).

Exercice 15 *Reprenons l'exemple des équations de Lotka-Volterra de la Section 1.2. Pourquoi si $N(0) > 0$ et $P(0) > 0$, alors $N(t) > 0$ et $P(t) > 0$ pour tout $t \geq 0$? En utilisant le fait que $H(N, P) = cN - d \ln N + bP - a \ln P$ est constant le long des trajectoires, tracer le portrait de phase pour l'équation de Lotka-Volterra. En déduire que les solutions sont périodiques en temps.*

Dans le cas autonome, le flot introduit dans la Définition 1 est en fait une fonction de deux variables. En effet, on a, pour tout $s, t > 0$ et tout $x_0 \in O$,

$$\phi(0, x_0; t) = \phi(s, x_0; t + s). \tag{8}$$

Dans le cas autonome, on notera

$$\phi(t, x_0) = \phi(0, x_0; t).$$

Exercice 16 *Prouver l'assertion* (8) *précédente.*

Définition 3 *On appelle* point d'équilibre*, ou* point fixe*, ou* point stationnaire *tout point x^* tel que $f(x^*) = 0$.*

Remarquer que si x^* est un point d'équilibre, alors l'orbite de x^* est réduit à x^* : $\mathcal{O}(x^*) = \{x^*\}$.

Dans le cas linéaire $f(x) = Ax$, les points d'équilibre sont les x tels que $Ax = 0$ (et 0 est donc toujours un point d'équilibre).

Définition 4 *On dit qu'un point fixe x^* est* stable *si pour tout $\varepsilon > 0$, il existe $\delta > 0$ tel que, si $\|x - x^*\| \leq \delta$, alors, pour tout $t \geq 0$, $\|\phi(t, x) - x^*\| \leq \epsilon$.*

En particulier, si x^* est un point fixe stable et x est suffisamment porche de x^*, la fonction $\phi(t, x)$ est bien définie pour tout temps $t \geq 0$ (puisqu'elle reste confinée dans un compact). On qualifie parfois cette notion de stabilité de stabilité au sens de Lyapunov.

Exercice 17 *Que dire de la stabilité du point d'équilibre $(0, 0)$ pour les deux systèmes suivants :* $\begin{cases} \dot{x} = -y \\ \dot{y} = x \end{cases}$ $\begin{cases} \dot{x} = -x \\ \dot{y} = y \end{cases}$

Définition 5 *On dit qu'un point fixe x^* est* asymptotiquement stable *si il est stable et qu'il existe un voisinage de x^* tel que pour tout x dans ce voisinage, $\lim_{t \to \infty} \phi(t, x) = x^*$.*

L'exercice 19 montre que la condition "il existe un voisinage de x^* tel que pour tout x dans ce voisinage, $\lim_{t\to\infty}\phi(t,x)=x^*$" n'implique pas la stabilité.

Remarquer que pour ces notions de stabilité, la flèche du temps compte (on regarde le comportement pour des temps longs *positifs*).

Exercice 18 *Comparer ces notions de stabilités à celle évoquée dans le Lemme 4.*

Exercice 19 *On considère le problème :*

$$\begin{cases} \dot{x} = x - y - x(x^2+y^2) + \frac{xy}{\sqrt{x^2+y^2}}, \\ \dot{y} = x + y - y(x^2+y^2) - \frac{x^2}{\sqrt{x^2+y^2}}. \end{cases}$$

Montrer que le point $x^*=(1,0)$ n'est pas stable, bien qu'il existe un voisinage de x^* tel que pour tout x dans ce voisinage, $\lim_{t\to\infty}\phi(t,x)=x^*$. (On pourra montrer que le système se réécrit en polaire $\dot{r}=r(1-r^2)$, $\dot{\theta}=2\sin^2(\theta/2)$.)

Exercice 20 *Le comportement qualitatif en temps long des solutions à une équations différentielle ordinaire peut beaucoup changer pour de petites variations de paramètres du problème (ceci est l'objet de la théorie des bifurcations). A titre d'exemple, on vérifiera les assertions suivantes :*

- *Le problème*
$$\dot{x} = \mu x - x^3$$
a un unique point fixe stable pour $\mu \leq 0$, qui devient instable pour $\mu > 0$, tandis que deux nouveaux points d'équilibre stables apparaissent, pour $\mu > 0$.
- *Le problème*
$$\dot{x} = \mu + x^2$$
a deux points fixes pour $\mu < 0$, un stable et un instable, qui se rejoignent pour $\mu = 0$ puis disparaissent pour $\mu > 0$.

Il est très important en pratique d'étudier la stablité des points d'équilibre. En effet, dans plusieurs modèles, les points d'équilibres correspondent à des "points de fonctionnement" du système, et savoir si on ne s'écarte pas trop de ces points de fonctionnement sous des petites perturbations est primordial pour assurer la fiabilité du système.

Il existe deux méthodes "systématiques" d'étude de la stabilité des points fixes : l'analyse par fonction de Lyapunov (Section 3.3), et l'analyse de stabilité linéaire (Section 3.4). Nous commençons par analyser en détail le cas linéaire.

3.2 Le cas linéaire

On considère le cas $f(x)=Ax$, et on étudie la stabilité du point fixe $x^*=0$.

Théorème 5 *Le point $x^*=0$ est asymptotiquement stable si et seulement si toutes les valeurs propres de A sont de partie réelle strictement négative.*

Preuve : Notons $\lambda_1, \lambda_2, ..., \lambda_p$ les valeurs propres (distinctes) de A, de multiplicités algébriques $m_1, m_2, ..., m_p$ (la matrice A est de dimension $m_1 + m_2 + \ldots + m_p$, et m_i est la puissance associée au monôme $(x - \lambda_i)$ dans le polynôme caractéristique $\det(A - \lambda \mathrm{Id})$ de A). Quitte à réordonner les valeurs propres, on peut supposer qu'il existe un entier q tel que $\lambda_1, \ldots, \lambda_{2q}$ ne sont pas réelles, et $\lambda_{2q+1}, \ldots, \lambda_p$ sont réelles. De même, on peut supposer que $\lambda_1 = \overline{\lambda_2}, \ldots, \lambda_{2q-1} = \overline{\lambda_{2q}}$ (où $\overline{\lambda}$ désigne le complexe conjugué de λ). En utilisant la décomposition de Jordan, on montre que les coefficients de $\exp(tA)$ sont des combinaisons linéaires (à coefficients réels) de fonctions du type $t^k \exp(\mathrm{Re}\,(\lambda_j)t)\cos(\mathrm{Im}\,(\lambda_j)t)$, $t^k \exp(\mathrm{Re}\,(\lambda_j)t)\sin(\mathrm{Im}\,(\lambda_j)t)$, pour $1 \leq j \leq 2q$, et de fonctions du type $t^k \exp(\lambda_j t)$ pour $2q+1 \leq j \leq p$, avec $0 \leq k \leq m_j - 1$. Le cas $k = 0$ correspond au cas où $\mathrm{Ker}(A - \lambda_i \mathrm{Id}) = \mathrm{Ker}(A - \lambda_i \mathrm{Id})^{m_i}$: le sous-espace propre est égal au sous-espace caractéristique.

Ces résultats reposent sur le fait que :

$$\exp\left(t\begin{bmatrix} \lambda_j & 1 & 0... & 0 \\ 0 & \lambda_j & 1 & 0 \\ 0 & ...\,0 & \lambda_j & 1 \\ 0 & ... & 0 & \lambda_j \end{bmatrix}\right) = \exp(\lambda_j t)\begin{bmatrix} 1 & t & ... & \frac{t^{m_j-1}}{(m_j-1)!} \\ 0 & 1 & t & ... \\ 0 & ...\,0 & 1 & t \\ 0 & ... & 0 & 1 \end{bmatrix}.$$

♢

Exercice 21 *Reprendre cette démonstration, pour montrer que le point $x^* = 0$ est stable si et seulement si les valeurs propres de A sont de partie réelle négative ou nulle, et que pour les valeurs propres de A de partie réelle nulle le sous-espace propre associé est égal au sous-espace caractéristique.*

Exercice 22 (Portraits de phase en dimension 2) *On considère le problème de Cauchy linéaire*

$$\dot{x} = Ax$$

avec x de dimension 2 et A une matrice 2×2. Tracer le portrait de phase en distingant les cas :

- *Les deux valeurs propres sont réelles et de signe opposé (point selle),*
- *Les deux valeurs propres sont réelles distinctes et de même signe (noeud impropre, attractif ou répulsif),*
- *Il y a une seule valeur propre (nécessairement réelle) et la matrice est diagonalisable (noeud propre, attractif ou répulsif),*
- *Il y a une seule valeur propre (nécessairement réelle) et la matrice n'est pas diagonalisable (noeud exceptionnel, attractif ou répulsif),*
- *Les deux valeurs propres sont complexes conjuguées et de partie réelle non nulle (foyer, attractif ou répulsif),*
- *Les deux valeurs propres sont complexes conjuguées et imaginaires pures (centre).*

3.3 Méthode par fonction de Lyapunov

Définition 6 *Une fonction de Lyapunov L pour l'équation différentielle ordinaire (7) en un point d'équilibre x^* est :*

- *une fonction* continue $L : V \to \mathbb{R}$, *définie sur un voisinage V de x^*,*

– *telle que x^* est* un minimum strict *de L sur V (on peut supposer sans restriction que la valeur au minimum est 0) : $L(x^*) = 0$ et $L(x) > 0$ si $x \neq x^*$,*
– *et telle que pour tout $x \in V$ tel que $x \neq x^*$, $t \mapsto L(\phi(t, x))$ est* strictement décroissante *(sur un intervalle ouvert autour de 0 pour lequel l'application est définie).*

Pour vérifier la troisième propriété, si L est de classe $\mathcal{C}^1$, il suffit de vérifier que $\nabla L(x) \cdot f(x) < 0$ pour tout $x \neq x^*$.

Théorème 6 *Si le système admet une fonction de Lyapunov au voisinage du point d'équilibre x^*, ce point d'équilibre est asymptotiquement stable.*

Preuve : On peut supposer V relativement compact (*i.e.* de fermeture compacte), quitte à restreindre V. Soit B une boule centrée en x^* incluse dans V, et $\varepsilon > 0$ le minimum de L sur $V \setminus B$. Il est clair que $L^{-1}[0, \varepsilon/2]$ est un compact inclus dans B. (Autrement dit, on peut supposer que le voisinage V dans la Définition 6 est en fait de la forme $L^{-1}[0, \eta)$, pour un $\eta > 0$.) En utilisant ce procédé, on démontre facilement la stabilité (à vérifier en exercice). Reste à prouver l'asymptotique stabilité.

Soit $x \in L^{-1}[0, \varepsilon/2]$. Etant donné que l'on sait que $t \mapsto L(\phi(t, x))$ est décroissante strictement, on a que $\phi(t, x)$ reste dans le compact $L^{-1}[0, \varepsilon/2]$. En particulier, le flot $\phi(t, x)$ est défini pour tout temps $t \in [0, \infty)$. On note a la limite de $L(\phi(t, x))$ quand $t \to \infty$. On peut construire une suite de points $t_n \to \infty$ telle que $\phi(t_n, x)$ tend vers un x_∞. On a évidemment $L(x_\infty) = a$.

Supposons que $x_\infty \neq x^*$. On a donc, pour un $t > 0$, $L(\phi(t, x_\infty)) < a$. Or $L(\phi(t, x_\infty)) = \lim_{n\to\infty} L(\phi(t, \phi(t_n, x))) = \lim_{n\to\infty} L(\phi(t + t_n, x)) = a$. D'où une contradiction. Par conséquent, $x_\infty = x^*$, et le point x^* est bien asymptotiquement stable. ♢

Exercice 23 *Que se passe-t'il si on considère des fonctions de Lyapunov L telles que pour tout $x \in V$ tel que $x \neq x^*$, $t \mapsto L(\phi(t, x))$ est décroissante sur un intervalle ouvert autour de 0 sur lequel l'application est bien définie (mais non pas strictement décroissante) ?*

Solution *: En reprenant la preuve, on montre facilement que le point d'équilibre est stable (et non pas asymptotiquement stable).*

Exercice 24 *On considère le système hamiltonien (2). On suppose H régulier. Les points d'équilibres sont les points critiques de H (que l'on suppose isolés). Montrer que les minima locaux de H sont des points d'équilibres stables. Appliquer ce résultat au cas du pendule sans frottement.*

Solution *: Prendre H moins sont minimum comme fonction de Lyapunov. Pour le pendule sans frottement, on voit donc que les points $(x, v) = (2k\pi, 0)$ sont stables.*

3.4 Méthode par linéarisation

Au voisinage d'un point d'équilibre x^*, le champ de vecteur f s'écrit $f(x) = Df(x^*)(x - x^*) + o(\|x - x^*\|)$ car $f(x^*) = 0$, où $Df(x^*)$ désigne la matrice de composantes $\left(\frac{\partial f_i}{\partial x_j}(x^*)\right)_{i,j}$. En pensant aux résultats de la Section 3.2, on comprend que la stabilité de x^* puisse être déduite de conditions sur le spectre de $Df(x^*)$.

Théorème 7 *Si toutes les valeurs propres de la matrice $Df(x^*)$ sont de parties réelles strictement négatives, alors le point d'équilibre est asympotiquement stable.*

Preuve : Quitte à effectuer un changement de variable $x \to x - x^*$, on peut supposer $x^* = 0$. La preuve consiste à exhiber une fonction de Lyapunov.

On commence par écrire un développement limité de f autour de $x^* = 0$:

$$f(x) = Df(0)x + |x|\epsilon(x),$$

avec ϵ une fonction définie sur un voisinage de 0 et telle $\lim_{x\to 0} |\epsilon(x)| = 0$. Dans la suite, on note

$$A = Df(0),$$

et $|\cdot|$ désigne la norme euclidienne pour les vecteurs, ou bien la norme triple associée pour les matrices. On introduit *le système linéarisé* :

$$\dot{y} = Ay.$$

D'après les résultats de la Section 3.2, on sait que si toutes les valeurs propres de la matrice $A = Df(0)$ sont de parties réelles strictement négatives, alors le point d'équilibre 0 est asympotiquement stable pour le système linéarisé. Il faut donc montrer que le système non-linéaire se comporte comme le système linéarisé au voisinage du point d'équilibre.

On introduit pour cela la fonction de Lyapunov

$$L(x) = \int_0^\infty |\exp(tA)x|^2 \, dt.$$

Noter que L est bien définie car on suppose que les valeurs propres de A sont de parties réelles strictement négatives, ce qui implique qu'il existe $\alpha, c > 0$ tel que

$$|\exp(tA)| \leq c \exp(-\alpha t).$$

Il est clair que 0 est un minimum strict de L sur un voisinage de 0. On va vérifier que $\nabla L \cdot f < 0$ en dehors de $x^* = 0$:

$$\begin{aligned}
\nabla L(x) \cdot f(x) &= \int_0^\infty 2 \exp(tA)x \cdot (\exp(tA) f(x)) \, dt, \\
&= \int_0^\infty 2 \exp(tA)x \cdot (\exp(tA) Ax) \, dt + \int_0^\infty 2 \exp(tA)x \cdot (\exp(tA)|x|\epsilon(x)) \, dt, \\
&\leq \int_0^\infty \frac{d}{dt} |\exp(tA)x|^2 \, dt + \int_0^\infty 2 |\exp(tA)|^2 |x|^2 |\epsilon(x)| \, dt, \\
&\leq -|x|^2 + C|x|^2|\epsilon(x)|.
\end{aligned}$$

Ceci montre qu'il existe un voisinage de 0 tel que $\nabla L \cdot f < 0$ en dehors de 0. La fonction L est donc une fonction de Lyapunov, et le point $x^* = 0$ est asymptotiquement stable d'après le Théorème 6. ♢

Exercice 25 *Reprendre la preuve précédente pour vérifier que si une des valeurs propres est de partie réelle strictement positive, alors le point d'équilibre est instable.*

Attention : si une des valeurs propres est de partie réelle nulle, la méthode par linéarisation ne permet pas de conclure sur la stabilité du point fixe. On pourra à titre d'exemple étudier la stabilité du point fixe $x^* = 0$ pour le problème $\dot{x} = \lambda x^3$, en fonction du signe de λ.

Remarque 7 *Cette méthode d'analyse de la stabilité en utilisant le spectre du problème linéarisé est spécifique à la dimension finie : si on considère un problème du type $\partial_t u = Au$ avec A un opérateur agissant sur la fonction u (u est une fonction du temps et d'une variable d'espace et A est un opérateur différentiel par exemple), et u^* un point d'équilibre pour l'opérateur A ($Au^* = 0$), alors, le fait que les valeurs propres de la différentielle de l'opérateur au point u^* soient de partie réelle strictement négative n'implique pas en général que u^* soit asymptotiquement stable. Les résultats de ce type en dimension infinie reposent plutôt en général sur des méthodes de type fonctions de Lyapunov. Les fonctions de Lyapunov s'appellent parfois énergie, ou entropie, car elles correspondent souvent à des quantités physiques.*

3.5 Le cas du pendule

On considère le problème du pendule sans frottement :

$$\begin{cases} \dot{x} = v, \\ \dot{v} = -\sin x, \end{cases} \tag{9}$$

où $x \in 2\pi\mathbb{R}/\mathbb{Z}$ et $v \in \mathbb{R}$. Ce système est un système hamiltonien, qui admet donc une intégrale première :

$$E(x, v) = \frac{v^2}{2} - \cos x.$$

Le fait que cette quantité soit conservée le long des trajectoires implique que les solutions sont définies pour tout temps $t \in \mathbb{R}$.

Pour tracer le portrait de phase dans le plan (x, v), on commence par tracer les lignes de niveau de la fonction E : $C_\alpha = \{(x, v),\ E(x, v) = \alpha\}$. Pour $\alpha > 1$, on obtient des lignes de niveaux correspondant à des trajectoires pour lesquelles le pendule effectue une rotation complète. Pour $-1 < \alpha < 1$, on obtient des lignes de niveaux correspondant à des trajectoires pour lesquelles le pendule oscille entre deux angles maximaux. Pour $\alpha = -1$, on obtient les points $(2k\pi, 0)$.

Le cas $\alpha = 1$ correspond à un cas très particulier. Remarquer que si $(x, v) \in C_1 \cap \{v > 0\}$, alors $v = \sqrt{2(1 + \cos(x))}$. On en déduit que pour une trajectoire telle que $x(0) \in C_1 \cap \{v > 0\}$, on a

$$t = \int_{x(0)}^{x(t)} \frac{1}{\sqrt{2(1 + \cos(u))}}\, du. \tag{10}$$

En effet, pour une telle trajectoire, $\frac{dx}{dt} = v(t) = \sqrt{2(1 + \cos(x(t)))}$. En faisant tendre t vers $+\infty$ dans (10), on voit que nécessairement $\lim_{t\to\infty} x(t) = \pi$. De même, on montre ainsi que $\lim_{t\to-\infty} x(t) = -\pi$. Cette trajectoire relie donc en temps infini les points $(-\pi, 0)$ et $(\pi, 0)$.

Exercice 26 *Réfléchir à quel mouvement du pendule correspond cette trajectoire.*

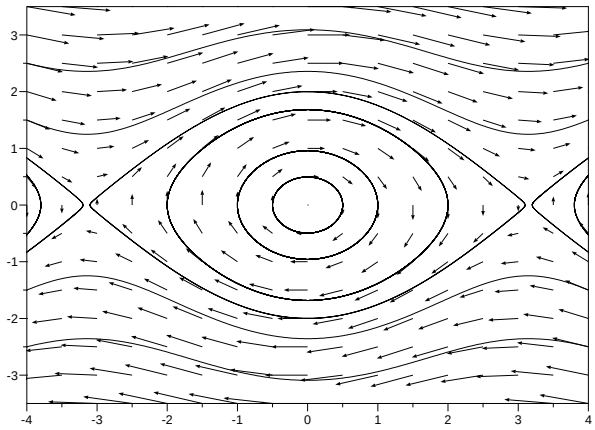

FIG. 4 – Portrait de phase pour le pendule sans frottement. EN abscisse, l'angle x, et en ordonnée, la vitesse v.

En regroupant ces résultats, on obtient le portrait de phase représenté sur la Figure 4.

Analysons maintenant les points d'équilibre du système : $x = k\pi$ ($k \in \mathbb{Z}$) et $v = 0$.

Les points d'équilibre $(x, v) = (2k\pi, 0)$ sont stables. En effet, on vérifie facilement qu'on reste dans un voisinage de $(2k\pi, 0)$ en utilisant le fait que les orbites ne se recoupent pas. On peut également le prouver en appliquant le critère de l'Exercice 23 avec E comme fonction de Lyapunov.

En revanche, les points $(x, v) = (\pi + 2k\pi, 0)$ sont instables. On peut le vérifier sur le portrait de phase en utilisant les trajectoires contenues dans C_1 que nous avons analysées ci-dessus. Ces trajectoires relient entre eux ces points d'équilibre qui sont donc instables[1]. On peut également utiliser un critère de stabilité linéaire. En effet, on vérifie que la différentielle du second membre de (9) s'écrit $\begin{bmatrix} 0 & 1 \\ -\cos x & 0 \end{bmatrix}$, et vaut donc au point $(\pi + 2k\pi, 0)$:

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Cette matrice admet une valeur propre strictement positive, et donc le point est instable d'après le résultat de l'Exercice 25.

Exercice 27 *Essayer d'appliquer le critère de stabilité linéaire pour prouver la stabilité des points $(x, v) = (2k\pi, 0)$. Interpréter en termes concrets les résultats obtenus.*

On considère maintenant le pendule avec frottement :

$$\begin{cases} \dot{x} = v, \\ \dot{v} = -\sin x - \lambda v, \end{cases} \tag{11}$$

[1]Une telle trajectoire qui relie deux points d'équilibre différents est appelée trajectoire hétérocline.

avec $\lambda > 0$. Quelques trajectoires typiques dans l'espace des phases sont tracées sur la Figure 5.

Exercice 28 *Interpréter les trajectoires observées sur la Figure 5 en terme de mouvements du pendule.*

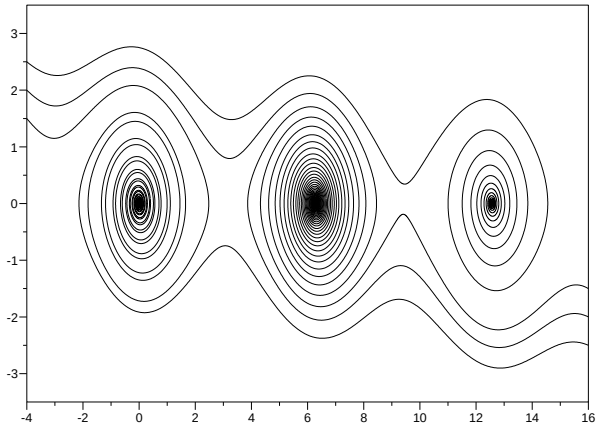

FIG. 5 – Quelques trajectoires $(x(t), v(t))_{t\in\mathbb{R}}$ pour le pendule avec frottement.

Les points d'équilibre de la dynamique sont toujours les $(x, v) = (k\pi, 0)$ ($k \in \mathbb{Z}$). En utilisant la fonction E comme fonction de Lyapunov et en appliquant le Théorème 6, on vérifie que les points $(2k\pi, 0)$ sont asymptotiquement stables. En effet

$$\nabla E \cdot \begin{pmatrix} v \\ -\sin x - \lambda v \end{pmatrix} = -\lambda v^2.$$

La fonction $t \mapsto E(\phi(t, (x, v)))$ (pour $(x, v) \neq (2k\pi, 0)$) est donc bien strictement décroissante sur une trajectoire car les points où $v = 0$ sont des points isolés de la trajectoire.

Exercice 29 *Essayer d'appliquer le critère de stabilité linéaire pour prouver la stabilité des points* $(x, v) = (2k\pi, 0)$.

Les points $(k + 2k\pi, 0)$ sont toujours instables. On peut par exemple le démontrer en regardant le spectre de la différentielle en ces points, qui s'écrit

$$\begin{bmatrix} 0 & 1 \\ 1 & -\lambda \end{bmatrix}.$$

On vérifie que cette matrice admet une valeur propre strictement positive, ce qui montre que les points sont instables (cf. Exercice 25).

4 Ergodicité

4.1 Chaos

Sur le problème de Lorenz (cf Section 1.3), on a observé que :

– la dynamique est extrêmement sensible aux conditions initiales,
– la dynamique a un comportement en temps long compliqué (ni convergence vers un point fixe, ni orbite périodique).

On parle de comportement chaotique. L'étude du problème de Lorenz dans sa généralité dépasse le cadre de ce cours introductif. On va se concentrer sur la dynamique associée aux maxima locaux de la variable z. On a observé que ces maxima vérifient

$$x_{n+1} = T(x_n)$$

avec T une application "en forme de tente". On se propose donc d'étudier le comportement de la dynamique des x_n, avec comme application T (cf. Figure 6) :

$$T : \begin{cases} [0,1] & \to & [0,1] \\ x & \mapsto & \begin{cases} 2x \text{ si } x \leq 1/2 \\ 2-2x \text{ si } x > 1/2 \end{cases} \end{cases}$$

Par analogie avec la Section précédente, on appelera orbite d'un point x l'ensemble

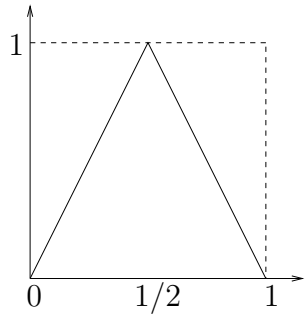

FIG. 6 – L'application tente T.

des points $\{x_n = T^n(x)\}_{n\geq 0}$ obtenus sur la trajectoire issue de x.

Observons tout d'abord que la dynamique des x_n présente les mêmes caractéristiques que la dynamique de Lorenz. Pour se faire, on va réécrire la dynamique T en décomposant les x_n en binaire. On introduit donc l'application $S : \{0,1\}^{\mathbb{N}\setminus\{0\}} \to [0,1]$ qui à une suite $(\alpha_j)_{j\geq 1}$ associe le réel

$$x = \sum_{j\geq 1} \alpha_j 2^{-j}.$$

L'application S n'est pas injective : par exemple, $S(1,0,0\ldots) = S(0,1,1\ldots) = 0.5$. On peut cependant la rendre injective en identifiant les suites $(\alpha_1, \alpha_2, ..., \alpha_N, 1, 0, 0, ...)$ et $(\alpha_1, \alpha_2, ..., \alpha_N, 0, 1, 1, ...)$.

Regardons comment l'application T s'écrit sur la représentation en binaire :

$$T \circ S(\alpha_1, \alpha_2, \ldots) = \begin{cases} S(\alpha_2, \alpha_3, \ldots) \text{ si } \alpha_1 = 0 \\ S(\overline{\alpha_2}, \overline{\alpha_3}, \ldots) \text{ si } \alpha_1 = 1 \end{cases}$$

où, pour un nombre $\alpha \in \{0,1\}$, $\overline{\alpha} = 1-\alpha$. De cette réécriture, nous pouvons déduire plusieurs conséquences sur la dynamique de $T^n(x)$:

– La dynamique est très sensible aux conditions initiales : en effet, si deux conditions initiales diffèrent dans leur développement en binaire seulement à partir de la N-ième décimale (elles sont donc très proches, puisque leur différence vaut au plus 2^{-N+1}), après N itérations, les deux trajectoires seront *a priori* complètement différentes.
– Les conditions initiales irrationnelles ont un comportement en temps long non périodique.

De plus, on vérifie que

– L'ensemble des orbites périodiques est dense dans $[0,1]$.

Exercice 30 *Représenter l'application T^n. En déduire que le système dynamique associé à T a 2^n points périodiques de période (au plus) n, et que les orbites périodiques sont denses dans $[0,1]$.*

Solution *: Les points périodiques de période au plus n sont exactement les points fixes de T^n. Par exemple, les points fixes de l'application T (points périodiques de période 1) sont $\{0, 2/3\}$. Les points périodiques de période 2 sont $\{0, 2/5, 2/3, 4/5\}$. De manière générale, on vérifie que $T^n(I_j^n) = [0,1]$ où $I_j^n = [j/2^n, (j+1)/2^n]$. Chaque intervalle I_j^n (pour $0 \leq j \leq 2^n - 1$) contient donc exactement un point fixe de l'application T^n. Il y a donc exactement 2^n points périodiques de période au plus n.*

Ceci montre aussi que pour tout $x \in [0,1]$ et tout $\epsilon > 0$, on trouve un point d'une orbite périodique dans l'intervalle $(x - \epsilon, x + \epsilon) \cap [0,1]$. Les orbites périodiques sont donc denses dans $[0,1]$.

Ces systèmes présentent une telle complexité de dynamique (on parle de systèmes chaotiques) qu'il semble naturel d'étudier des moyennes sur les trajectoires plutôt que les trajectoires elles-mêmes. De même, du fait de la sensibilité aux conditions initiales, il semble plus naturel d'étudier comment une distribution de conditions initiales va évoluer, ou bien comment on peut caractériser la dynamique pour presque toute condition initiale (pour une certaine mesure) plutôt que de regarder l'évolution de conditions initiales données ponctuellement. On va donc vers une description de nature probabiliste de la dynamique. On aura alors besoin de munir $[0,1]$ d'une mesure de probabilité, et d'une σ-algèbre.

Un autre exemple que l'on considèrera dans la suite est la rotation sur le tore :

$$R_\alpha \left\{ \begin{array}{ccc} \mathbb{T} & \to & \mathbb{T} \\ x & \mapsto & x + \alpha \end{array} \right. ,$$

où $\mathbb{T} = \mathbb{R}/\mathbb{Z}$ est le tore de dimension 1, et $\alpha \in \mathbb{R} \backslash \mathbb{Q}$. La dynamique est très différente du cas précédent : toute condition initiale engendre une orbite dense. Il n'y a pas d'orbites périodiques, pas de sensibilité aux conditions initiales.

Exercice 31 *Vérifier les assertions précédentes pour la dynamique de rotation sur le tore.*

4.2 Le théorème ergodique de Birkhoff

Généralisons l'étude. On considère un espace X muni d'une tribu (ou σ-algèbre) $\mathcal{A}$ et d'une mesure μ de probabilité définie sur $\mathcal{A}$, et une dynamique de la forme

$$x_{n+1} = T(x_n)$$

avec $T : X \to X$ une application mesurable qui préserve la mesure μ : pour tout ensemble $A \in \mathcal{A}$

$$\mu(T^{-1}(A)) = \mu(A).$$

On parle de *système dynamique mesuré* pour désigner la donnée $(X, \mathcal{A}, \mu, T)$. La suite des x_n est une trajectoire obtenue en itérant l'application T.

Remarque 8 (Mesure image) *Soit $T : X \to Y$ une application mesurable de $(X, \mathcal{A})$ dans $(Y, \mathcal{B})$. On se donne une mesure μ sur X. On peut alors définir la mesure image de μ par T notée $\mu \circ T^{-1}$ et définie par : pour tout $B \in \mathcal{B}$,*

$$\mu \circ T^{-1}(B) = \mu(T^{-1}(B)).$$

L'apparition de ce T^{-1} peut surprendre mais on peut se convaincre que c'est la bonne définition en utilisant une fonction test : pour toute fonction $f : Y \to \mathbb{R}$ mesurable et telle que $f \circ T \in L^1(X, \mathcal{A}, \mu)$,

$$\int f \, d(\mu \circ T^{-1}) = \int f \circ T d\mu.$$

Dire que T préserve μ est donc exactement dire que la mesure image de μ par T est μ. Dans ce cas, on a donc $\int f \circ T d\mu = \int f d\mu$ pour toute fonction $f \in L^1(X, \mathcal{A}, \mu)$.

Remarque 9 *En pratique, on connaît souvent $(X, \mathcal{A})$ et T. Trouver une mesure de probabilité μ invariante n'est pas toujours facile. Il peut y avoir plusieurs mesures invariantes ! Ceci dit, si X est un espace métrique compact, muni de la tribu des boréliens et T une application continue (donc mesurable), il existe toujours au moins une mesure invariante borélienne. C'est le théorème de Krylov-Bogolioubov.*

Remarque 10 *Pour faire des calculs en physique statistique, la situation inverse à la remarque précédente se produit souvent : on connaît $(X, \mathcal{A}, \mu)$, et on construit une dynamique T qui admet μ comme mesure invariante (et même qui est ergodique par rapport à la mesure μ, cf. Définition 7 ci-dessous).*

Par exemple, dans le modèle d'Ising (qui cherche à comprendre le comportement des aimants et des matériaux ferromagnétiques), on attache à chaque noeud d'un réseau une valeur ± 1 qui donne la direction d'un spin en ce noeud. L'espace d'état est donc un espace fini $X = \{-1, +1\}^S$ où, par exemple, $S = \{1, 2, \ldots, n\}^2$. On introduit l'énergie associée à une configuration (x_i) :

$$V(x) = -H \sum_{i \in S} x_i - J \sum_{i \sim j \in S} x_i x_j$$

où $i \sim j$ désigne des noeuds de S qui sont voisins (on décide par exemple que les voisins d'un noeud de S sont les quatre noeuds situés à gauche, à droite, au dessus et en dessous). Le paramètre H modélise l'influence d'un champ magnétique extérieur, et J la force de l'interaction entre les spins ($J > 0$ pour un matériau ferromagnétique, $J < 0$ pour un matériau anti-ferromagnétique). Dans l'ensemble canonique NVT, les configurations sont distribuées suivant la mesure de Boltzmann-Gibbs :

$$\mu(x) = Z^{-1} \exp(-\beta V(x)),$$

où $\beta = 1/(k_B T)$, k_B étant la constante de Boltzmann et T la température, et $Z = \sum_{x\in X} \exp(-\beta V(x))$ est la fonction de partition. Il faut bien comprendre que le cardinal de X est rapidement très grand, et que donc il est illusoire de vouloir simuler directement par méthode de Monte Carlo des échantillons tirés suivant μ. Les méthodes pour calculer par exemple l'énergie moyenne $\mathbb{E}_\mu(V(x))$ ou l'aimantation moyenne $\frac{1}{n^2}\mathbb{E}_\mu(\sum_{i\in S} x_i)$ reposent typiquement sur la construction d'une application T qui admet μ comme mesure invariante. On renvoie par exemple au cours [5] pour plus de détails.

Exercice 32 *Revenons sur notre exemple de l'application tente T. On munit $[0,1]$ de la tribu des boréliens. L'application T est continue donc mesurable. Donner quelques exemples de mesures de probabilité invariantes pour la dynamique, absolument continues par rapport à la mesure de Lebesgue (c'est-à-dire qui s'écrivent sous la forme $\mu = f(x)\,dx$ avec $f \in L^1(0,1)$), ou pas... Montrer qu'il existe une unique mesure de probabilité invariante absolument continue par rapport à la mesure de Lebesgue (on considèrera le cas où $\mu = f(x)\,dx$ avec f une fonction continue pour simplifier).*

Solution *: A partir des orbites périodiques, on peut construire de nombreuses mesures invariantes qui ne sont pas absolument continues par rapport à la mesure de Lebesgue. Par exemple, δ_0 est une mesure invariante, ou bien $(1/2)(\delta_{2/5} + \delta_{4/5})$. La mesure de Lebesgue sur $[0,1]$ est aussi invariante par la dynamique, puisque T est une application de Jacobien 1. Pour l'unicité de la mesure invariante absolument continue par rapport à la mesure de Lebesgue, on introduit la fonction de répartition associée à la mesure invariante. On voit qu'elle vérifie l'équation fonctionnelle $F(x) = F(x/2) + 1 - F(1-x/2)$. On sait de plus que $F(1) = 1$ et que $F(0) = 0$. En dérivant, on a donc $f(x) = \frac{1}{2}\left(f(x/2) + f(1-x/2)\right)$. Supposons que f est une fonction continue. Pour tout $x \in \mathbb{R}\setminus\mathbb{Q}$, on a*

$$\begin{aligned} f(x) &= \frac{1}{2}\left(f(x/2) + f(1-x/2)\right), \\ &= \frac{1}{4}\left(f(x/4) + f(1-x/4) + f(1/2-x/2) + f(1/2-x/4)\right), \\ &= \frac{1}{2^n}\sum_{i=0}^{2^n-1} f(x_i), \end{aligned}$$

où, pour $0 \le i \le 2^n - 1$, $x_i \in (i/2^n, i+1/2^n)$. En passant à la limite en n, on voit que $f(x) = \int_0^1 f(x)\,dx$, et donc f est une constante. Le fait que μ soit une mesure de probabilité impose $f = 1$. On peut également se référer à l'Exercice 35 ci-dessous pour traiter le cas général où $f \in L^1$.

Pour toute fonction "test" $f : X \to \mathbb{R}$ mesurable[2], on note

$$S_n(f)(x) = \frac{1}{n}\sum_{i=0}^{n-1} f \circ T^i(x)$$

la moyenne de f sur une trajectoire de longueur n. $S_n(f)$ est la somme de Birkhoff d'ordre n de f (pour T). Le théorème fondamental de cette section est :

[2]La tribu des ensembles mesurables sur $\mathbb{R}$ est implicitement l'ensemble des boréliens.

Théorème 8 (Théorème ergodique de de Birkhoff) *Soit $(X, \mathcal{A}, \mu, T)$ un système dynamique mesuré. Soit $f \in L^1(X, \mathcal{A}, \mu)$. Alors les sommes de Birkhoff $S_n(f)$ convergent (dans la limite $n \to \infty$) p.p. vers une fonction f^* T-invariante p.p. De plus, la convergence a lieu dans $L^1(X, \mathcal{A}, \mu)$, et donc $f^* \in L^1(X, \mathcal{A}, \mu)$, et $\int f^* d\mu = \int f d\mu$. Si $f \in L^p(X, \mathcal{A}, \mu)$, avec $p \in [1, \infty)$, alors $S_n(f)$ converge dans L^p vers f^*.*

Exercice 33 *En écrivant $S_n(f) \circ T - S_n(f)$, montrer que si $f \in L^p$ et $S_n(f)$ converge dans L^p (où $p \in [1, \infty)$), alors, nécessairement, $f^* \circ T = f^*$ p.p.*

Solution *: $S_n(f) \circ T - S_n(f) = \frac{1}{n}(f \circ T^n - f)$, donc, en norme L^p, tend vers 0 ; passer à la limite $n \to \infty$ dans $S_n(f) \circ T - S_n(f)$.*

4.3 Preuve de la convergence en norme L^2

On va montrer la convergence en norme L^2 dans le cas $f \in L^2$ (cette version du théorème s'appelle le théorème ergodique de Von Neumann). Pour la démonstration de la convergence p.p., on renvoie par exemple à [1, Theorem 6.21] ou [11].

Remarque 11 *La convergence en norme L^2 est en fait valide même dans le cas où μ n'est pas une mesure finie, contrairement à la convergence p.p.*

L'espace $L^2(X, \mathcal{A}, \mu)$ est un espace de Hilbert. On introduit

$$U : \left\{ \begin{array}{ccc} L^2(X, \mathcal{A}, \mu) & \to & L^2(X, \mathcal{A}, \mu) \\ f & \mapsto & f \circ T \end{array} \right. .$$

On a $S_n(f) = \frac{1}{n} \sum_{i=0}^{n-1} U^i f$. L'application U est linéaire et unitaire : $\int (Uf)^2 d\mu = \int f^2 d\mu$ donc $\|U\| = 1$. On va alors montrer que (dans L^2)

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=0}^{n-1} U^i(f) = \pi_I(f) \tag{12}$$

où π_I est la projection orthogonale sur l'espace vectoriel fermé I des fonctions f invariantes par U (c'est-à-dire telles que $f \circ T = f$) :

$$I = \{f \in L^2, Uf = f \, p.p.\}.$$

Ceci terminera donc la preuve du théorème.

Remarque 12 *Le théorème ergodique permet donc d'identifier la limite f^* comme la projection de f sur les fonctions invariantes par U. En termes probabilistes, f^* est l'espérance conditionnelle de f (pour la mesure μ) par rapport à la tribu $\mathcal{I}$ (appelée la tribu des invariants) définie par :*

$$\mathcal{I} = \{A \in \mathcal{A}, \, T^{-1}(A) = A \, p.p.\}.$$

En effet, on vérifie que f est $\mathcal{I}$-mesurable si et seulement si $f \circ T = f$ p.p., c'est-à-dire si et seulement si $f \in I$. L'espérance conditionnelle se définit naturellement comme la projection dans l'espace $L^2(X, \mathcal{A}, \mu)$ sur I. On peut en fait aussi la définir pour $f \in L^1(X, \mathcal{A}, \mu)$, et dans ce cas, la limite f^* du théorème ergodique s'identifie encore à l'espérance conditionnelle de f par rapport à $\mathcal{I}$.

Ceci nécessite quelques rappels sur les espaces de Hilbert. Un espace de Hilbert E est un espace vectoriel muni d'un produit scalaire[3] $\langle ., . \rangle$ (qui définit la norme sur E : $\|f\|^2 = \langle f, f \rangle$) pour lequel E est complet.

Théorème 9 (Projection sur un convexe) *Soit C un convexe fermé d'un Hilbert $(E, \langle ., . \rangle)$. Alors pour tout $x \in E$, il existe un unique $\pi_C(x)$ (projection de x sur C) tel que $\pi_C(x) \in C$ et*

$$\|x - \pi_C(x)\| = \inf\{\|x - y\|,\, y \in C\} = d(x, C).$$

Le point $\pi_C(x)$ est le seul point $y \in C$, tel que $\forall z \in C$,

$$\langle x - y, z - y \rangle \leq 0. \tag{13}$$

L'application π_C est 1-Lipschitzienne.

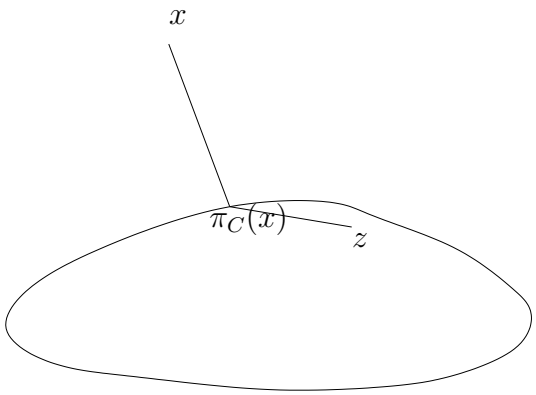

FIG. 7 – Projection sur un convexe.

Preuve : On utilise l'identité du parallélogramme : $\forall x, y \in E$,

$$\|x + y\|^2 + \|x - y\|^2 = 2\left(\|x\|^2 + \|y\|^2\right).$$

On considère $\alpha = \inf\{\|x - y\|,\, y \in C\}$ et y_n une suite minimisante d'éléments de C associée. On a : $\forall n, m$,

$$\|2x - y_n - y_m\|^2 + \|y_n - y_m\|^2 = 2\left(\|x - y_n\|^2 + \|x - y_m\|^2\right)$$

et donc, comme $(y_n + y_m)/2 \in C$,

$$\begin{aligned} \|y_n - y_m\|^2 &= 2\left(\|x - y_n\|^2 + \|x - y_m\|^2 - 2\|x - (y_n + y_m)/2\|^2\right), \\ &\leq 2\left(\|x - y_n\|^2 + \|x - y_m\|^2 - 2\alpha^2\right). \end{aligned} \tag{14}$$

Par conséquent, pour n et m assez grand, le second membre est aussi petit que l'on veut : la suite (y_n) est donc de Cauchy. Comme C est fermé dans E complet, (y_n) converge vers un élément noté $\pi_C(x)$ de C, et on a donc

$$\|x - \pi_C(x)\| = \min\{\|x - y\|,\, y \in C\}.$$

[3]On suppose ici que le produit scalaire est à valeur dans $\mathbb{R}$, mais tout ce qui suit peut se généraliser au cas d'un produit scalaire à valeur dans $\mathbb{C}$ (espace hermitien).

Il existe une unique limite possible car si y_1 et y_2 sont tels que $\|x-y_1\| = \|x-y_2\| = \alpha$, alors, par (14)

$$\|y_1 - y_2\|^2 \leq 2\left(\|x-y_1\|^2 + \|x-y_2\|^2 - 2\alpha^2\right)$$

et donc $y_1 = y_2$.

Prouvons maintenant (13). Par ce qui précède, il est clair que $\pi_C(x)$ est le seul point de C tel que (cf. Figure 7) :

$$\forall z \in C \text{ et } \forall t \in [0,1], \|x - \pi_C(x)\| \leq \left\|x - \left((1-t)\pi_C(x) + tz\right)\right\|. \qquad (15)$$

Or,

$$\begin{aligned}\left\|x - \left((1-t)\pi_C(x) + tz\right)\right\|^2 &= \|(x-\pi_C(x)) - t(z-\pi_C(x))\|^2, \\ &= \|x-\pi_C(x)\|^2 - 2t\langle x-\pi_C(x), z-\pi_C(x)\rangle + t^2\|z-\pi_C(x)\|^2.\end{aligned}$$

Donc (15) est équivalent à

$$\forall z \in C \text{ et } \forall t \in [0,1], 0 \leq -2t\langle x-\pi_C(x), z-\pi_C(x)\rangle + t^2\|z-\pi_C(x)\|^2,$$

et il est facile de vérifier que ceci implique

$$\forall z \in C, \langle x-\pi_C(x), z-\pi_C(x)\rangle \leq 0.$$

Il reste à montrer que l'application π_C est 1-lipschitzienne. On a :

$$\langle x-\pi_C(x), \pi_C(y)-\pi_C(x)\rangle \leq 0$$

$$\langle y-\pi_C(y), \pi_C(x)-\pi_C(y)\rangle \leq 0$$

d'où, en sommant,

$$\langle x-\pi_C(x) - y + \pi_C(y), \pi_C(y)-\pi_C(x)\rangle \leq 0.$$

On en déduit que

$$\|\pi_C(y)-\pi_C(x)\|^2 \leq \langle x-y, \pi_C(x)-\pi_C(y)\rangle$$

et on conclut en appliquant l'inégalité de Cauchy Schwarz $\langle x-y, \pi_C(x)-\pi_C(y)\rangle \leq \|x-y\|\,\|\pi_C(x)-\pi_C(y)\|$. ♢

Théorème 10 (Orthogonal d'un sous-espace vectoriel) *Soit F un sous-espace vectoriel fermé d'un Hilbert $(E, \langle .,. \rangle)$. On introduit le sous-espace vectoriel fermé*

$$F^\perp = \{x \in E, \forall y \in F, \langle x, y\rangle = 0\}.$$

On introduit π_F la projection sur F, définie dans le Théorème 9. La projection π_F est un opérateur linéaire et continu, et tel que $x - \pi_F(x)$ appartient à $F^\perp$: π_F s'appelle aussi la projection orthogonale sur F.

On a $E = F \oplus F^\perp$, c'est-à-dire que tout $x \in E$ s'écrit de manière unique comme la somme d'un élément de F et d'un élément de $F^\perp$:

$$x = \pi_F(x) + (x - \pi_F(x)).$$

On a la relation de Pythagore $\|x\|^2 = \|\pi_F(x)\|^2 + \|x-\pi_F(x)\|^2$.

Enfin, pour tout sous-espace vectoriel G de E, on a

$$(G^\perp)^\perp = \overline{G}$$

et, en particulier, si G est fermé, $(G^\perp)^\perp = G$.

Preuve : Il est clair que $F^\perp$ est bien un sous-espace vectoriel fermé de E (en utilisant la linéarité et la continuité de l'application produit scalaire).

F est un sous-espace fermé et convexe de E, ce qui permet de bien définir $\pi_F(x)$, par le Théorème 9. En utilisant (13), et en remarquant que l'ensemble des $\{z-y, z \in F\}$ est exactement F (car F est un espace vectoriel), on voit que $\pi_F(x)$ est en fait l'unique point $y \in F$ tel que, $\forall z \in F$, $\langle x-y, z\rangle \leq 0$, c'est-à-dire (puisque si $z \in F$, alors $-z \in F$) l'unique point $y \in F$ tel que, $\forall z \in F$,

$$\langle x-y, z\rangle = 0.$$

On en déduit que $x - \pi_F(x) \in F^\perp$, et que $\pi_F(x+\lambda y) = \pi_F(x) + \lambda \pi_F(y)$ puisque $\pi_F(x) + \lambda\pi_F(y) \in F$ et $\forall z \in F$,

$$\langle (x+\lambda y) - (\pi_F(x) + \lambda\pi_F(y)), z\rangle = \langle x - \pi_F(x), z\rangle + \lambda\langle y - \pi_F(y), z\rangle = 0.$$

La relation $x = \pi_F(x) + (x - \pi_F(x))$ montre que $E = F + F^\perp$, et l'unicité de la décomposition vient du fait que $F \cap F^\perp = \{0\}$. La relation de Pythagore découle du fait que $\langle \pi_F(x), x - \pi_F(x)\rangle = 0$ (puisque $x - \pi_F(x) \in F^\perp$).

Si G est un sous-espace vectoriel fermé de E, on a $E = G \oplus G^\perp$, d'après ce qui précède, et donc $E = G^\perp \oplus (G^\perp)^\perp$. Or, $G \subset (G^\perp)^\perp$, donc $G = (G^\perp)^\perp$. Maintenant, si G est un sous-espace vectoriel quelconque de E, on a $G \subset \overline{G}$, donc $\overline{G}^\perp \subset G^\perp$. De plus, comme G est dense dans $\overline{G}$, on a aussi $G^\perp \subset \overline{G}^\perp$ d'où

$$G^\perp = \overline{G}^\perp.$$

On en déduit que $(G^\perp)^\perp = (\overline{G}^\perp)^\perp$. Ceci termine la preuve car $(\overline{G}^\perp)^\perp = \overline{G}$, puisque $\overline{G}$ est fermé. $\diamondsuit$

Exercice 34 (Théorème de Riesz) *1) Soit E un espace vectoriel, et S et T deux formes linéaires sur E. On suppose que $Ker\ S \subset Ker\ T$. En déduire que S et T sont proportionnelles, c'est-à-dire qu'il existe λ tel que $S = \lambda T$.*

Correction : *Si $S = 0$, alors $Ker\ S = E$, donc $Ker\ T = E$, donc $T = 0$ et donc on a bien $S = \lambda T$. Sinon, soit x_0 tel que $S(x_0) \neq 0$, et, quitte à renormaliser, on peut supposer $S(x_0) = 1$. On a alors, pour tout $x \in E$, $(x - S(x)x_0) \in Ker\ S$, donc $(x - S(x)x_0) \in Ker\ T$, et donc*

$$T(x) = T(x_0)S(x).$$

Cette propriété est purement algébrique (elle ne nécessite aucune hypothèse de continuité sur S et T).

2) On suppose maintenant que E est un espace de Hilbert. Montrer que les formes linéaires continues sur E sont exactement les applications de la forme $x \mapsto \langle a, x\rangle$, avec $a \in E$.

Correction : *Il est clair que pour tout $a \in E$, l'application $x \mapsto \langle a, x\rangle$ est une forme linéaire continue. Soit maintenant une forme linéaire continue $l : E \to \mathbb{R}$. Si $l = 0$, on peut choisir $a = 0$. Sinon, soit $a \in (Ker\ l)^\perp$, de norme 1. Le noyau de la forme linéaire $x \mapsto \langle a, x\rangle$ contient le noyau de la forme linéaire l, donc elles sont proportionnelles.*

On termine par la définition de l'adjoint d'une application linéaire :

Proposition 2 *Soit U une application linéaire continue sur un espace de Hilbert E. On définit l'adjoint U^* de U par : $\forall x, y, \in E$,*

$$\langle Ux, y\rangle = \langle x, U^* y\rangle.$$

U^* est une application linéaire continue. On a $\|U^*\| = \|U\|$ et $(U^*)^* = U$.

Preuve : Il est clair que la relation $\forall x, y, \in E$, $\langle Ux, y\rangle = \langle x, U^* y\rangle$ définit bien U^*, puisque $x \mapsto \langle Ux, y\rangle$ est une forme linéaire continue, donc (par le théorème de Riesz), il existe un unique a tel que $U^* y = a$. La linéarité de U^* découle de la linéarité du produit scalaire. Pour montrer que $\|U^*\| = \|U\|$, on remarque que pour tout $x \in E$, $\|U^* x\|^2 = \langle UU^* x, x\rangle \leq \|U\|\|U^* x\|\|x\|$ et donc $\|U^* x\| \leq \|U\|\|x\|$, ce qui montre que $\|U^*\| \leq \|U\|$. Un raisonnement similaire permet de vérifier que $\|U\| \leq \|U^*\|$. Enfin, on vérifie facilement que $(U^*)^* = U$. ♢

Nous avons maintenant les outils pour démontrer (12), que nous énonçons de manière générale sous la forme

Lemme 5 *Soit $U : E \to E$ un opérateur linéaire continu sur un espace de Hilbert E. On suppose que $\|U\| \leq 1$. Alors, $\forall f \in E$,*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} U^i(f) = \pi_I(f)$$

où π_I désigne la projection orthogonale sur le sous-espace vectoriel fermé

$$I = Ker\ (Id - U).$$

Preuve : Il est clair que l'espace vectoriel I est un sous-espace vectoriel fermé de $L^2(X, \mathcal{A}, \mu)$, ce qui permet de bien définir π_I. De plus, il est clair que si $f \in I$, alors $S_n(f) = \frac{1}{n}\sum_{i=0}^{n-1} U^i(f) = f$, et donc $S_n(f) = \pi_I(f)$. Il reste donc à démontrer que si $f \in I^\perp$, alors $\lim_{n\to\infty} S_n(f) = 0$. La propriété $\|U\| \leq 1$ va ici intervenir de manière cruciale.

Le preuve se fait en trois étapes :

Etape 1 : On montre que $I^\perp = \overline{\text{Im } (\text{Id} - U)}$.

D'après le Théorème 10, il suffit de montrer que $I = \text{Im } (\text{Id} - U)^\perp$, c'est-à-dire que $\text{Ker } (\text{Id}-U) = \text{Im } (\text{Id}-U)^\perp$. Or $f \in \text{Im } (\text{Id}-U)^\perp$ si et seulement si, pour tout $g \in E$, $\langle f, g - Ug\rangle = 0$, c'est-à-dire si et seulement si pour tout $g \in E$, $\langle f - U^* f, g\rangle = 0$, c'est-à-dire si et seulement si $f = U^* f$:

$$\text{Im } (\text{Id} - U)^\perp = \text{Ker } (\text{Id} - U^*).$$

Donc, montrer que $\text{Ker } (\text{Id}-U) = \text{Im } (\text{Id}-U)^\perp$ revient à montrer que $\text{Ker } (\text{Id}-U) = \text{Ker } (\text{Id} - U^*)$, c'est-à-dire que $f = Uf$ si et seulement si $f = U^* f$. On écrit alors

$$\begin{aligned} \|Uf - f\|^2 &= \|Uf\|^2 + \|f\|^2 - 2\langle Uf, f\rangle \\ &\leq 2\left(\|f\|^2 - \langle f, U^* f\rangle\right). \end{aligned}$$

Par conséquent, si $U^*f = f$, alors $\|Uf - f\| = 0$ et donc $Uf = f$: Ker $(\mathrm{Id} - U) \subset$ Ker $(\mathrm{Id} - U^*)$. Comme $\|U^*\| = \|U\| \leq 1$ et $(U^*)^* = U$, par le même raisonnement, on a l'inclusion inverse, et ceci termine donc la première étape.

Etape 2 : Si $f \in \mathrm{Im}\ (\mathrm{Id} - U)$, alors $\lim_{n\to\infty} S_n(f) = 0$.
Si $f \in \mathrm{Im}\ (\mathrm{Id} - U)$, alors il existe un $g \in E$ tel que $f = g - Ug$. On a donc $S_n(f) = \frac{1}{n}\sum_{i=0}^{n-1} U^i(g - Ug) = \frac{1}{n}(g - U^n g)$, d'où $\|S_n(f)\| \leq \frac{2}{n}\|g\|$, ce qui termine l'étape 2.

Etape 3 : Si $f \in \overline{\mathrm{Im}\ (\mathrm{Id} - U)}$, alors $\lim_{n\to\infty} S_n(f) = 0$.
Ce résultat découle du fait que les applications $S_n = \frac{1}{n}\sum_{i=0}^{n-1} U^i$ sont uniformément bornées :

$$\|S_n\| \leq 1.$$

Soit $f \in \overline{\mathrm{Im}\ (\mathrm{Id} - U)}$ et $\varepsilon > 0$. On choisit un $f_\varepsilon \in \mathrm{Im}\ (\mathrm{Id} - U)$ tel que $\|f - f_\varepsilon\| \leq \varepsilon/2$. On a alors

$$\begin{aligned}\|S_n(f)\| &= \|S_n(f - f_\varepsilon) + S_n(f_\varepsilon)\|,\\ &\leq \|f - f_\varepsilon\| + \|S_n(f_\varepsilon)\|,\\ &\leq \varepsilon/2 + \|S_n(f_\varepsilon)\|.\end{aligned}$$

Pour n assez grand, le second terme est plus petit que $\varepsilon/2$, et ceci termine la preuve
$\diamondsuit$

4.4 Ergodicité

Nous avons montré que, pour un système dynamiqe mesuré $(X, \mathcal{A}, \mu, T)$, et une fonction test $f \in L^1(X, \mathcal{A}, \mu, T)$, les sommes de Birkhoff $\frac{1}{n}\sum_{i=0}^{n-1} f \circ T^i$ converge p.p. vers une fonction f^* T-invariante. Sous une hypothèse supplémentaire dite d'ergodicité, cette limite est en fait une constante (donc $f^* = \int f d\mu$ p.p. car on sait que $\int f^* d\mu = \int f d\mu$).

Définition 7 *Soit $(X, \mathcal{A}, \mu, T)$ un système dynamique mesuré. On dit que T est ergodique si $\forall A \in \mathcal{A}$, $T^{-1}(A) = A$ p.p. implique $\mu(A) = 0$ ou $\mu(A) = 1$.*

Il est très important de remarquer que la notion d'ergodicité fait appel à la fois à l'application T et la mesure μ. La phrase "T est ergodique" n'a a priori pas de sens (ou bien elle nécessite que la mesure μ soit implicitement définie).

Proposition 3 *Soit $(X, \mathcal{A}, \mu, T)$ un système dynamique mesuré, pour lequel T est une tranformation ergodique. Alors, pour toute fonction $f \in L^1(X, \mathcal{A}, \mu)$, les sommes de Birkhoff convergent p.p. et dans L^1 vers la moyenne de f par rapport à μ :*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} f \circ T^i(x) = \int f d\mu.$$

Si $f \in L^p(X, \mathcal{A}, \mu)$, avec $p \in [1, \infty)$, la convergence a lieu dans L^p.

Preuve : Il suffit de montrer que toute fonction f^* T-invariante et intégrable est nécessairement p.p. constante si T est ergodique.On considère $A = \{x, f^*(x) > \int f^* d\mu\}$. A est clairement mesurable et T invariante, donc $\mu(A) = 0$ ou $\mu(A) = 1$. Si $\mu(A) = 1$, on a $\int f^* d\mu > \int f^* d\mu$, ce qui est impossible, donc $\mu(A) = 0$. En raisonnant de même sur $A = \{x, f^*(x) < \int f^* d\mu\}$, on obtient que $f^* = \int f^* d\mu$ p.p. $\diamondsuit$

On a montré dans la preuve précédente que si T est ergodique pour la mesure μ alors une fonction $f \in L^1(X, \mathcal{A}, \mu)$ T-invariante p.p. est constante p.p.. Cette propriété est en fait une caractérisation de l'ergodicité :

Proposition 4 *Soit $(X, \mathcal{A}, \mu, T)$ un système dynamique mesuré et $p \in [1, \infty)$. Alors T est ergodique pour la mesure μ si et seulement si pour toute fonction $f \in L^p(X, \mathcal{A}, \mu)$ T-invariante p.p., f est égale à une constante p.p.*

Preuve : Pour le sens direct, voir la preuve précédente. Pour la réciproque, il suffit de remarquer que les fonctions indicatrices d'ensemble mesurable sont bien intégrables. $\diamondsuit$

Remarque 13 *En terme de l'opérateur U défini à la section précédente, l'ergodicité revient donc à dire que l'espace propre associé à la valeur propre* 1 *est réduit aux fonctions constantes.*

On a donc l'égalité entre une moyenne sur une trajectoire et une moyenne en espace. Ce théorème est à la base de méthodes numériques pour calculer des moyennes : on cherche alors à construire une dynamique donc on sait qu'elle est ergodique par rapport à une mesure μ que l'on veut échantillonner, et on fait des moyennes trajectorielles pour calculer des moyennes par rapport à μ. Ce théorème est également fondamental en physique statistique pour étudier des systèmes à l'équilibre thermodynamique (cf. [5] et Remarque 10).

Il est clair que le critère d'ergodicité est nécessaire : pour tout sous-espace A de X invariant par T, la suite des points x_n reste dans A si $x_0 \in A$, et ne visite donc pas A^c. Par conséquent, si $\mu(A^c) > 0$ on ne pourra jamais obtenir de moyennes par rapport à μ en faisant une moyenne sur des points de A. Il faut donc que soit $\mu(A) = 0$ (la probabilité de partir d'une condition initiale dans A est nulle), soit $\mu(A) = 1$, et donc $\mu(A^c) = 0$ (la partie de X qui n'est pas visitée par la dynamique "ne compte pas" dans une moyenne par rapport à μ). Le critère d'ergodicité est donc une hypothèse garantissant que l'espace des phases ne peut pas se séparer en deux morceaux disjoints de mesure strictement positive. Remarquer que si cela est possible, on peut toujours considérer la restriction de l'application à un des deux morceaux, et étudier l'ergodicité sur chacune de ces composantes. Ceci peut d'ailleurs se formaliser en un théorème de décomposition en composantes ergodiques. Schématiquement, on peut dire que les différentes composantes ergodiques sont obtenues en considérant, pour chaque condition initiale x, la mesure définie par la moyenne sur la trajectoire partant de x.

Remarque 14 (suite de la Remarque 12) *En terme probabiliste, le critère d'ergodicité fait que toute fonction f $\mathcal{A}$-mesurable est indépendante de la tribu des invariants $\mathcal{I}$, et donc f^* qui est l'espérance conditionnelle de f par rapport à $\mathcal{I}$ est simplement l'espérance de f.*

Remarque 15 (Unique ergodicité) *Quand une application admet une unique mesure invariante, alors elle est ergodique pour cette mesure car les probabilités ergodiques sont les points extrémaux du convexe constitué des mesures de probabilités invariantes. On parle dans ce cas d'unique ergodicité. Un exemple de système uniquement ergodique est donné par l'application* $R_\alpha \left\{ \begin{array}{ccc} \mathbb{T} & \to & \mathbb{T} \\ x & \mapsto & x+\alpha \end{array} \right.$ *, où* $\mathbb{T} = \mathbb{R}/\mathbb{Z}$ *est le tore de dimension 1, et* $\alpha \in \mathbb{R} \setminus \mathbb{Q}$.

Exercice 35 *1. Montrer que si* μ *est une mesure ergodique pour* T*, et si* ν *est une mesure* T*-invariante absolument continue par rapport à* μ*, alors* $\mu = \nu$*. On rappelle que* μ *est absolument continue par rapport à* ν *si et seulement si tout ensemble de mesure nulle pour* μ *est de mesure nulle pour* ν*. Le théorème de Radon Nikodym montre qu'alors il existe une fonction* f *mesurable et* μ*-intégrable telle que* $d\nu = f d\mu$ *(dans notre cas, on a de plus* $\int f d\mu = 1$*).*

2. En déduire que pour un système dynamique mesuré $(X, \mathcal{A}, \mu, T)$*,* μ *est un point extrémal de l'ensemble des mesures de probabilités invariantes par* T *si et seulement si* $(X, \mathcal{A}, \mu, T)$ *est ergodique. On rappelle qu'un point* M *est extrémal dans ensemble convexe* K *inclus dans un sous-espace vectoriel si et seulement si, si* $M = tA + (1-t)B$*, avec* A *et* B *dans* K *et* $t \in (0,1)$*, alors* $A = B = M$*. Ici l'ensemble convexe est l'ensemble des mesures de probabilités invariantes par* T*.*

Solution *:*

1. *On vérifie que* $(X, \mathcal{A}, \nu, T)$ *est ergodique car* $\mu(A) = 0$ *(resp.* $\mu(A) = 1$*) implique* $\nu(A) = 0$ *(resp.* $\nu(A) = 1$*). Par la Proposition 3, on a donc que pour tout* $A \in \mathcal{A}$*, les sommes de Birkhoff* $\frac{1}{n}\sum_{i=0}^{n-1} 1_A \circ T^i$ *convergent* μ*-presque sûrement (donc* ν*-presque sûrement) vers* $\mu(A)$*, et* ν*-presque sûrement vers* $\nu(A)$*. On en déduit que* $\nu(A) = \mu(A)$*, et donc que* $\mu = \nu$*.*
2. *Si* $(X, \mathcal{A}, \mu, T)$ *est ergodique, et que* $\mu = t\nu_1 + (1-t)\nu_2$*, avec* ν_i *des mesures de probabilités invariantes par* T *et* $t \in (0,1)$*, alors* ν_1 *et* ν_2 *sont absolument continues par rapport à* μ*. On en déduit que* $\nu_1 = \nu_2 = \mu$ *par le résultat de la première question. Réciproquement, si* μ *n'est pas ergodique, il existe* $A \in \mathcal{A}$ *tel que* $\mu(A) \in (0,1)$*. On pose alors* $t = \mu(A)$*, et on définit* ν_1 *(resp.* ν_2*) comme la mesure de probabilité* μ *sachant* A *(resp.* A^c*) (ou bien restreinte à* A *(resp.* A^c*)) :* $\forall B \in \mathcal{A}$*,* $\nu_1(B) = \frac{\mu(B \cap A)}{\mu(A)}$ *(resp.* $\nu_2(B) = \frac{\mu(B \cap A^c)}{\mu(A^c)}$*). On a clairement* $\mu = t\nu_1 + (1-t)\nu_2$*, et les mesures* ν_i *sont bien* T*-invariantes, ce qui montre que* μ *n'est pas extrémal.*

Revenons à nos problèmes modèles.

La rotation $R_\alpha \left\{ \begin{array}{ccc} \mathbb{T} & \to & \mathbb{T} \\ x & \mapsto & x+\alpha \end{array} \right.$ est mesurable et μ est invariante si on munit $\mathbb{T}$ de l'ensemble des boréliens, et si μ désigne la mesure de Lebesgue sur $\mathbb{T}$. De plus R_α est ergodique pour la mesure de Lebesgue car si $f \circ R_\alpha = f$ (pour une fonction $f \in L^2(\mathbb{T})$), on a en série de Fourier

$$f = \sum_{n \in \mathbb{Z}} c_n \exp(2i\pi n x)$$

et

$$f \circ R_\alpha = \sum_{n\in\mathbb{Z}} c_n \exp(2i\pi n\alpha)\exp(2i\pi nx)$$

et donc, par unicité de la décomposition en série de Fourier, $c_i = 0$ pour tout $i \neq 0$, soit $f = c_0$. On a en fait dans ce cas unique ergodicité (cf. Remarque 15).

Considérons maintenant l'application tente : $T : \left\{ \begin{array}{ccc} [0,1] & \to & [0,1] \\ x & \mapsto & \left\{ \begin{array}{l} 2x \text{ si } x \leq 1/2 \\ 2-2x \text{ si } x > 1/2 \end{array} \right. \end{array} \right.$.

L'application T est mesurable et μ est invariante si on munit $[0,1]$ de l'ensemble des boréliens, et si μ désigne la mesure de Lebesgue sur $[0,1]$. De plus, T est ergodique pour cette mesure. Pour prouver ce point, on remarque que si D est l'application $D : \left\{ \begin{array}{ccc} [0,1] & \to & [0,1] \\ x & \mapsto & \left\{ \begin{array}{l} 2x \text{ si } x \leq 1/2 \\ 2x-1 \text{ si } x > 1/2 \end{array} \right. \end{array} \right.$, on a $T \circ T = T \circ D$. Donc, si f est une fonction de carré intégrable T invariante, alors f est aussi une fonction D-invariante (car $f = f \circ T \circ T = f \circ T \circ D = f \circ D$). Or, en passant en série de Fourier, on a

$$f = \sum_{n\in\mathbb{Z}} c_n \exp(2i\pi nx)$$

et

$$f \circ D = \sum_{n\in\mathbb{Z}} c_n \exp(4i\pi nx)$$

ce qui montre que les c_n sont tous nuls, sauf c_0, soit g est constante (puisque $\sum |c_n|^2 = \int |f|^2 < \infty$, et donc (c_n) est une suite qui tend vers 0). On en déduit que f est constante, et donc que T est ergodique pour la mesure μ.

On peut donc en déduire que pour presque toute condition initiale, $\frac{1}{n}\sum_{i=0}^{n-1} f \circ T^i(x)$ converge vers $\int_0^1 f\, d\mu$, pour f une fonction intégrable. Ceci est bien valable pour *presque* toute condition initiale, car si $x_0 = 0$, on voit bien que les sommes de Birkhoff ne convergent pas vers la moyenne de f. Ce problème se pose dès que l'on part d'une condition initiale pour laquelle la dynamique est périodique. Ceci dit, l'ensemble de ces conditions initiales est de mesure nulle pour la mesure de Lebesgue : il n'y a donc pas de contradiction avec la Proposition 3.

Remarque 16 *La Proposition 3 dit que pour presque toute condition initiale, la moyenne sur une trajectoire converge vers la moyenne en espace. Une question naturelle est de savoir si il y a des cas où pour toute condition initiale (sans le presque), on a convergence. Un exemple est fourni par le résultat suivant : Si $(X, \mathcal{A}, \mu, T)$ est un système dynamique mesuré avec X est un espace métrique compact et si X admet une unique mesure invariante, alors pour toute fonction f continue, les sommes de Brikhoff $S_n(f)$ convergent uniformément sur X.*

Remarque 17 (Système dynamique discret et continu) *On a étudié des dynamiques dites discrètes, c'est-à-dire pour lequel "le temps" est indexé par $\mathbb{N}$. Pour passer du cas temps continu au cas temps discret, on considère le flot, ou une section de Poincaré, ou bien encore l'exemple de Lorenz...*

Exercice 36 *Montrer comment la loi forte des grands nombres peut être vue comme un corollaire du théorème ergodique de Birkhoff.*

Correction : *Soit $(Y_i)_{i\geq 1}$ une suite de variables aléatoires réelles i.i.d. définie sur un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$. On suppose Y_1 intégrable, de moyenne $\mathbb{E}(Y_1) = m$. On note $X = \mathbb{R}^{\mathbb{N}\setminus\{0\}}$ l'espace des valeurs pour la suite Y, et on munit X de la mesure image $\mu = \mathbb{P} \circ Y^{-1}$. La mesure μ est la mesure produit : $\mu = \nu^{\otimes \mathbb{N}\setminus\{0\}} = \nu \otimes \nu \otimes ...$, où ν est la loi commune des v.a. i.i.d. On considère sur X l'application* shift *T : $T(x_1, x_2, ...) = (x_2, x_3, ...)$. La mesure μ est clairement invariante par T. (L'ensemble X est muni des boréliens $\mathcal{B}$, pour la topologie produit. Les boréliens sont engendrés par les cylindres $\mathcal{B}(x_1, \varepsilon_1) \times ... \times \mathcal{B}(x_n, \varepsilon_n)$.) On remarque que*

$$\mathbb{P}\left\{\omega,\ \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} Y_k = m\right\} = \mu\left\{(x_i)_{i\geq 1},\ \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} x_k = m\right\}. \tag{16}$$

Or, si on considère $f : X \to \mathbb{R}$ l'application projection sur la première coordonnée $f((x_i)_{i\geq 1}) = x_1$, on observe que

$$\frac{1}{n}\sum_{k=1}^{n} x_k = \frac{1}{n}\sum_{k=1}^{n} f(T^k x).$$

*Par ailleurs, on a $\int f d\mu = m$. Il est donc clair que montrer la loi forte des grands nombres (*i.e. *que $\frac{1}{n}\sum_{k=1}^{n} Y_k$ converge presque sûrement vers m, autrement dire que les deux probabilités définies dans* (16) *sont égales à 1) revient à montrer que le système dynamique $(X, \mathcal{B}, \mu, T)$ est ergodique.*

Montrons maintenant l'ergodicité du système dynamique $(X, \mathcal{B}, \mu, T)$. Soit f et g deux fonctions de carré intégrable, et $\pi_n : X \to X$ la projection sur les n premières composantes ($\pi_n(x_1, ..., x_n, x_{n+1}, ...) = (x_1, ..., x_n, 0, ...)$). On a

$$\begin{aligned}
\int f \circ T^n g \, d\mu - \int f \, d\mu \int g \, d\mu &= \int f \circ T^n g \, d\mu - \int f \circ T^n g \circ \pi_{n-1} \, d\mu \\
&\quad + \int f \circ T^n g \circ \pi_{n-1} \, d\mu - \int f \, d\mu \int g \, d\mu \\
&= \int f \circ T^n (g - g \circ \pi_{n-1}) d\mu \\
&\quad + \int f \circ T^n d\mu \int g \circ \pi_{n-1} \, d\mu \int f \, d\mu \int g \, d\mu \\
&= \int f \circ T^n (g - g \circ \pi_{n-1}) d\mu + \int f d\mu \int (g \circ \pi_{n-1} - g) \, d\mu
\end{aligned}$$

où on a utilisé le fait que $f \circ T^n(x)$ ne dépend que des $(x_k)_{k\geq n}$ alors que $g \circ \pi_{n-1}(x)$ ne dépend que de $(x_0, x_1, ..., x_{n-1})$. Par Cauchy Schwarz, on a donc :

$$\left|\int f \circ T^n g \, d\mu - \int f \, d\mu \int g \, d\mu\right| \leq \sqrt{\int f^2 d\mu}\sqrt{\int (g - g \circ \pi_{n-1})^2 d\mu} + \left|\int f d\mu\right| \int |g \circ \pi_{n-1} - g| \, d\mu,$$

et donc, on a

$$\lim_{n\to\infty} \int f \circ T^n g \, d\mu = \int f \, d\mu \int g \, d\mu.$$

Ceci implique l'ergodicité : si A est un ensemble mesurable T-invariant, $f = g = 1_A$ permet d'obtenir $\mu(A) = \mu(A)^2$, soit $\mu(A) = 0$ ou $\mu(A) = 1$.

On a en fait ici montrer une propriété plus forte que l'ergodicité : le système dynamique est fortement mélangeant.

Exercice 37 *On considère l'application logistique $Q(x) = 4x(1-x)$ et sa dynamique associée sur $[0,1]$ muni de l'ensemble des boréliens et de la mesure de Lebesgue. Soit $h(x) = sin^2(\pi x/2)$. Vérifier que $Q \circ h = h \circ T$, où T est l'application tente. En déduire des propriétés sur le système dynamique associé à Q.*

Solution *: On vérifie facilement que $Q \circ h = h \circ T$. Par ailleurs, on vérifie que la mesure image de la mesure de Lebesgue λ sur $[0,1]$ par h est $\mu = \lambda \circ h^{-1}$ avec $\mu = \frac{1}{\pi\sqrt{x(1-x)}}dx$. On en déduit que $([0,1], \mathcal{B}, \mu, Q)$ est ergodique.*

5 Approximation numérique et problèmes raides

On veut calculer une approximation de la solution x du problème de Cauchy

$$\begin{cases} \dot{x} = f(t,x), \\ x(0) \text{ donné.} \end{cases}$$

On suppose que $x \in \mathbb{R}^d$.

On suppose que le problème continu est bien posé pour $t \in \mathbb{R}$, c'est-à-dire qu'il admet une unique solution. On a vu qu'une hypothèse raisonnable pour avoir ce caractère bien posé est que f soit L-lipschitzienne, ce que l'on suppose dans toute la suite :

$$\begin{array}{c} f \text{ est une fonction continue des variables } (t,x) \\ \text{et } L\text{-lipschitzienne par rapport à la variable } x. \end{array} \tag{17}$$

Comme dans la Section 2, on note $\phi(t_*, x_*; t)$ le flot associé à l'équation différentielle ordinaire, c'est-à-dire que $t \mapsto \phi(t_*, x_*; t)$ est la solution du problème de Cauchy, avec comme condition initiale $x(t_*) = x_*$.

Pour des raisons de simplicité, on considère une grille uniforme en temps de pas de temps δt. On pose $t_n = n\delta t$. La généralisation de ce qui suit à des pas de temps variables est facile. On va considérer des schémas numériques qui construisent une approximation $x_n \approx x(t_n)$.

On chercher à répondre à deux types de question :

- A T fixé, que dire de la solution exacte par rapport à la solution approchée dans la limite $\delta t \to 0$? C'est une question de convergence, que l'on va traiter dans la Section 5.1.
- A δt fixé, que dire de la solution exacte par rapport à la solution approchée dans la limite $T \to \infty$? C'est une question liée à la stabilité (absolue) du schéma, et aux systèmes raides, problèmes que nous aborderons dans la Section 5.2.

5.1 Convergence

Trois aspects : les schémas à un pas et leur convergence / les méthodes de type prédicteur-correcteur / les méthodes d'extrapolation.

5.1.1 Schémas à un pas

On veut approcher la solution du problème de Cauchy sur un intervalle de temps $[0, T]$, et on pose $N(\delta t) = T/\delta t$ le dernier pas de temps du schéma. On suppose que δt est tel que $N(\delta t) \in \mathbb{N} \setminus \{0\}$.

On se restreint dans la suite aux méthodes à un pas : x_{n+1} ne dépend que de x_n :

$$x_{n+1} = x_n + \delta t \Psi(t_n, x_n; t_{n+1})$$

Evidemment, la fonction incrément Ψ dépend de la fonction f (même si nous omettons d'indiquer clairement cette dépendance pour ne pas alourdir les notations). La fonction $\Phi(t_*, x_*; t) = x + \delta t \Psi(t_*, x_*; t)$ est la fonction "flot numérique" sur l'intervalle de temps $[t_*, t]$. C'est une approximation du flot ϕ associé à l'équation différentielle ordinaire. Il faut cependant faire attention au fait que en général, $\Phi(t_*, x_*; t_* + 2\delta t)$ n'est pas égal à $\Phi(t_* + \delta t, \Phi(t_*, x_*; t_* + \delta t); t_* + 2\delta t)$ (d'ailleurs, la différence entre les deux donne des informations sur l'erreur du schéma, cf. Section 5.1.5).

Quelques exemples :

– La méthode d'Euler explicite

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

– La méthode d'Euler implicite

$$x_{n+1} = x_n + \delta t f(t_{n+1}, x_{n+1})$$

– La méthode des trapèzes (appelée aussi méthode de Crank - Nicolson)

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

– La méthode de Heun

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

– La méthode du point milieu

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

Toutes ces méthodes font partie d'une classe plus importante de schémas, appelés schémas de Runge et Kutta. On renvoie à [3] pour une présentation de ces méthodes, ainsi qu'au cours [5].

Définition 8 *Un schéma est dit explicite si, pour toute fonction f, le calcul de x_{n+1} à partir de la donnée de x_n est de la forme $x_{n+1} = T(x_n)$ avec T une application dont on connaît l'expression analytique en fonction de f. Au contraire, un schéma est dit implicite si le calcul de x_{n+1} à partir de la donnée de x_n nécessite de résoudre un problème (typiquement non-linéaire si la fonction f est non-linéaire) de la forme $H(x_n, x_{n+1}) = 0$.*

Remarque 18 *On appelle parfois la méthode d'Euler explicite (resp. implicite) la méthode d'Euler progressive ou* forward *(resp. rétrograde ou* backward*). Il faut faire attention à ces dénominations car elles supposent que l'on résout le problème en temps croissant, ce qui n'est pas toujours le cas en pratique (cf. notamment l'équation de Black Scholes en finance).*

Exercice 38 *Dans les exemples précédents, distinguer les schémas explicites des schémas implicites.*

5.1.2 Schémas implicites et méthode prédicteur-correcteur

Pour les schémas implicites, la fonction Ψ n'a donc pas en général d'expression analytique simple. Cependant, on supposera dans la suite qu'elle est bien définie, c'est-à-dire que x_{n+1} est bien défini en fonction de x_n. Ceci ne va pas de soi.

Regardons un peu plus précisément, à titre d'exemple, le schéma d'Euler implicite sur une équation différentielle ordinaire en dimension 1 ($d = 1$) : pour x_n donné, x_{n+1} satisfait

$$x_{n+1} = x_n + \delta t f(t_{n+1}, x_{n+1}). \tag{18}$$

Une première question (théorique) est de savoir si ce problème admet une unique solution. Une seconde question (numérique) est de savoir comment la calculer. On peut répondre à ces deux questions par le résultat suivant :

Proposition 5 *Si δt est suffisamment petit de telle sorte que*

$$\forall x,\ \delta t \left| \frac{\partial f}{\partial x} \right| (t_n, x) < 1 \tag{19}$$

alors il existe un unique x_{n+1} tel que $x_{n+1} = x_n + \delta t f(t_{n+1}, x_{n+1})$, ce qui montre que l'application Φ est bien définie. De plus le schéma de type point fixe

$$x_{n+1}^{(k+1)} = x_n + \delta t f(t_{n+1}, x_{n+1}^{(k)})$$

est tel que $x_{n+1}^{(k)}$ converge (quand $k \to \infty$) vers x_{n+1} quelque soit $x_{n+1}^{(0)}$.

Preuve : cf. les rappels sur le théorème de point fixe de Picard.

La condition (19) découle de la condition $\delta t L < 1$, puisque f est L-lipschitzienne. Cette condition de petitesse sur le pas de temps n'est pas en général contraignante, sauf dans le cas des problèmes raides (cf. Section 5.2).

Exercice 39 *Montrer que dans le cas du schéma d'Euler implicite, si on suppose que $\delta t L < 1$, alors la fonction incrément Ψ est Lipschitzienne de constante de Lipschitz $\frac{L}{1-\delta t L}$.*

D'un point de vue numérique, on peut également envisager d'utiliser une méthode de Newton pour trouver x_{n+1} solution de (18) :

$$x_{n+1}^{(k+1)} = x_{n+1}^{(k)} - \frac{x_{n+1}^{(k)} - x_n - \delta t f(t_{n+1}, x_{n+1}^{(k)})}{1 - \delta t \frac{\partial f}{\partial x}(t_{n+1}, x_{n+1}^{(k)})}.$$

On rappelle que la méthode de Newton converge plus vite que la méthode de point fixe. Par contre, on observe en pratique qu'elle nécessite souvent de partir d'une condition initiale très proche de la solution pour converger.

Dans les deux cas (méthode de point fixe, ou méthode de Newton), il est très important en pratique d'avoir une très bonne approximation $x_{n+1}^{(0)}$ de la solution x_{n+1} (*initial guess*) pour amorcer l'algorithme. Dans notre contexte, une idée naturelle est d'estimer x_{n+1} par un pas d'une méthode explicite. Une telle stratégie s'appelle une stratégie prédicteur-correcteur (on prédit la valeur de x_{n+1} puis on la corrige).

Exercice 40 *Réinterpréter la méthode de Heun comme une méthode de prédicteur-correcteur.*

Solution *: prédicteur = Euler explicite / correcteur = Crank Nicolson.*

Au vu de ce qui précède, on peut dire que la différence entre méthode implicite et méthode explicite est qu'une méthode explicite ne nécessite pour calculer l'incrément Ψ que l'évaluation de f un nombre fini de fois, alors qu'une méthode implicite nécessite *a priori* un nombre infini d'évaluations de f. Bien sûr, en pratique, on n'utilise qu'un nombre fini d'itérations de ces algorithmes (jusqu'à ce qu'un critère de convergence numérique soit rempli), mais on retiendra qu'un pas d'une méthode implicite nécessite typiquement beaucoup plus d'évaluations de f qu'un pas d'une méthode explicite.

Il est important de préciser qu'en pratique, on considère que le temps CPU d'une méthode numérique pour résoudre une équation différentielle ordinaire est proportionnel au nombre d'évaluations de la fonction f : toutes les autres opérations sont négligeables. On évalue donc le coût calcul d'un pas d'une méthode par le nombre d'évaluations de la fonction f. Une méthode implicite est donc typiquement beaucoup plus coûteuse qu'une méthode explicite. En anticipant sur la suite, précisons que l'intérêt des méthodes implicites est lié à des problèmes de stabilité : les méthodes explicites sont inutilisables en pratique pour les problèmes raides (cf. Section 5.2).

5.1.3 Convergence

Une question naturelle à ce stade est : est-ce que x_n est une bonne approximation de $x(t_n)$? Pour répondre à cette question, il faut analyser deux phénomènes :

- Quelle est l'erreur commise à chaque pas de temps. Cette erreur s'appelle l'erreur de troncature, ou erreur de consistance.
- Comment ces erreurs commises à chaque pas de temps interagissent au cours des itérations ? Autrement dit, ne peut-on pas avoir des phénomènes d'amplification des erreurs ? Cette analyse est reliée à une notion de stabilité.

Prenons l'exemple du schéma d'Euler explicite. Sur un pas de temps $[t_n, t_{n+1}]$, on a, pour la solution approchée

$$y_{n+1} = y_n + \delta t f(t_n, y_n),$$

et pour la solution exacte

$$y(t_{n+1}) = y(t_n) + \int_{t_n}^{t_{n+1}} f(s, y(s))\, ds.$$

Si on note $e_n = y(t_n) - y_n$ l'erreur à l'instant t_n, on a donc

$$\begin{aligned} e_{n+1} &= e_n + \int_{t_n}^{t_{n+1}} f(s, y(s))\, ds - \delta t f(t_n, y_n) \\ &= e_n + \int_{t_n}^{t_{n+1}} f(s, y(s)) - f(t_n, y(t_n))\, ds + \delta t (f(t_n, y(t_n)) - f(t_n, y_n)). \end{aligned}$$

On note $\varepsilon_n = \int_{t_n}^{t_{n+1}} f(s, y(s)) - f(t_n, y(t_n))\, ds$ l'erreur commise par le schéma sur le pas de temps $[t_n, t_{n+1}]$, en partant du point $y(t_n)$ à l'instant t_n : c'est l'erreur de consistance. En utilisant le fait que f est Lipschitz, on a donc

$$|e_{n+1}| \leq |e_n|(1 + L\delta t) + |\varepsilon_n|,$$

et on en déduit par un Lemme de Gronwall discret

$$\begin{aligned} |e_n| &\leq |e_0|(1 + L\delta t)^n + \sum_{k=0}^{n-1} (1 + L\delta t)^{n-1-k} |\varepsilon_k|, \\ &\leq |e_0| \exp(L t_n) + \sum_{k=0}^{n-1} \exp(L(t_n - t_{k+1}) |\varepsilon_k|. \end{aligned}$$

Ceci permet donc d'analyser comment les erreurs commises à chaque pas de temps s'accumulent. Pour terminer la preuve de convergence, il reste à estimer l'erreur de consistance ε_k. Si on suppose simplement f continue et Lipschitz, on écrit par exemple

$$\begin{aligned} |\varepsilon_n| &= \left| \int_{t_n}^{t_{n+1}} f(s, y(s)) - f(t_n, y(t_n))\, ds \right|, \\ &\leq \int_{t_n}^{t_{n+1}} |\dot{y}(s) - \dot{y}(t_n)|\, ds, \\ &\leq \delta t \sup_{|t-s| \leq \delta t} |\dot{y}(t) - \dot{y}(s)|, \end{aligned}$$

et on a donc, $\forall n \in \{0, \ldots, N(\delta t)\}$,

$$\begin{aligned} |e_n| &\leq |e_0| \exp(L t_n) + \sup_{|t-s| \leq \delta t} |\dot{y}(t) - \dot{y}(s)|\, \delta t \sum_{k=0}^{n-1} \exp(L(t_n - t_{k+1}), \\ &\leq |e_0| \exp(L t_n) + \sup_{|t-s| \leq \delta t} |\dot{y}(t) - \dot{y}(s)|\, \delta t \frac{\exp(L t_n) - 1}{\exp(L\delta t) - 1}, \\ &\leq |e_0| \exp(L t_n) + \sup_{|t-s| \leq \delta t} |\dot{y}(t) - \dot{y}(s)|\, \frac{\exp(L t_n) - 1}{L}. \end{aligned}$$

Ceci montre la convergence du schéma numérique, car le module de continuité $\sup_{|t-s| \leq \delta t} |\dot{y}(t) - \dot{y}(s)|$ de la solution $\dot{y}$ tend vers 0 quand δt tend vers 0.

Si on suppose un peu plus de régularité sur la fonction f et donc sur la solution y, on peut estimer la vitesse de convergence. Supposons par exemple $f \in \mathcal{C}^1$, et donc

$y \in \mathcal{C}^2$. On écrit dans ce cas

$$\begin{aligned}
|\varepsilon_n| &= \left| \int_{t_n}^{t_{n+1}} f(s, y(s)) - f(t_n, y(t_n))\, ds \right|, \\
&\leq \int_{t_n}^{t_{n+1}} |\dot{y}(s) - \dot{y}(t_n)|\, ds, \\
&\leq \int_{t_n}^{t_{n+1}} \int_{t_n}^{s} |\ddot{y}|\,(r)\, dr\, ds, \\
&\leq \delta t \int_{t_n}^{t_{n+1}} |\ddot{y}|\,(r)\, dr.
\end{aligned}$$

On obtient alors l'estimée : $\forall n \in \{0, \ldots, N(\delta t)\}$,

$$|e_n| \leq |e_0| \exp(Lt_n) + \delta t \int_0^{t_n} \exp(L(t_n - s))\, |\ddot{y}|\,(s)\, ds.$$

On obtient donc une convergence en $O(\delta t)$: on parle de convergence à l'ordre 1.

Nous avons introduits sur cet exemple simple les ingrédients essentiels de la preuve de convergence d'un schéma d'approximation d'équation différentielle ordinaire. Généralisons celà pour des schémas plus compliqués. On définit tout d'abord l'erreur de troncature (appelée aussi erreur de consistance) :

Définition 9 *On appelle erreur de troncature (sur le pas de temps (t_n, t_{n+1}), et pour une condition initiale x) la quantité $\varepsilon(t_n, x; t_{n+1}) = y(t_n + \delta t) - y_{n+1}$ où y est solution du problème de Cauchy :*

$$\begin{cases} \dot{y} = f(t, y), \\ y(t_n) = x, \end{cases}$$

et y_{n+1} est obtenu par le schéma numérique

$$y_{n+1} = x + \delta t \Psi(t_n, x; t_{n+1}).$$

En terme de flot, on a donc $\varepsilon(t_*, x_*; t) = \phi(t_*, x_*; t) - \Phi(t_*, x_*; t)$.

On dit que le schéma numérique est consistant[4] *si, quelque soit la fonction f continue Lipschitz, quelque soit (t_*, x_*), $\varepsilon(t_*, x_*; t+\delta t) = o(\delta t)$ dans la limite $\delta t \to 0$. Autrement dit, $\lim_{\delta t \to 0} \dfrac{\varepsilon(t_*, x_*; t_* + \delta t)}{\delta t} = 0$.*

Pour un entier $p \geq 1$, on dit que le schéma numérique est consistant d'ordre p si, quelque soit la fonction f de classe $\mathcal{C}^p$, quelque soit (t_*, x_*),

$$\varepsilon(t_*, x_*; t_* + \delta t) = O(\delta t^{p+1})$$

dans la limite $\delta t \to 0$. Autrement dit, $\exists \delta t^*, \exists K > 0, \forall \delta t \leq \delta t^*, \left| \dfrac{\varepsilon(t_*, x_*; t_* + \delta t)}{\delta t^{p+1}} \right| \leq K$.

[4] sous-entendu avec l'équation différentielle ordinaire $\dot{x} = f(t, x)$

Dans le cas où f ne dépend pas du temps, l'erreur de troncature $\varepsilon(t_*, x_*; t)$ est simplement une fonction de x_* et $t - t_*$: $\varepsilon(t_*, x_*; t) = \varepsilon(0, x_*; t - t_*)$. En une phrase, on peut retenir le slogan : **l'erreur de troncature est l'erreur résiduelle sur un pas de temps quand on applique le schéma numérique à la solution exacte**.

L'erreur de consistance est typiquement obtenue par des développements de Taylor sur l'erreur, en supposant autant de régularité que nécessaire sur f (ou de manière équivalente, sur la solution exacte).

Prenons l'exemple du schéma d'Euler explicite. Si y est la solution exacte, telle que $y(t) = x$, et que $y_1 = x + \delta t f(t, x)$, on a : $\exists \theta \in [0, 1]$,

$$\begin{aligned} y(t + \delta t) &= y(t) + \delta t \dot{y}(t) + \ddot{y}(t + \theta \delta t) \frac{\delta t^2}{2}, \\ &= x + \delta t f(t, x) + \ddot{y}(t + \theta \delta t) \frac{\delta t^2}{2}. \end{aligned}$$

On en déduit que

$$\begin{aligned} |\varepsilon(t, x; t + \delta t)| &= |y(t + \delta t) - y_1|, \\ &\leq \sup_{\theta \in [0,1]} |\ddot{y}(t + \theta \delta t)| \frac{\delta t^2}{2}. \end{aligned}$$

Or, on a $\ddot{y} = \partial_t f(t, y) + \partial_x f(t, y) f(t, y)$, et donc, en supposant que f est $\mathcal{C}^1$, on a bien $\sup_{\theta \in [0,1]} |\ddot{y}(t + \theta \delta t)| < \infty$. Ceci montre que le schéma est consistant d'ordre 1.

Exercice 41 *Etudier la consistance des schémas introduits au début de cette section.*

Solution : *Pour la méthode des trapèzes, par exemple, on a, pour le schéma numérique sur le pas de temps (t_n, t_{n+1}), et en partant du point x :*

$$\begin{aligned} y_{n+1} &= x + \frac{\delta t}{2} \left(f(t_n, x) + f(t_{n+1}, y_{n+1}) \right), \\ &= x + \frac{\delta t}{2} \left(f(t_n, x) + f\left(t_{n+1}, x + \frac{\delta t}{2} \left(f(t_n, x) + f(t_{n+1}, y_{n+1}) \right) \right) \right), \\ &= x + \frac{\delta t}{2} \left(f(t_n, x) + f(t_n, x) + \delta t \partial_t f(t_n, x) + \delta t f \partial_x f(t_n, x) + O(\delta t^2) \right), \\ &= x + \delta t f(t_n, x) + \frac{\delta t^2}{2} \partial_t f(t_n, x) + \frac{\delta t^2}{2} f \partial_x f(t_n, x) + O(\delta t^3). \end{aligned}$$

Pour la solution exacte valant x à l'instant t_n, on a de même

$$\begin{aligned} y(t_{n+1}) &= y(t_n) + \dot{y}(t_n)\, \delta t + \ddot{y}(t_n)\, \frac{\delta t^2}{2} + O(\delta t^3), \\ &= x + f(t_n, x)\, \delta t + \left(\partial_t f(t_n, x) + f \partial_x f(t_n, x) \right) \frac{\delta t^2}{2} + O(\delta t^3). \end{aligned}$$

On voit donc que $\varepsilon(t_n, x; t_{n+1}) = O(\delta t^3)$, et donc que le schéma est d'ordre 2. La constante dans $O(\delta t^3)$ dépend de dérivées de f jusqu'à l'ordre 2.

Remarque 19 *On comprend sur les exemples de l'exercice précédent que vérifier la consistance pour des schémas d'ordre plus élevé devient compliqué, car cela nécessite de calculer des dérivées d'ordre élevé de la solution en fonction de f. On peut alors recourir à des notations sous forme d'arbres pour faciliter les calculs (cf. Chapitre II.2 dans [3]).*

Pour prouver la convergence, il reste à analyser comment les erreurs s'accumulent au cours des itérations. Il y a deux approches possibles :

– soit utiliser le fait que le flot $\phi(t,\cdot;t+\delta t)$ est $C^{\delta t}$-lipschitzien,

– soit utiliser le fait que le fot $\Phi(t,\cdot;t+\delta t)$ est $(1+C\delta t)$-lipschitzien

(pour des constantes C positives). Ces deux propriétés correspondent à des propriétés de stabilité du problème.

Commençons par la première approche. On rappelle que flot $\phi(t,\cdot;t+\delta t)$ est $\exp(L\delta t)$-lipschitzien, puisque f est globalement L-lipschitzienne (cf. (17)), ce qui traduit une stabilité du problème continu (cf. Lemme 4).

Proposition 6 *On suppose que f vérifie (17), et que le schéma est consistant d'ordre p au sens suivant : soit $B=\{(t,x(t)), t\in[0,T]\}$ la trajectoire de la solution exacte, et $\mathcal{V}(B,\varepsilon)$ un ε-voisinage de B. On suppose qu'il existe $\varepsilon>0$ et $K>0$ tel que $\forall (t,x)\in\mathcal{V}(B,\varepsilon)$,*

$$|\varepsilon(t,x;t+\delta t)|\leq K\delta t^{p+1}.$$

On suppose par ailleurs que $x(0)=x_0$. Alors, le schéma est convergent : $\exists C>0$, $\exists\delta t^*>0$, $\forall\delta t<\delta t^*$,

$$\sup_{0\leq n\leq N(\delta t)}|x(t_n)-x_n|\leq C\delta t^p. \tag{20}$$

On peut prendre $C=\frac{K}{L}(\exp(LT)-1)$, et $\delta t^*=\left(\frac{\varepsilon L}{K(\exp(LT)-1)}\right)^{1/p}$.

Remarque 20 *Pour un schéma d'ordre p, la constante dans l'estimation de l'erreur de consistance en $O(\delta t^{p+1})$ dépend de dérivées de f jusqu'à l'ordre p. En particulier, si on suppose que ces dérivées sont bornées par une constante K, on peut prendre $\varepsilon=\infty$ et $\delta t^*=\infty$.*

Preuve : On écrit

$$\begin{aligned}|x(t_{n+1})-x_{n+1}|&\leq|x(t_{n+1})-\phi(t_n,x_n;t_{n+1})|+|\phi(t_n,x_n;t_{n+1})-x_{n+1}|,\\&\leq|\phi(t_n,x(t_n);t_{n+1})-\phi(t_n,x_n;t_{n+1})|+|\phi(t_n,x_n;t_{n+1})-\Phi(t_n,x_n;t_{n+1})|,\\&\leq\exp(L\delta t)|x(t_n)-x_n|+|\varepsilon(t_n,x_n;t_{n+1})|.\end{aligned}$$

On a utilisé le Lemme 4 pour estimer le premier terme au second membre.

On considère maintenant l'hypothèse de récurrence :

$$\mathrm{HR}(n)\qquad |x(t_n)-x_n|\leq\frac{K}{L}\big(\exp(Ln\delta t)-1\big)\delta t^p.$$

Il est clair que HR(0) est vrai. Supposons maintenant HR(n). Comme $|x(t_n)-x_n|\leq\frac{K}{L}(\exp(Ln\delta t)-1)\delta t^p\leq\frac{K}{L}(\exp(LT)-1)\delta t^p$, si $\delta t<\delta t^*=\left(\frac{\varepsilon L}{K(\exp(LT)-1)}\right)^{1/p}$, on

a $|x(t_n) - x_n| < \varepsilon$ et donc $(t_n, x_n) \in \mathcal{V}(B, \varepsilon)$, de sorte que $\varepsilon(t_n, x_n; t_{n+1}) \leq K\delta t^{p+1}$. On a donc

$$\begin{aligned}
|x(t_{n+1}) - x_{n+1}| &\leq \exp(L\delta t)|x(t_n) - x_n| + |\varepsilon(t_n, x_n; t_{n+1})| \\
&\leq \exp(L\delta t)\left(\frac{K}{L}(\exp(Ln\delta t) - 1)\delta t^p\right) + K\delta t^{p+1} \\
&\leq \frac{K}{L}\exp(L(n+1)\delta t)\delta t^p - \frac{K}{L}\delta t^p\left(\exp(L\delta t) - L\delta t\right) \\
&\leq \frac{K}{L}\big(\exp(L(n+1)\delta t) - 1\big)\delta t^p,
\end{aligned}$$

puisque $\exp(x) \geq 1 + x$. Ceci démontre HR$(n+1)$. ♢

On comprend maintenant pourquoi dans la définition de la consistance à l'ordre p, on utilise un δt^{p+1} : la somme des erreurs qui s'accumulent tue un des δt, et la convergence est donc bien en δt^p.

Exercice 42 *Si le schéma est seulement consistant (sans ordre de consistence), montrer que le schéma est convergent (sans ordre de convergence).*

Dans la démonstration précédente, on a estimé l'erreur $|x(t_{n+1}) - x_{n+1}|$ au temps t_{n+1} en fonction de l'erreur au temps t_n en introduisant $\phi(t_n, x_n; t_{n+1})$. On a ensuite dû utiliser une propriété de stabilité du flot ϕ.

On peut aussi reprendre cette démonstration en introduisant $\Phi(t_n, x(t_n); t_{n+1})$, et en utilisant cette fois une propriété de stabilité du flot numérique Φ (on comparera avec le Lemme 4). C'est la deuxième approche mentionnée ci-dessus. On renvoie à la Figure 8 pour une illustration graphique des deux méthodes d'analyse d'erreur.

Proposition 7 *On suppose que la fonction incrément $\Psi(t, x; t + \delta t)$ est Lipschitzienne par rapport à la variable x, avec une constante de Lipschitz $\overline{L}$ indépendante de t et δt. Alors, le schéma numérique est stable au sens suivant : si (x_n) vérifie*

$$x_{n+1} = x_n + \delta t\Psi(t_n, x_n; t_{n+1})$$

et si (y_n) vérifie

$$y_{n+1} = y_n + \delta t\Psi(t_n, y_n; t_{n+1}) + \varepsilon_n,$$

alors on a

$$\max_{0 \leq n \leq N(\delta t)} |x_n - y_n| \leq \exp(\overline{L}T)\left(|x_0 - y_0| + \sum_{k=0}^{N(\delta t)-1} |\varepsilon_k|\right).$$

Preuve : On note $e_n = x_n - y_n$. On a

$$e_{n+1} = e_n + \delta t(\Psi(t_n, x_n; t_{n+1}) - \Psi(t_n, y_n; t_{n+1})) - \varepsilon_n.$$

On en déduit

$$|e_{n+1}| \leq (1 + \overline{L}\delta t)|e_n| + |\varepsilon_n|,$$

et donc, par un Lemme de Gronwall discret

$$|e_n| \leq (1+\overline{L}\delta t)^n |e_0| + \sum_{k=0}^{n-1} (1+\overline{L}\delta t)^{n-1-k} |\varepsilon_k|,$$

d'où le résultat en utilisant le fait que $(1+\overline{L}\delta t) \leq \exp(\overline{L}\delta t)$. $\diamondsuit$

En utilisant cette stabilité, on démontre le résultat de convergence suivant :

Proposition 8 *On suppose que f vérifie (17), et que le schéma est consistant à l'ordre p au sens suivant : pour tout $t \in [0,T]$,*

$$|\varepsilon(t, x(t); t+\delta t)| \leq K \delta t^{p+1}.$$

On suppose que la fonction incrément $\Psi(t,x;t+\delta t)$ est Lipschitzienne par rapport à la variable x, avec une constante de Lipschitz $\overline{L}$ indépendante de t et δt. On suppose par ailleurs que $x(0) = x_0$. Alors, le schéma est convergent : $\exists C > 0$, $\forall \delta t > 0$,

$$\sup_{0 \leq n \leq N(\delta t)} |x(t_n) - x_n| \leq C \delta t^p. \tag{21}$$

On peut prendre $C = \frac{K}{\overline{L}}(\exp(\overline{L}T) - 1)$.

Remarque 21 *En pratique, il est possible que la propriété de consistance et le caractère Lipschitzien de Ψ ne soient vérifiés que pour des pas de temps δt plus petits qu'un δt^* (penser par exemple aux schémas implicites). Dans tous ces théorèmes de convergence, c'est seulement l'asymptotique $\delta t \to 0$ qui nous intéresse.*

Preuve : La preuve se calque sur celle de la Proposition 6, mais en introduisant cette fois le point $\Phi(t_n, x(t_n); t_{n+1})$, et en utilisant le fait que la fonction $\Phi(t,x;t+\delta t)$ est $(1+\overline{L}\delta t)$-Lipschitzienne par rapport à la variable x. On écrit

$$\begin{aligned}
|x(t_{n+1}) - x_{n+1}| &\leq |x(t_{n+1}) - \Phi(t_n, x(t_n); t_{n+1})| + |\Phi(t_n, x(t_n); t_{n+1}) - x_{n+1}|, \\
&\leq |\phi(t_n, x(t_n); t_{n+1}) - \Phi(t_n, x(t_n); t_{n+1})| + |\Phi(t_n, x(t_n); t_{n+1}) - \Phi(t_n, x_n; t_{n+1})|, \\
&\leq |\varepsilon(t_n, x(t_n); t_{n+1})| + (1+\overline{L}\delta t)|x(t_n) - x_n|, \\
&\leq K\delta t^{p+1} + (1+\overline{L}\delta t)|x(t_n) - x_n|.
\end{aligned}$$

On en déduit :

$$\begin{aligned}
|x(t_n) - x_n| &\leq (1+\overline{L}\delta t)^n |x(0) - x_0| + K\delta t^{p+1} \sum_{k=0}^{n-1} (1+\overline{L}\delta t)^{n-1-k}, \\
&\leq K\delta t^{p+1} \frac{(1+\overline{L}\delta t)^n - 1}{\overline{L}\delta t},
\end{aligned}$$

d'où le résultat en utilisant le fait que $(1+\overline{L}\delta t) \leq \exp(\overline{L}\delta t)$. $\diamondsuit$

Remarque 22 (Erreurs d'arrondis) *La notion de stabilité introduite dans la Proposition 7 est fondamentale : elle montre en fait que le schéma numérique définit un problème* bien posé *au sens ou une petite perturbation sur les données implique une petite perturbation du résultat. Ceci est par exemple important en pratique car le calcul en virgule flottante sur ordinateur introduit inévitablement des erreurs d'arrondis. La Proposition 7 permet de contrôler comment ces erreurs se propagent.*

Exercice 43 *Reprendre la preuve de convergence en introduisant des erreurs d'arrondis à chaque pas de temps. Que devient l'énoncé de la convergence du schéma dans ce cas ?*

Pour les schémas numériques à un pas consistants, le fait que Ψ soit Lipschitzien découle typiquement de l'hypothèse (17) sur f. Dans le cas des schémas implicites, on a vu que la fonction Ψ est définie seulement si $\delta t \leq \delta t^*$, auquel cas la fonction Ψ est lipschitzienne et le schéma est stable seulement pour les pas de temps $\delta t \leq \delta t^*$ (cf. par exemple l'Exercice 39). On peut donc dire essentiellement que pour les schémas à un pas, le fait que f soit Lipschitzienne implique que le flot discret (méthode numérique) et le flot continu sont stables, et donc que les deux approches pour la preuve de convergence (Propositions 6 et 8) sont essentiellement équivalentes.

Par contre, cette approche au niveau discret est nécessaire pour analyser d'autres types de schémas, comme les schémas multi-pas, pour lesquels la consistance de la méthode et la stabilité du flot au niveau continu ne suffisent plus à assurer la convergence de la méthode. Pour de telles méthodes, des solutions parasites peuvent provoquer des instabilités (cf. Exercice 44). De manière plus générale, l'analyse de la stabilité (en un sens à préciser) du schéma discret s'avère crucial pour certaines classes de problèmes (les problèmes raides) pour lesquels les estimations obtenus ci-dessus sont inutiles, du fait du comportement du second membre dans (20) ou (21) quand T augmente (cf. Section 5.2).

Des résultats précédents, on peut retenir le slogan : **stabilité et consistance impliquent convergence**.

En regardant dans le détail les preuves précédentes, on comprend que la constante en facteur de δt^p (disons pour une méthode d'ordre p) dans les estimations d'erreur (20) et (21) dépend en fait de la régularité de la solution, et donc de la régularité de f. Pour un schéma d'ordre p, la constante dépend typiquement de normes L^∞ de dérivées d'ordre $p+1$ de la solution exacte, ou de manière équivalente, d'ordre p de f. Si la solution n'est pas suffisamment régulière, on ne peut pas atteindre une convergence d'ordre p, même si le schéma est consistant d'ordre p. De même, plus les normes des dérivées de f, jusqu'à l'ordre p, sont grandes, plus l'erreur de consistance sera grande. On en déduit qu'il est inutile d'utiliser des schémas d'ordre élevé si on sait que la solution est irrégulière. Pour obtenir une meilleure approximation dans ce cas, il vaut mieux diminuer le pas de temps δt. On peut d'ailleurs se servir d'une estimation de l'erreur de consistance pour dériver une méthode d'adaptation du pas de temps : on parle d'estimation d'erreur *a posteriori* pour désigner une estimation de l'erreur calculée à partir du résultat approché. On peut utiliser par exemple les normes des dérivées de f, (cf. p. 169 [3]), mais on peut faire beaucoup mieux pour estimer l'erreur *a posteriori* (cf. Section 5.1.5).

Remarque 23 *L'hypotèse f Lipschitz joue un rôle primordial dans cette section. On peut se convaincre qu'une hypothèse suffisante pour obtenir les théorèmes de convergences est que f soit Lipschitz simplement au voisinage de la solution exacte.*

5.1.4 Autres schémas

Nous nous sommes concentrés sur les schémas à un pas, du type Runge et Kutta. La forme la plus générale d'un schéma de Runge et Kutta est la suivante : pour t_n

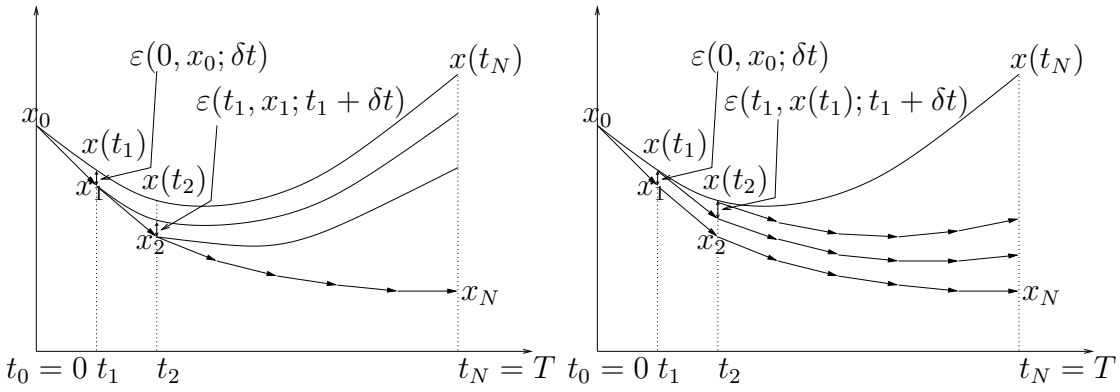

FIG. 8 – Les deux méthodes d'analyse de l'erreur. Une ligne continue indique une solution obtenue par le flot exact. Une ligne fléchée indique une solution obtenue en appliquant le schéma numérique. Pour la Proposition 6 (schéma de gauche), on introduit les solutions exactes partant des points (t_i, x_i) et on regarde comment les erreurs de consistance $\varepsilon(t_i, x_i; t_{i+1})$ sont transportées par le flot exact. Pour la Proposition 8 (schéma de droite), on introduit les solutions approchées partant des points $(t_i, x(t_i))$ et on regarde comment les erreurs de consistance $\varepsilon(t_i, x(t_i); t_{i+1})$ sont transportées par le schéma numérique.

et x_n donnés, x_{n+1} est défini par

$$\begin{cases} k_i = f(t_n + c_i\delta t, x_n + \delta t \sum_{j=1}^{s} a_{i,j}k_j), \text{ pour } i \in \{1, \dots, s\}, \\ x_{n+1} = x_n + \delta t \sum_{i=1}^{s} b_i k_i, \end{cases} \tag{22}$$

s est le nombre d'étapes de la méthode, $(b_i)_{1\le i\le s}$ et $(a_{i,j})_{1\le i,j\le s}$ sont des nombres réels et, pour $1 \le i \le s$, $c_i = \sum_{j=1}^{s} a_{i,j}$. Le schéma est explicite si et seulement si $a_{i,j} = 0$ pour $i \le j$. Dans ce type de schémas, on cherche à monter en ordre typiquement en itérant la fonction f (penser par exemple à la méthode de Heun).

On peut également chercher à monter en ordre tout en n'utilisant que des fonctions linéaires des x_k et des $f(t_k, x_k)$. La contre-partie est qu'il faut faire appel, pour calculer x_{n+1}, non pas à x_n mais à l'ensemble des x_k, pour $n - p \le k \le n$: on parle de schéma à $p + 1$ pas. De telles méthodes s'écrivent typiquement sous la forme

$$x_{n+1} = \sum_{j=0}^{p} a_j x_{n-j} + \delta t \sum_{j=0}^{p} b_j f(t_{n-j}, x_{n-j}) + \delta t b_{-1} f(t_{n+1}, x_{n+1}). \tag{23}$$

La méthode est explicite si et seulement si $b_{-1} = 0$. On voit que pour démarrer la méthode, on a besoin non seulement de $x(0)$, mais aussi de $x(\delta t), \dots, x((p-1)\delta t)$. En pratique, on utilise d'autres schémas numériques (par exemple à un pas) pour estimer ces conditions initiales.

On généralise la notion de consistance pour des schémas multi-pas de la manière suivante : l'erreur de consistance est la différence entre la solution exacte et la solution approchée obtenue par le schéma numérique, en prenant comme conditions initiales pour le schéma numérique les valeurs exactes. A titre d'exemple, on peut considérer le schéma

$$x_{n+1} = x_{n-1} + 2\delta t f(t_n, x_n)$$

dont on vérifie qu'il est d'ordre 2. En effet, sur l'intervalle de temps (t_{n-1}, t_{n+1}), pour la solution exacte, on a :

$$x(t_{n+1}) = x(t_{n-1}) + \dot{x}(t_{n-1})2\delta t + \ddot{x}(t_{n-1})2\delta t^2 + x^{(3)}(t_{n-1})\frac{4}{3}\delta t^3 + O(\delta t^4).$$

Pour la solution approchée partant des valeurs exactes $x(t_n)$ et $x(t_{n+1})$ on a :

$$\begin{aligned}
x_{n+1} &= x(t_{n-1}) + 2\delta t f(t_n, x(t_n)) \\
&= x(t_{n-1}) + 2\delta t \dot{x}(t_n) \\
&= x(t_{n-1}) + 2\delta t \left(\dot{x}(t_{n-1}) + \ddot{x}(t_{n-1})\delta t + x^{(3)}(t_{n-1})\frac{1}{2}\delta t^2 + O(\delta t^3) \right) \\
&= x(t_{n-1}) + \dot{x}(t_{n-1})2\delta t + \ddot{x}(t_{n-1})2\delta t^2 + x^{(3)}(t_{n-1})\delta t^2 + O(\delta t^3).
\end{aligned}$$

On a donc $|x(t_{n+1}) - x_{n+1}| = O(\delta t^3)$, ce qui montre que le schéma est bien d'ordre 2.

Toutes les notions que nous avons introduites précédemment peuvent être généraliser pour les schémas multipas, et l'analyse de convergence repose typiquement sur un schéma de preuve de type "stabilité et consistance impliquent convergence". Pour les schémas multi-pas, il faut simplement analyser précisément la stabilité du schéma numérique : pour une fonction f Lipschitz, le schéma peut être consistant, alors que des solutions numériques parasites polluent la solution et le rendent instable, donc inutilisable en pratique. Pour analyser ce phénomène, il faut en fait introduire la notion de zéro-stabilité : un schéma est zéro-stable si la solution du schéma numérique appliqué à l'équation différentielle ordinaire avec second membre nulle ($f = 0$) reste bornée. Cette propriété est naturellement vérifiée pour les schémas à un pas, mais pas pour les schémas multipas. On illustre cette notion sur l'Exercice 44, et on renvoie à [3] pour plus de détails.

Exercice 44 *On considère le schéma numérique multi-pas suivant :*

$$x_{n+2} + 4x_{n+1} - 5x_n = \delta t(4f(t_{n+1}, x_{n+1}) + 2f(t_n, x_n)).$$

1. *Vérifier que ce schéma est consistant d'ordre 3.*
   Solution : *Sur l'intervalle de temps (t_n, t_{n+2}), pour la solution exacte, on a :*

$$x(t_{n+2}) = x(t_n) + \dot{x}(t_n)2\delta t + \ddot{x}(t_n)2\delta t^2 + x^{(3)}(t_n)\frac{4}{3}\delta t^3 + x^{(4)}(t_n)\frac{2}{3}\delta t^4 + O(\delta t^5)$$

Pour la solution approchée partant des valeurs exactes $x(t_n)$ et $x(t_{n+1})$ on a :

$$\begin{aligned}
x_{n+2} &= -4x(t_{n+1}) + 5x(t_n) + \delta t(4f(t_{n+1}, x(t_{n+1})) + 2f(t_n, x(t_n))), \\
&= -4\left(x(t_n) + \dot{x}(t_n)\delta t + \ddot{x}(t_n)\frac{1}{2}\delta t^2 + x^{(3)}(t_n)\frac{1}{6}\delta t^3 + x^{(4)}(t_n)\frac{1}{24}\delta t^4 + O(\delta t^5)) \right) \\
&\quad + 5x(t_n) + \delta t(4\dot{x}(t_{n+1}) + 2\dot{x}(t_n)), \\
&= x(t_n) - 2\dot{x}(t_n)\delta t - 2\ddot{x}(t_n)\delta t^2 - x^{(3)}(t_n)\frac{2}{3}\delta t^3 - x^{(4)}(t_n)\frac{1}{6}\delta t^4 + O(\delta t^5) \\
&\quad + 4\delta t\left(\dot{x}(t_n) + \ddot{x}(t_n)\delta t + x^{(3)}(t_n)\frac{1}{2}\delta t^2 + x^{(4)}(t_n)\frac{1}{6}\delta t^3 + O(\delta t^4) \right), \\
&= x(t_n) + 2\dot{x}(t_n)\delta t + 2\ddot{x}(t_n)\delta t^2 + x^{(3)}(t_n)\frac{4}{3}\delta t^3 + x^{(4)}(t_n)\frac{1}{2}\delta t^4 + O(\delta t^5).
\end{aligned}$$

On a donc $x(t_{n+2}) - x_{n+2} = O(\delta t^4)$, et la méthode est donc d'ordre 3.

2. *On considère le problème de Cauchy $\dot{x} = 0$ avec condition initiale $x(0) = 1$, sur l'intervalle de temps $[0, 1]$: la solution est donc constante. Montrer que la solution du schéma numérique appliqué à ce problème s'écrit*

$$x_n = A + B(-5)^n,$$

où A et B sont deux constantes déterminées par les conditions initiales x_0 et x_1. A quelles conditions sur (x_0, x_1) va-t'on bien obtenir une solution constante ? Pourquoi, en pratique, même avec cette condition initiale, la solution numérique va être mauvaise ?

3. *On considère maintenant le problème de Cauchy $\dot{x} = x$ avec condition initiale $x(0) = 1$, sur l'intervalle de temps $[0, 1]$. Le schéma ci-dessus appliqué à ce problème s'écrit*

$$x_{n+2} + 4(1 - \delta t)x_{n+1} - (5 + 2\delta t)x_n = 0.$$

On obtient la solution générale sous la forme

$$x_n = A\lambda_1(\delta t)^n + B\lambda_2(\delta t)^n,$$

où A et B sont deux constantes déterminées par les conditions initiales ($x_0 = 1, x_1 = e^{\delta t}$), et où $\lambda_1(\delta t) > 0 > \lambda_2(\delta t)$ sont les deux racines du trinôme

$$\lambda^2 + 4(1 - \delta t)\lambda - (5 + 2\delta t) = 0.$$

Vérifier que $\lambda_1(\delta t) = 1 + \delta t + O(\delta t^2)$ et que $\lambda_2(\delta t) = -5 + O(\delta t)$. En déduire que la partie $B\lambda_2(\delta t)^n$ détériore fortement la solution numérique quand $\delta t \to 0$ (on parle de solution parasite). Programmer la méthode pour vérifier cela numériquement.

.

5.1.5 Une analyse plus fine de l'erreur, méthodes d'extrapolation

L'analyse que l'on a faite précédemment peut être un peu plus affinée : on peut obtenir des développements asymptotiques de l'erreur $|x(t_n) - x_n|$ en fonction du pas de temps δt. L'intérêt de cette analyse est que l'on peut ensuite utiliser des méthodes d'extrapolation pour augmenter l'ordre d'une méthode, ou estimer *a posteriori* l'erreur faite par un schéma.

Théorème 11 *On suppose que l'erreur de troncature $\varepsilon(t, x; t + \delta t)$ admet en tout point x un développement limité de la forme*

$$\varepsilon(t, x; t + \delta t) = d_{p+1}(t)\delta t^{p+1} + \ldots + d_{P+1}(t)\delta t^{P+1} + O(\delta t^{P+2}).$$

On suppose de plus que la fonction incrément satisfait la condition de consistance $\Psi(t, x; t) = f(t, x)$ et est suffisamment régulière. Alors, l'erreur globale admet également un développement limité de la forme

$$x(t_n) - x_n = e_p(t_n)\delta t^p + \ldots + e_P(t_n)\delta t^P + O(\delta t^{P+1}).$$

Preuve : On se propose de montrer qu'il existe une fonction $e_p(t)$ régulière telle que $x(t_n)-x_n = e_p(t_n)\delta t^p + O(\delta t^{p+1})$. La démonstration pour obtenir les termes suivants se fait de la même façon.

Pour une solution exacte $(x(t))_{0\leq t\leq T}$ fixée, on cherche une fonction $e_p(t)$ telle que $x(t_n) - x_n = e_p(t_n)\delta t^p + O(\delta t^{p+1})$. On raisonne par condition nécessaire. Supposons que l'on connaisse cette fonction $e_p(t)$, et posons

$$\hat{x}_n = x_n + e_p(t_n)\delta t^p.$$

On voit que $\hat{x}_n$ est une approximation de $x(t_n)$ d'ordre $(p+1)$. Autrement dit, il faut que l'on construise e_p de telle sorte que le schéma numérique vérifié par $\hat{x}_n$ soit d'ordre $(p+1)$.

La fonction incrément $\hat{\Psi}$ associée à la suite des $\hat{x}_n$ s'obtient en écrivant :

$$\begin{aligned}
\hat{x}_{n+1} &= x_{n+1} + e_p(t_{n+1})\delta t^p, \\
&= x_n + \delta t\Psi(t_n, x_n; t_{n+1}) + e_p(t_{n+1})\delta t^p, \\
&= \hat{x}_n + \delta t\left(\Psi(t_n, \hat{x}_n - e_p(t_n)\delta t^p; t_{n+1}) + (e_p(t_{n+1}) - e_p(t_n))\delta t^{p-1}\right).
\end{aligned}$$

On obtient donc

$$\hat{\Psi}(t, x; t+\delta t) = \Psi(t, x - e_p(t)\delta t^p; t+\delta t) + (e_p(t+\delta t) - e_p(t))\delta t^{p-1}.$$

L'erreur de consitance associée à la fonction incrément $\hat{\Psi}$ est donc :

$$\begin{aligned}
\hat{\varepsilon}(t, x; t+\delta t) &= \phi(t, x; t+\delta t) - \left(x + \delta t\hat{\Psi}(t, x; t+\delta t)\right), \\
&= \phi(t, x; t+\delta t) - (x + \delta t\Psi(t, x; t+\delta t)) + \delta t\left(\Psi(t, x; t+\delta t) - \hat{\Psi}(t, x; t+\delta t)\right), \\
&= \varepsilon(t, x; t+\delta t) + \delta t\Big(\Psi(t, x; t+\delta t) - \Psi(t, x - e_p(t)\delta t^p; t+\delta t) \\
&\qquad\qquad - (e_p(t+\delta t) - e_p(t))\delta t^{p-1}\Big), \\
&= d_{p+1}(t)\delta t^{p+1} + \delta t\left(\frac{\partial\Psi}{\partial x}(t, x; t+\delta t)e_p(t)\delta t^p - e_p'(t)\delta t^p\right) + O(\delta t^{p+2}), \\
&= \left(d_{p+1}(t) + \frac{\partial\Psi}{\partial x}(t, x; t)e_p(t) - e_p'(t)\right)\delta t^{p+1} + O(\delta t^{p+2}).
\end{aligned}$$

On voit donc que si e_p satisfait l'équation différentielle ordinaire (on utilise le fait que $\Psi(t, x; t) = f(t, x)$)

$$e_p'(t) = d_{p+1}(t) + \frac{\partial f}{\partial x}(t, x(t))e_p(t),$$

et en posant par ailleurs $e_p(0) = 0$, on a, pour tout $t \in [0, T]$,

$$\hat{\varepsilon}(t, x(t); t+\delta t) = O(\delta t^{p+2}).$$

Par la Proposition 8, on obtient que $\hat{x}_n - x(t_n) = O(\delta t^{p+1})$, et donc $x(t_n) - x_n = e_p(t_n)\delta t^p + O(\delta t^{p+1})$, ce qui termine la preuve. ♢

Expliquons maintenant l'utilité de ce développement limité de l'erreur.

Une première utilisation de ce développement limité est *l'extrapolation*. Pour un schéma donné d'ordre p, notons $x_n^{\delta t}$ la solution obtenue avec un pas de temps δt au n-ième pas de temps, et donc au temps $n\delta t$. On a

$$x(t_n) - x_n^{\delta t} = e_p(t_n)\delta t^p + O(\delta t^{p+1}),$$

$$x(t_n) - x_{2n}^{\delta t/2} = e_p(t_n)(\delta t/2)^p + O(\delta t^{p+1}),$$

et donc

$$x(t_n) - \frac{2^p x_{2n}^{\delta t/2} - x_n^{\delta t}}{2^p - 1} = O(\delta t^{p+1}). \tag{24}$$

En combinant les résultats obtenus par le schéma avec deux pas de temps différents, on peut donc gagner un ordre dans la vitesse de convergence.

Une deuxième utilisation, plus importante en pratique, est *l'estimation d'erreur a posteriori*. En effet, la relation (24) se réécrit :

$$x(t_n) - x_{2n}^{\delta t/2} = \frac{x_{2n}^{\delta t/2} - x_n^{\delta t}}{2^p - 1} + O(\delta t^{p+1}).$$

Cette estimée peut être utilisée comme estimée *a posteriori* pour choisir le pas de temps de manière adaptative. Considérons sans perte de généralité le premier pas de temps où l'on calcule une approximation de $x(\delta t)$. On calcule $x_2^{\delta t/2}$ (2 pas du schéma avec un pas de temps $\delta t/2$) et $x_1^{\delta t}$ (1 pas du schéma avec un pas de temps δt), et on a

$$x(\delta t) - x_2^{\delta t/2} = \frac{x_2^{\delta t/2} - x_1^{\delta t}}{2^p - 1} + O(\delta t^{p+1}).$$

La quantité $\frac{x_2^{\delta t/2} - x_1^{\delta t}}{2^p - 1}$ fournit une approximation calculable (estimée *a posteriori*) de l'erreur. On diminue le pas de temps si cette erreur est supérieure à un seuil fixé à l'avance, jusqu'à obtenir une erreur inférieure à ce seuil. La même technique d'estimation d'erreur s'applique ensuite à chaque pas de temps.

5.2 Stabilité absolue et problèmes raides

5.2.1 Stabilité absolue

Les analyses de convergence faites à la Section 5.1.3 peuvent s'avérer inutilisables en pratique car les constantes C dans les Propositions 6 et 8 peuvent être très grandes, et les estimations peuvent donc être en pratique très mauvaises. En particulier, ces estimées permettent de comparer la solution exacte et la solution approchée sur un intervalle de temps borné, mais ne disent rien sur le comportement en temps long (puisque la constante C dépend exponentiellement du temps d'intégration T).

Illustrons ces difficultés sur un problème test très simple. On considère la solution du problème $\dot{x} = -\mu x$, avec $\mu > 0$, soit $x(t) = x(0)\exp(-\mu t)$. Utilisons maintenant un schéma d'Euler explicite pour discrétiser l'équation :

$$x_{n+1} = x_n - \mu\delta t x_n.$$

On a donc $x_n = (1 - \mu\delta t)^n x_0$. En particulier, si $\delta t > 2/\mu$, on voit que la suite des (x_n) est non bornée alors que la solution exacte tend vers 0 quand $t \to \infty$.

On voit que dans de telles cas, l'analyse du comportement en temps long est cruciale (et même plus importante que l'analyse de la consistance du schéma) pour obtenir des approximations raisonnables. On étudie dans ce cadre *la stabilité absolue* (où A-stabilité) du schéma (à ne pas confondre avec la stabilité du schéma introduite précédemment dans la Proposition 7). La notion de stabilité absolue est très liée à la notion de stabilité asymptotique introduite dans la Section 3 pour les problèmes continus en temps.

Pour cette analyse, on suppose δt fixé, et on fait tendre T et donc $N(\delta t) = T/\delta t$ vers l'infini. Pour comprendre ce qui se passe, on considère un problème test :

$$\dot{x} = \lambda x, \tag{25}$$

avec $\lambda \in \mathbb{C}$.

Pour justifier l'importance de ce problème modèle, on peut considérer qu'à chaque pas de temps, après linéarisation du second membre et diagonalisation du jacobien (en le supposant diagonalisable), on a essentiellement des problèmes de ce type à résoudre.

Pour le problème continu, il est clair que $x(t)$ tend vers 0 quand $t \to \infty$ pour toute condition initiale dès que $\mathrm{Re}\ \lambda < 0$. On dira d'une méthode numérique qu'elle est absolument stable si elle vérifie la même propriété :

Définition 10 *Une méthode numérique est dite absolument stable si, quand elle est appliquée au problème test (25) avec* $\mathrm{Re}\ \lambda < 0$*, elle vérifie : pour toute condition initiale* x_0*,*

$$\lim_{n\to\infty} |x_n| = 0.$$

Remarque 24 *L'idée de chercher des schémas numériques tels que les approximations vérifient des propriétés satisfaites par le problème continu est une problématique très répandue en analyse numérique. On s'intéresse ici au comportement en temps long du schéma. On peut chercher par exemple à conserver au niveau discret des invariants de la dynamique continue. Dans le cas des systèmes hamiltoniens, de bons schémas pour vérifier cette propriété sont les schémas symplectiques (cf. [5]).*

Les schémas à un pas que nous avons considérés ci-dessus appliqués au problème test (25) s'écrivent :

$$x_{n+1} = R(\lambda\delta t)x_n$$

où $R : \mathbb{C} \to \mathbb{C}$ est appelé la fonction de stabilité de la méthode.

Remarque 25 *Pour des raisons d'homogénéité, il est normal que les paramètres* λ *et* δt *n'apparaissent dans les schémas numériques que sous la forme du produit* $\lambda\delta t$*.*

Clairement, pour qu'une méthode numérique soit absolument stable, il faut que si $\mathrm{Re}\ \lambda < 0$ (ce qui est équivalent à $\mathrm{Re}\ z < 0$, pour $z = \lambda\delta t$), alors $|R(z)| < 1$.

Définition 11 *L'ensemble des* $z \in \mathbb{C}$ *tels que* $|R(z)| < 1$ *s'appelle la région de stabilité absolue de la méthode numérique :*

$$\mathcal{A} = \{z \in \mathbb{C},\ |R(z)| < 1\}.$$

On a clairement

Proposition 9 *Un schéma numérique est absolument stable quelque soit les paramètres λ et δt si et seulement si $\{z \in \mathbb{C}, \mathrm{Re}\, z < 0\} \subset \mathcal{A}$. On parle alors de schéma (inconditionnellement) absolument stable.*

En pratique, on peut utiliser le résultat suivant qui découle du principe du maximum pour les fonctions analytiques (cf. [4, p.43]) :

Proposition 10 *Un schéma numérique est absolument stable si et seulement si $|R(i\lambda)| < 1$ pour tout $\lambda \in \mathbb{R}$, et R est analytique sur le demi-plan complexe $\{z \in \mathbb{C}, \mathrm{Re}\,(z) < 0\}$.*

Remarque 26 *Pour certains auteurs (cf. par exemple [4, 9]), la région de stabilité absolue est $\{z \in \mathbb{C}, |R(z)| \leq 1\}$ et une méthode est (inconditionnellement) absolument stable si $\{z \in \mathbb{C}, \mathrm{Re}\, z \leq 0\} \subset \{z \in \mathbb{C}, |R(z)| \leq 1\}$. Dans ce cas, la méthode est absolument stable si, quand elle est appliquée au problème test (25) avec $\mathrm{Re}\, \lambda \leq 0$, elle vérifie : pour toute condition initiale x_0, $\exists C > 0$, $\forall n > 0$, $|x_n| < C$. On comparera avec les deux définitions de stabilité pour les points fixes : Définition 4 et Définition 5.*

Regardons quelques exemples. Pour le schéma d'Euler explicite, on a $R(z) = (1 + z)$. La région de stabilité est le disque de centre -1 et de rayon 1. Le schéma est donc stable sous la condition $|1 + z| < 1$, ce qui impose une restriction sur le pas de temps. On dit que le schéma est conditionnellement absolument stable.

Pour le schéma d'Euler implicite, $R(z) = (1 - z)^{-1}$, et donc la région de stabilité est l'ensemble des z tels que $|1 - z| > 1$ (le complémentaire du disque de centre 1 et de rayon 1). Le schéma est inconditionnellement absolument stable.

Exercice 45 *Etudier la stabilité absolue des schémas numériques introduits précédemment. A quelle condition sur $\theta \in [0, 1]$ les θ-schémas :*

$$x_{n+1} = x_n + \delta t f\left(t_n + \theta\delta t, (1 - \theta)x_n + \theta x_{n+1}\right)$$

ou

$$x_{n+1} = x_n + \delta t\left((1 - \theta)f(t_n, x_n) + \theta f(t_{n+1}, x_{n+1})\right)$$

sont A-stables ?

Solution *: Pour la méthode des trapèzes (Crank-Nicolson) ou la méthode du point milieu (qui sont les mêmes méthodes quand on les applique à un problème linéaire), on obtient $R(z) = \frac{1+z/2}{1-z/2}$. On a donc $\mathcal{A} = \{z \in \mathbb{C}, |z+2| < |z-2|\} = \{z \in \mathbb{C}, \mathrm{Re}\, z < 0\}$, et le schéma est donc A-stable. De manière générale pour les θ-schémas, on obtient $R(z) = \frac{1+(1-\theta)z}{1-\theta z}$, et on vérifie que le schéma est A-stable si et seulement si $\theta \in [0.5, 1]$. Il existe donc des méthodes implicites qui ne sont pas A-stables.*

Pour Heun, on obtient $R(z) = 1 + z + z^2/2$, et le schéma n'est donc pas A-stable car si $z \in \mathbb{R}$ tend vers $-\infty$, alors $|R(z)|$ tend vers $+\infty$.

Ces exemples sont caractéristiques de ce que l'on observe en général : les schémas explicites ne peuvent qu'être conditionnellement absolument stable (cf. [6, p.366], ou [7, p.482]), et les schémas implicites sont en général inconditionnellement stables.

Exercice 46 *On considère un schéma multi-pas linéaire à k pas du type (23) que l'on écrit sous la forme générale :*

$$\alpha_k x_{n+k} + \alpha_{k-1} x_{n+k-1} + \ldots + \alpha_0 x_n = \delta t(\beta_k f(t_{n+k}, x_{n+k}) + \ldots + \beta_0 f(t_n, x_n).$$

Etudier la stabilité absolue de ce schéma.

Solution *: On applique le schéma au problème modèle (25) et on obtient :*

$$\alpha_k x_{n+k} + \alpha_{k-1} x_{n+k-1} + \ldots + \alpha_0 x_n = \lambda\delta t(\beta_k x_{n+k} + \ldots + \beta_0 x_n).$$

On sait qu'une solution générale de ce problème s'écrit sous la forme

$$x_n = p_1(n)\zeta_1^n + \ldots + p_l(n)\zeta_l^n$$

où les $(\zeta_i)_{1\leq i\leq l}$ *sont les racines (dans* $\mathbb{C}$*) de multiplicité respectives* $(m_i)_{1\leq i\leq l}$ *du polynôme caractéristique*

$$(\alpha_k - \lambda\delta t\beta_k)\zeta^k + \ldots + (\alpha_0 - \lambda\delta t\beta_0)$$

et les fonctions $(p_i)_{1\leq i\leq l}$ *sont des polynômes de degré* $m_i - 1$*. Le domaine de stabilité est donc :*

$$\mathcal{A} = \Big\{ z \in \mathbb{C},\ |\zeta_i(z)| < 1 \text{ avec } (\zeta_i(z))_{1\leq i\leq l} \text{ les racines du polynôme } p(\zeta) = (\alpha_k - z\beta_k)\zeta^k + \ldots + (\alpha_0 - z\beta_0) \Big\}.$$

5.2.2 Problèmes raides

Considérons l'exemple suivant :

$$\dot{x} = Mx \tag{26}$$

avec

$$M = \begin{bmatrix} -1 & 0 \\ 0 & -\mu \end{bmatrix}$$

avec μ un nombre positif très grand. La solution s'écrit $x(t) = (x_1(0)e^{-t}, x_2(0)e^{-\mu t})$. La deuxième composante converge beaucoup plus vite vers 0 que la première composante : il y a deux échelles de temps séparées dans le système, ce qui est une caractéristique typique des *problèmes raides*. Si $x_2(0)$ est très petit, ou bien t assez grand, on a $x(t) \approx (x_1(0)e^{-t}, 0)$.

Utilisons maintenant un schéma d'Euler explicite pour résoudre ce problème :

$$x_{n+1} = x_n + \delta t M x_n.$$

On obtient $x_n = (x_1(0)(1-\delta t)^n, x_2(0)(1-\mu\delta t)^n)$. On voit que pour obtenir une solution raisonnable, il faut satisfaire deux conditions :

$$(1-\delta t)^n \approx e^{-n\delta t},$$

et

$$|1 - \mu\delta t| < 1.$$

La première condition est une condition de précision. La deuxième condition est une condition de stabilité. Les deux imposent une limitation sur le pas de temps si on veut bien approcher la solution exacte $x(t)$. Si μ est grand, on s'aperçoit que c'est la condition de stabilité $\delta t < 2/\mu$ qui est limitante, alors que la deuxième composante de $x(t)$ (où ce paramètre μ apparaît) a un effet négligeable sur la solution. Bien sûr, ce problème modèle est un peu trop simple, puisque la variable rapide du système est en fait une composante du vecteur $x(t)$, et que l'on pourrait ignorer cette composante puisqu'elle n'influe pas sur le résultat de manière significative. Cependant, de manière générale, il n'est pas toujours aussi simple de séparer ainsi les variables rapides des variables lentes : penser au même problème, avec une matrice M diagonalisable et avec un spectre très étendu.

Exercice 47 *On considère l'équation de la chaleur*

$$\partial_t u = \partial_{x,x} u$$

pour $t \in [0,T]$ et $x \in [0,1]$, avec des conditions de Dirichlet homogènes au bord $u(t,0) = u(t,1) = 0$ et une condition initiale u_0. On discrétise le problème en espace par une méthode de différences finies sur un maillage uniforme de pas $\delta x = 1/(I+1)$ du segment $[0,1]$, où $(I+1)$ est le nombre d'intervalles. On obtient alors le problème suivant

$$\partial_t U = AU,$$

avec $U(t)$ un vecteur de dimension I : pour $1 \leq i \leq I$, $U_i(t)$ est une approximation de $u(t, i\delta x)$ (on pose donc $U_i(0) = u_0(i\delta x)$). La matrice A est la matrice de taille $I \times I$:

$$A = \frac{1}{\delta x^2} \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & (0) \\ & \ddots & \ddots & \ddots & \\ (0) & & 1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix}.$$

Montrer que la matrice A est diagonalisable, de valeurs propres, pour $1 \leq k \leq I$, $\lambda_k = \frac{2}{\delta x^2}(1 - \cos(k\pi\delta x))$ et de vecteurs propres associés $X_k = (\sin(k\pi j\delta x))_{1\leq j\leq N}$. Que peut-on en déduire sur la discrétisation en temps de l'équation de la chaleur ?

Solution *: On observe que $\frac{\lambda_I}{\lambda_1} = \frac{\sin^2(I\pi/(2(I+1)))}{\sin^2(\pi\delta x/(2(I+1)))} \sim \frac{4I^2}{\pi^2}$, dans la limite $I \to \infty$. Plus le pas d'espace est petit, plus le problème devient raide. Par conséquent, on rencontrera certainement des problèmes de stabilité si on utilise une méthode explicite pour discrétiser le problème en temps.*

Ces observations nous amènent à la "définition" suivante des problèmes raides :

Définition 12 *Un problème raide est un problème pour lequel, quand on l'approche par un schéma numérique avec une région de stabilité absolue de taille finie (donc typiquement par une méthode explicite), c'est l'exigence de stabilité absolue qui limite le pas de temps, au lieu de l'objectif de précision de l'approximation numérique.*

Autrement dit, "Les problèmes raides sont des problèmes pour lesquels les méthodes explicites ne marchent pas" [4, p. 2]. Ces définitions ne sont pas vraiment

mathématiques mais plutôt heuristiques. On trouve donc beaucoup de variantes suivant les auteurs.

Pour les problèmes linéaires du type (26), un problème est raide si le spectre de la matrice M contient des valeurs propres très petites et des valeurs propres très grandes. Autrement dit, le système présente des échelles de temps très petites et des échelles de temps très grandes. C'est cette "caractérisation" en termes d'échelles de temps qui est la plus générale, et qui s'applique également aux problèmes non-linéaires. Pour des problèmes non-linéaires, on retiendra cependant que ces différentes échelles de temps ne se lisent pas forcément sur le spectre du Jacobien du second membre. De manière très générale, les problèmes de raideur sont liés au comportement dynamique de la solution exacte.

Au vu de l'analyse précédente, on comprend qu'une manière de traiter les problèmes raides est d'utiliser des schémas implicites. Si on reprend l'exemple du début de cette section, et que l'on approxime la solution par un schéma d'Euler implicite, on obtient

$$x_{n+1} = x_n + \delta t M x_{n+1}$$

et donc $x_n = (x_1(0)(1+\delta t)^{-n}, x_2(0)(1+\mu\delta t)^{-n})$. Cette fois, même si $\mu\delta t$ est grand, la seconde composante tend vers 0 dans la limite n grand. On n'a donc plus de problème de stabilité. Bien sûr, le prix à payer est de résoudre le problème implicite à chaque pas de temps (ce qui peut être difficile si le pas de temps est trop grand cf. Section 5.1.2).

5.2.3 Réduction de systèmes

Dans cette section, on se propose d'aborder le problème des problèmes raides sous un angle plus analytique. On a vu que dans un problème, certaines variables rapides rendent la résolution numérique difficile, alors que le détail de la dynamique de ces variables n'est pas importante pour le système global. Une idée est alors d'essayer d'éliminer ces variables rapides du système complet, pour obtenir un système plus simple car de dimension plus petite, et moins raide. On peut ensuite discrétiser ce problème plus simple de manière plus efficace que le problème original.

On suppose donc dans cette section que l'on connaisse les variables rapides du système, et que le problème peut s'écrire sous une forme particulière permettant de réduire le système. L'objectif est plus de donner des idées sur la manière de traiter le problème, plutôt que d'énoncer un résultat rigoureux. On renvoie par exemple à [2] pour des résultats plus précis et une bibliographie complète.

Considérons par exemple le système, pour un paramètre $\varepsilon > 0$

$$\begin{cases} \dot{x_\varepsilon}(t) = f(x_\varepsilon(t), y_\varepsilon(t), \varepsilon), \\ \varepsilon \dot{y_\varepsilon}(t) = g(x_\varepsilon(t), y_\varepsilon(t), \varepsilon), \end{cases} \tag{27}$$

avec comme condition initiale $(x_\varepsilon(0), y_\varepsilon(0)) = (\alpha, \beta)$. Quand $\varepsilon \to 0$, la variable y_ε varie de plus en plus vite : c'est la variable rapide du système. Une question naturelle est de savoir quelle est la limite lorsque ε tend vers 0 du système (27). Il est naturel d'introduire le système

$$\begin{cases} \dot{x_0}(t) = f(x_0(t), y_0(t), 0), \\ 0 = g(x_0(t), y_0(t), 0), \end{cases} \tag{28}$$

avec comme condition initiale $(x_0(0), y_0(0)) = (\alpha, \beta)$. Il n'est pas clair que la solution de (27) converge vers la solution de (28). En particulier les résultats de stabilité de la solution d'une équation différentielle ordinaire par rapport à une perturbation du type de ceux du Lemme 4 ne s'appliquent pas car le petit paramètre est en facteur de la dérivée en temps : on parle de *perturbation singulière*.

Pour comprendre ce qui se passe, on regarde le problème à une autre échelle en temps en introduisant le temps rapide $\tau = \frac{t}{\varepsilon}$. On pose $\overline{x}_\varepsilon(\tau) = x_\varepsilon(\varepsilon\tau)$ et $\overline{y}_\varepsilon(\tau) = y_\varepsilon(\varepsilon\tau)$. On a donc

$$\left\{ \begin{array}{l} \partial_\tau \overline{x}_\varepsilon(\tau) = \varepsilon f(\overline{x}_\varepsilon(\tau), \overline{y}_\varepsilon(\tau), \varepsilon), \\ \partial_\tau \overline{y}_\varepsilon(\tau) = g(\overline{x}_\varepsilon(\tau), \overline{y}_\varepsilon(\tau), \varepsilon), \end{array} \right. \tag{29}$$

avec comme condition initiale $(\overline{x}_\varepsilon(0), \overline{y}_\varepsilon(0)) = (\alpha, \beta)$. Cette fois, on peut (sous de bonnes hypothèses de régularité sur f et g) considérer que ce problème est une perturbation régulière du problème

$$\left\{ \begin{array}{l} \partial_\tau \overline{x}_0(\tau) = 0, \\ \partial_\tau \overline{y}_0(\tau) = g(\overline{x}_0(\tau), \overline{y}_0(\tau), 0), \end{array} \right. \tag{30}$$

avec conditions initiales $(\overline{x}_0(0), \overline{y}_0(0)) = (\alpha, \beta)$. On voit donc que y_ε varie d'abord très rapidement et peut être approché dans une couche limite autour de $t = 0$ par la solution du problème

$$\partial_\tau \overline{y}_0(\tau) = g(\alpha, \overline{y}_0(\tau), 0), \tag{31}$$

avec comme condition initiale $\overline{y}_0(0) = \beta$, alors que x_ε reste sur cette échelle de temps approximativement constant égal à α. L'équation (31) est appelée l'équation rapide du système. On peut réduire le problème initial à condition que la solution de ce problème rapide ait un comportement en temps long "simple".

Supposons par exemple que la solution de (31) tende vers une solution stationnaire asymptotiquement stable $\overline{y} = \xi(\alpha)$ avec

$$g(\alpha, \xi(\alpha), 0) = 0. \tag{32}$$

Si ce problème admet plusieurs solutions, la fonction ξ peut dépendre de la condition initiale β qui détermine dans quel bassin d'attraction la variable rapide va tomber. On a besoin que la solution stationnaire soit asymptotiquement stable, et ceci peut être typiquement vérifié en regardant si les valeurs propres du jacobien $\partial_y g(x, \xi(x), 0)$ sont bien de parties réelles strictement négatives. La variété des $\{(x, \xi(x))\}$ s'appelle la variété lente du problème. Une transition rapide amène donc la solution de (27) au voisinage de la variété lente.

Ensuite, la solution évolue sur la variété lente, et est bien approchée par le problème réduit (28). On comprend de cette analyse que l'on s'attend typiquement à une convergence uniforme sur tout intervalle de temps $t \in [0, T]$ de x_ε vers x_0, et sur tout intervalle de temps $t \in [t_1, T]$ (pour $t_1 > 0$) de y_ε vers y_0. Ceci nécessite que pour tout $t \in [0, T]$, $y_0(t) = \xi(x_0(t))$ est une racine isolée stable (au sens suivant : les valeurs propres du jacobien $\partial_y g(x_0(t), y_0(t), 0)$ sont de parties réelles strictement négatives) de l'équation $g(x_0(t), y_0(t), 0) = 0$.

Numériquement, il faut utiliser des techniques particulières pour résoudre le problème réduit (28), qui est une équation différentielle algébrique.

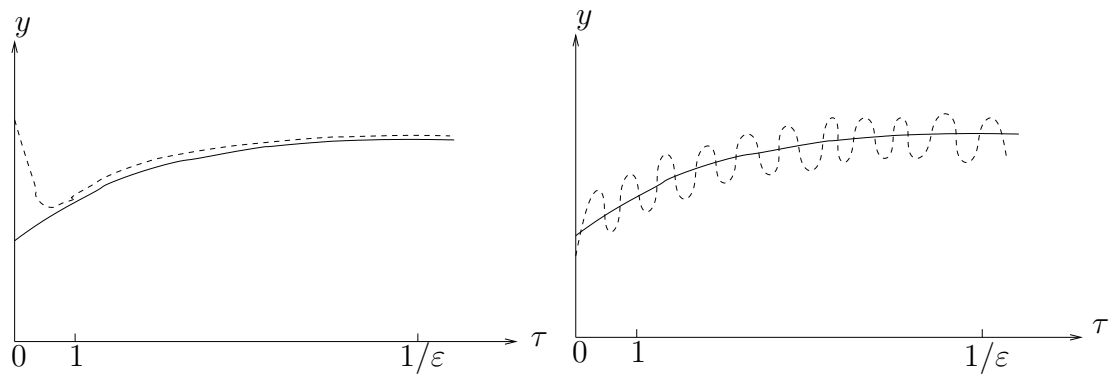

FIG. 9 – Dynamiques typique pour y_ε (en pointillé), et dynamiques effectives associées (en trait plein). A gauche, après une période transitoire très courte autour de $\tau = 0$, les échelles rapides se stabilisent et on évolue sur la variété lente du problème. A droite, les échelles rapides ne sont pas asymptotiquement stables, mais se moyennisent autour d'une dynamique effective.

Théorème 12 (Tikhonov) *On suppose que l'équation $g(x, y, 0) = 0$ admet une solution $y = \xi(x)$, avec ξ une fonction régulière et $\partial_y g(x, \xi(x), 0)$ une matrice dont les valeurs propres sont à parties réelles strictement négatives. On suppose que le système réduit*

$$\dot{x}(t) = f(x(t), \xi(x(t)), 0), \tag{33}$$

avec condition initiale $x(0) = \alpha$ admet une solution pour $t \in [0, T]$. Alors, pour ε suffisamment proche de 0, le système (27) admet une solution (x_ϵ, y_ϵ) sur $[0, T]$, si $y_\epsilon(0)$ appartient au bassin d'attraction du point d'équilibre $\xi(\alpha)$ du sous système rapide

$$\partial_\tau \overline{y}_0(\tau) = g(\alpha, \overline{y}_0(\tau), 0).$$

De plus, on a x_ε et y_ε convergent vers la solution du problème réduit : la convergence est uniforme sur $[0, T]$ pour x_ε, et uniforme sur $[t_1, T]$ (pour tout $t_1 > 0$) pour y_ε.

Pour la preuve, on renvoie à [10].

Exercice 48 *Simuler le problème suivant sur ordinateur*

$$\begin{cases} \dot{x} = y, \\ \varepsilon \dot{y} = x - g(y), \end{cases} \tag{34}$$

et explorer les comportements des solutions pour différentes fonctions g.

Exercice 49 *On considère le problème proie-prédateur suivant :*

$$\begin{cases} \dot{x} = xy - x, \\ \varepsilon \dot{y} = y(2 - y) - xy, \end{cases} \tag{35}$$

avec x_0 et y_0 deux réels positifs, et ε un paramètre petit. La solution est telle que $x(t)$ et $y(t)$ sont positifs. Tracer le portrait de phase du problème et expliquer le comportement des solutions tracées sur la Figure 10 en terme de dynamique effective.

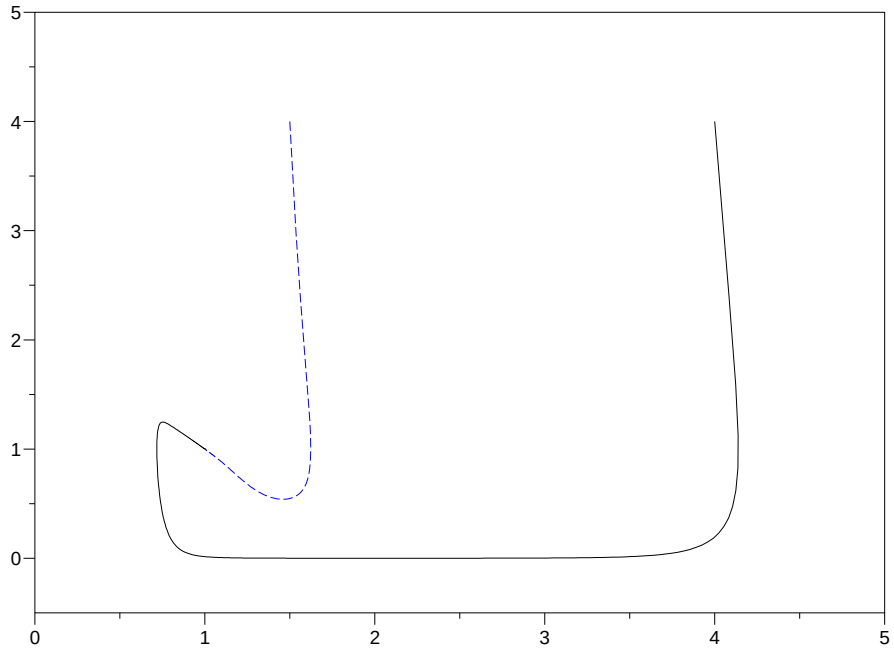

FIG. 10 – Trajectoires des solutions $(x(t), y(t))$ du système (35) dans le plan de phase pour $\varepsilon = 0.1$, $t \in [0, 50]$ et deux conditions initiales : $(x(0), y(0)) = (4, 4)$ et $(x(0), y(0)) = (1.5, 4)$. Dans les deux cas, les trajectoires convergent en temps long vers le point $(1, 1)$.

Solution : *La variété lente a deux composantes :* $\{(x,0)\}$ *et* $\{(x,2-x)\}$. *La composante* $\{(x,0)\}$ *de la variété est asymptotiquement stable si* $x>2$, *et instable si* $0<x<2$. *La composante* $\{(x,2-x)\}$ *de la variété est asymptotiquement stable si* $0<x<2$. *L'équation réduite sur* $\{(x,0)\}$ *est*

$$\dot{x} = -x,$$

et l'équation réduite sur $\{(x,2-x)\}$ *est*

$$\dot{x} = x(1-x).$$

Nous avons compris quelle était la bonne dynamique effective dans le cas où l'équation rapide du système converge vers un point fixe asymptotiquement stable. Il existe une autre situation pour laquelle il est possible d'écrire une dynamique effective : supposons que la dynamique rapide (31) soit ergodique par rapport à une mesure μ_α, c'est-à-dire que la solution de (31) est telle que : pour tout fonction test ϕ mesurable bornée,

$$\lim_{\tau\to\infty} \frac{1}{\tau}\int_0^\tau \phi(\overline{y}_0(\theta))\,d\theta = \int \phi d\mu_\alpha.$$

Dans ce cas, la dynamique effective s'écrit

$$\dot{X}_0(t) = F(X_0), \tag{36}$$

avec

$$F(x) = \int f(x,\cdot,0)d\mu_x. \tag{37}$$

Nous avons dans ce cas un phénomène de moyennisation.

Références

[1] L. Breiman. *Probability*. SIAM, 1992.

[2] D. Givon, R. Kupferman, and A. Stuart. Extracting macroscopic dynamics : model problems and algorithms. *Nonlinearity*, 17(6) :R55–R127, 2004.

[3] E. Hairer, S.P. Nørsett, and G. Wanner. *Solving ordinary differential equations I.* Springer, 1992.

[4] E. Hairer and G. Wanner. *Solving ordinary differential equations II.* Springer, 2002.

[5] F. Legoll and M. Lewin. Mathématiques des modèles multi-échelles. Cours ENPC.

[6] A. Quarteroni, R. Sacco, and F. Saleri. *Méthodes numériques pour le calcul scientifique*. Springer, 2000.

[7] A. Quarteroni, R. Sacco, and F. Saleri. *Numerical mathematics*. Springer, 2000.

[8] J.-N. Roux and G. Stoltz. Physique statistique et mécanique quantique. Cours ENPC.

[9] A.M. Stuart and A.R. Humphries. *Dynamical systems and numerical analysis*. Cambridge University Press, 1998.

[10] A. Tikhonov, A. Vasil'eva, and A. Sveshnikov. *Differential equations*. Springer, 1980.

[11] P. Walters. *Introduction to ergodic theory*. Springer-Verlag, 2000.