Outils d'Analyse

pour les

Equations Differentielles Ordinaires

*
* *

Ludovic Godard-Cadillac

BORDEAUX INP Enseirb-Matmeca

Année Universitaire 2025-2026

Table des matières

Presentation

Ce polycopié consacré à l'analyse des Équations Différentielles Ordinaires (EDO) a été conçu pour les étudiants en première année du département de mathématiques pour la mécanique de l'école d'ingénieur ENSEIRB-MATMECA (Groupe Bordeaux-INP). Il s'adresse plus généralement à toutes celles et ceux qui souhaitent acquérir des bases solides dans l'analyse de ces équations, tant sous les aspects relatifs à l'étude des solutions exactes que l'implémentation de méthodes numériques pour la résolution approchée. Le point-de-vue adopté dans ce document est volontiers théorique ; les éléments d'application de la théorie à des équations issues de la physique sera largement abordée lors des cours magistraux, des séances de travaux dirigés et de travaux pratiques.

 Son contenu a été construit sur la base du matériau pédagogique légué par les précédents responsables de ce cours, à savoir les professeurs Annabelle Collin et Kevin Santugini.

Important : Les éléments du polycopié signalés par le symbole **(†)** (appelé "*dague*" ou "*obèle*") sont les éléments les plus importants du cours, qu'il faut travailler en priorité. Il s'agit de résultats fondamentaux que l'on peut utiliser sans refaire la démonstration lors de la résolution des exercices. Le symbole **(‡)** ("*double-dague*" ou "*double-obèle*") signale en plus que ce résultat tombera comme question de cours le jour de l'examen... Les autres résultats (qui ne sont pas marqués) ne peuvent PAS être utilisés directement ; on s'en inspire à titre méthodologique, mais il faut savoir le refaire au cas-par-cas.

Chapitre 0

Rappels et compléments

0.1 Normes et topologie (†)

Définition 0.1 Espace vectoriel normé

Soit E un espace vectoriel réel et soit $\mathcal{N} : E \to \mathbb{R}_+$. On dit que l'application $\mathcal{N}$ est une *norme* sur E ssi elle vérifie les propriétés suivantes :

- **homogénéité :** $\forall\, \lambda \in \mathbb{R}, \quad \forall\, x \in E, \quad \mathcal{N}(\lambda x) = |\lambda| \mathcal{N}(x).$
- **séparation :** $\mathcal{N}(x) = 0 \quad \Longrightarrow \quad x = 0.$
- **inégalité triangulaire :** $\forall\, x, y \in E, \quad \mathcal{N}(x+y) \leq \mathcal{N}(x) + \mathcal{N}(y).$

Pour les normes, nous adopterons la notation usuelle suivante : $\|\_\|_E$.

L'inégalité triangulaire existe sous deux formulations équivalentes :

$$\forall\, x, y \in E, \quad \|x+y\|_E \leq \|x\|_E + \|y\|_E \qquad \Longleftrightarrow \qquad \forall\, x, y \in E, \quad \Big| \|x\|_E - \|y\|_E \Big| \leq \|x-y\|_E.$$

Exemple de normes en dimension finie : les normes ℓ^p pour $p \in [1, +\infty]$ sont définies sur l'espace $E = \mathbb{R}^d$ (en séparant le cas $p < +\infty$ et $p = +\infty$) par

$$\|x\|_{\ell^p} := \sqrt[p]{\sum_{k=1}^{d} |x_k|^p}\,, \qquad \text{ou} \qquad \|x\|_{\ell^\infty} := \max_{k=1,\dots,d} |x_k|. \tag{1}$$

Pour simplifier les notations par la suite, la norme euclidienne canonique sur $\mathbb{R}^d$, à savoir la norme ℓ^2, sera simplement notée $|\_|$ par analogie avec la valeur absolue sur $\mathbb{R}$.

Exemple de normes en dimension infinie : les normes L^p pour $p \in [1, +\infty]$ sont définies sur l'espace des fonctions continues par morceaux $E = \mathcal{C}^0_{pm}(\Omega; \mathbb{R}^p)$ par :

$$\|f\|_{L^p(\Omega)} := \sqrt[p]{\int_\Omega \big|f(x)\big|^p \,\mathrm{d}x}\,, \qquad \text{ou} \qquad \|f\|_{L^\infty(\Omega)} := \sup_{x \in \Omega} |f(x)|. \tag{2}$$

Définition 0.2 Convergence en norme

Soit $(E, \|\_\|_E)$ un espace vectoriel normé. Soit $(x_n)_{n\in\mathbb{N}}$ une suite d'éléments de E. On dit que (x_n) *converge vers* $x \in E$ ssi $\|x_n - x\|_E \longrightarrow 0$ lorsque $n \to +\infty$. Autrement dit, pour tout $\varepsilon > 0$, il existe un $N \in \mathbb{N}$ tel que pour tout $n \geq N$, on a $\|x_n - x\|_E \leq \varepsilon$.

En dimension finie, la notion de convergence est indépendante du choix de la norme :

THÉORÈME 0.1 Comparaison des normes en dimension infinie

Si $\|\_\|$ est une norme sur $\mathbb{R}^d$ alors $\quad \exists\, K > 0, \quad \forall\, X \in \mathbb{R}^d, \quad \dfrac{1}{K}|X| \;\leq\; \|X\| \;\leq\; K|X|.$

Dans le cas où la suite (x_n) est à valeurs réelles, on dispose de la notion plus faible de *limite supérieur* et de *limite inférieure* :

$$\limsup_{n\to+\infty} x_n \;:=\; \lim_{n\to+\infty} \sup_{k\geq n} x_k, \qquad \text{et} \qquad \liminf_{n\to+\infty} x_n \;:=\; \lim_{n\to+\infty} \inf_{k\geq n} x_k, \tag{3}$$

La *limite supérieure* et la *limite inférieure* sont toujours bien définies. La limite existe si et seulement si les limites supérieure et inférieure sont égales.

DÉFINITION 0.3 Boule ouverte et boule fermée

Soit $(E, \|\_\|_E)$ un espace vectoriel normé. Soit $x \in E$ et $r \geq 0$. On définit la *boule ouverte* de centre x et de rayon r par

$$\mathcal{B}(x, r) := \{y \in E \;:\; \|x - y\|_E < r\}.$$

On définit la *boule fermée* de centre x et de rayon r par

$$\overline{\mathcal{B}}(x, r) := \{y \in E \;:\; \|x - y\|_E \leq r\}.$$

DÉFINITION 0.4 Ensemble ouvert et ensemble fermé

Soit $(E, \|\_\|_E)$ un espace vectoriel normé et soit V un sous-ensemble de E. On dit que V est un *ensemble ouvert* ssi

$$\forall\, x \in V, \quad \exists\, r > 0, \quad \mathcal{B}(x, r) \subseteq V.$$

On dit que V est un *ensemble fermé* ssi son complémentaire est ouvert.

Dans la définition précédente, le $r > 0$ dépend implicitement du choix de $x \in V$.

Remarque : Il existe des ensembles qui ne sont ni ouverts ni fermés. Il existe des ensembles qui sont et ouverts et fermés.

DÉFINITION 0.5 Adhérence d'un ensemble

Soit $(E, \|\_\|_E)$ un espace vectoriel normé et soit V un sous-ensemble de E. On définit *l'adhérence* de V, notée $\overline{V}$, par

$$\overline{V} \;:=\; \left\{x \in E \;:\; \exists\, (x_n) \in V^{\mathbb{N}},\; x_n \longrightarrow x\right\}$$

L'adhérence de V est le plus petit ensemble fermé qui contient V.

Définition 0.6 Intérieur d'un ensemble

Soit $(E, \|\_\|_E)$ un espace vectoriel normé et soit V un sous-ensemble de E. On définit *l'intérieur* de V, noté $\overset{\circ}{V}$, par

$$\overset{\circ}{V} := \Big\{ x \in E \,:\, \exists\, r_x > 0,\ \mathcal{B}(x, r_x) \subseteq V \Big\}.$$

L'intérieur de V est le plus grand ensemble ouvert contenu dans V. Par ailleurs, on peut démontrer que $(\overset{\circ}{V})^c = \overline{V^c}$, où l'exposant ^{c} désigne le passage au complémentaire. On définit aussi la frontière de V par $\partial V := \overline{V} \setminus \overset{\circ}{V}$.

Définition 0.7 Ensemble compact

Soit $(E, \|\_\|_E)$ un espace vectoriel normé et soit V un sous-ensemble de E. On dit que V est un *ensemble compact* ssi pour toute suite (x_n) d'éléments de V on peut extraire une sous-suite qui converge dans V.

En dimension finie, un ensemble est compact ssi il est fermé et borné. En dimension infinie, la réciproque devient fausse : il est possible de trouver des ensembles fermés et bornés qui ne sont pas compacts. Par exemple, la boule unité fermée de l'espace des fonctions L^2 n'est *pas* compacte.

Pour Ω un ouvert de $\mathbb{R}^d$, on dit que $K \subset \mathbb{R}^d$ est *compactement inclus* dans Ω (et on le note $K \subset\subset \Omega$) ssi $\overline{K}$ est un compact inclus dans Ω.

Définition 0.8 Ensemble connexe

Soit $(E, \|\_\|_E)$ un espace vectoriel normé et soit V un sous-ensemble ouvert (resp. fermé) de E. On dit que V est un *ensemble connexe* ssi il n'est pas possible de l'écrire comme une réunion disjointe de deux ouverts (resp. fermés).

De manière intuitive, un ensemble (ouvert ou fermé) est connexe ssi il est "*en un seul morceau*". Pour un ensemble V donné, on peut définir les *composantes connexes* de V comme étant les sous-ensembles connexes de V maximaux pour l'inclusion (c'est-à-dire tels que tout sur-ensemble inclus dans V est nécessairement non-connexe).

Lemme 0.1 Ensemble ouvert connexe par arc

Soit $(E, \|\_\|_E)$ un espace vectoriel normé de dimension finie et soit $V \subseteq E$.

(i) Si on suppose V ouvert : alors V est connexe si et seulement si il est *connexe par arc* ; c'est-à-dire que pour tout $X_1, X_2 \in V$, il existe un chemin continu $\gamma : [0,1] \to V$ qui relie X_1 à X_2 au sens où $\gamma(0) = X_1$ et $\gamma(1) = X_2$.

(ii) Si on suppose V fermé : alors si V connexe, il est connexe par arc.

Dans le cas des ensembles fermés, il existe des contre-exemples pour la réciproque (un contre-exemple célèbre sur $\mathbb{R}^2$ s'appelle le *cercle polonais*).

0.2 Régularité des fonctions (†)

0.2.1 Continuité, dérivabilité, caractère $\mathcal{C}^1$

Définition 0.9 Fonction continue

Une fonction $\mathcal{F} : \Omega \subseteq \mathbb{R}^d \to \mathbb{R}^p$ est *continue* en un point $X \in \Omega$ ssi $\mathcal{F}(Y)$ tend vers $\mathcal{F}(X)$ lorsque $Y \in \Omega$ tend vers X.

Définition 0.10 Fonction dérivable

Une fonction $f : I \subseteq \mathbb{R} \to \mathbb{R}$ avec I un intervalle ouvert, est *dérivable* en un point $x \in I$ ssi l'accroissement fini

$$h \longmapsto \frac{f(x+h) - f(x)}{h}. \tag{4}$$

admet une limite lorsque $h \to 0$.

Cette limite est la dérivée de f en x, noté $\frac{df}{\mathrm{d}x}(x)$, ou parfois $f'(x)$. En physique, lorsque la variable représente le temps t, on note aussi avec un point : $\dot{f}(t)$. Si une fonction est dérivable en x alors elle est continue en x.

La notion de *dérivabilité* est bien définie pour les fonctions d'une seule variable réelle. Pour les fonctions de plusieurs variables, on utilise la notion de *dérivée partielle.* La k^e dérivée partielle d'une fonction $f : \Omega \to \mathbb{R}$ (avec Ω ouvert de $\mathbb{R}^d$) est la dérivée de f par rapport à la k^e composante de la variable, noté $\frac{\partial f}{\partial x_k}(X)$. On note $\nabla f(X)$ (le gradient de f) le vecteur colonne formé de chacune des dérivées partielles de f au point $X \in \Omega$.

Définition 0.11 fonction $\mathcal{C}^1$

$f : \Omega \to \mathbb{R}$ (Ω ouvert de $\mathbb{R}^d$) est *de classe* $\mathcal{C}^1$ ssi $X \in \Omega \mapsto \nabla f(X)$ est continue.

On définit de manière analogue les fonctions $\mathcal{C}^k$ et $\mathcal{C}^\infty$.

Remarque : Le gradient ∇f est également bien défini pour les fonctions à valeurs vectorielles : $\mathcal{F} = (f_1, \ldots, f_p)$. Il suffit pour cela de concaténer les gradients de chaque fonction partielle f_k (on obtient alors une matrice). On définit souvent la *matrice Jacobienne* comme étant la transposée du gradient :

$$\mathcal{J}ac[\mathcal{F}](X) := \nabla\mathcal{F}(X)^T = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_p}{\partial x_1} & \cdots & \frac{\partial f_p}{\partial x_d} \end{pmatrix}. \tag{5}$$

On peut également définir la matrice Hessienne pour les fonctions f à valeurs scalaires (c'est-à-dire à valeurs dans $\mathbb{R}$), comme étant le gradient du gradient :

$$\mathcal{H}ess[f](X) := \nabla^2 f(X) = \begin{pmatrix} \frac{\partial^2 f}{\partial {x_1}^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial {x_2}^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial {x_n}^2} \end{pmatrix}. \tag{6}$$

La matrice Hessienne de f est une matrice symétrique dès que la fonction f est de classe $\mathcal{C}^2$ (Théorème de Schwarz).

Toutes ces notions de gradient, jacobienne, hessienne, se généralisent à des fonctions à valeurs vectorielles, matricielles etc... dans le cadre de la théorie des tenseurs (hors programme).

0.2.2 Fonctions lipschitziennes

Dans le cadre de l'analyse des équations différentielles, nous allons avoir besoin d'une notion de régularité légèrement plus faible que la régularité dérivable. Pour cela, en s'inspirant de la formule (4), on introduit le *taux d'accroissement* de la fonction $\mathcal{F} : \Omega \subseteq \mathbb{R}^d \to \mathbb{R}^p$ au point $X \in \mathbb{R}^d$ par :

$$\tau_{\mathcal{F}}(X) \ := \ \limsup_{H \to 0} \ \frac{|\mathcal{F}(X+H) - \mathcal{F}(X)|}{|H|} \quad \in \mathbb{R}_+ \cup \{+\infty\}. \tag{7}$$

L'intérêt de la notion de taux d'accroissement réside dans le fait qu'il est bien défini même dans des cas où la fonction n'est pas $\mathcal{C}^1$ et nous permets d'étudier les variations d'une telle fonction aussi efficacement qu'avec la notion de dérivée. La définition du taux d'accroissement (7) implique que, localement au voisinage de $X \in \Omega$, le graphe de la fonction n'intersecte pas les cône de révolution issus de X et de pentes supérieures à $\tau_{\mathcal{F}}(X)$. Plus précisément, nous avons la propriété suivante (voir aussi Figure 1) :

Proposition 0.1 Cone d'exclusion

Soit $\mathcal{F} : \Omega \to \mathbb{R}$ (Ω ouvert de $\mathbb{R}^d$). Pour $X_0 \in \Omega$ et $\tau \geq 0$, on définit le cône suivant :

$$\Gamma^{\tau}_{X_0} \ := \ \Big\{ (X, y) \in \Omega \times \mathbb{R} \ : \ |y - \mathcal{F}(X_0)| \geq \tau \big| X - X_0 \big| \Big\}. \tag{8}$$

Alors, pour tout $\tau > \tau_{\mathcal{F}}(X_0)$, il existe un rayon r_τ tel que le graphe de $\mathcal{F}$ sur la boule $\mathcal{B}(X_0, \tau)$ n'intersecte pas ce cône. Autrement dit :

$$\Big\{ \big(X, f(X)\big) \ : \ X \in \Omega \cap \mathcal{B}(X_0, r_\tau) \Big\} \ \cap \ \Gamma^{\tau}_{X_0} \ = \ \emptyset. \tag{9}$$

A partir de la définition du taux d'accroissement, on définit alors naturellement la *constante de Lipschitz* de la fonction $\mathcal{F}$ comme étant le plus grand taux d'accroissement de cette fonction :

$$\lambda_{\mathcal{F}} \ := \ \sup_{X \in \Omega} \ \tau_{\mathcal{F}}(X). \tag{10}$$

Définition 0.12 Fonction lipschitzienne

Soit $\mathcal{F} : \Omega \to \mathbb{R}^p$ (Ω ouvert de $\mathbb{R}^d$). On dit que $\mathcal{F}$ est *lipschitzienne* ssi $\lambda_{\mathcal{F}} < +\infty$.

On dit que $\mathcal{F}$ est *localement lipschitzienne* ssi pour tout $K \subset\subset \Omega$ on a $\mathcal{F}_{|_K}$ lipschitzienne.

Pour la prononciation, dire : "*lip-chit-sienne*". On dit aussi souvent "lipschitz" par simplicité. Pour éviter les ambiguïtés, on précise parfois qu'une fonction est *globalement* lipschitzienne car les fonctions lipschitziennes sont localement lipschitzienne mais la réciproque est fausse.

Au niveau *intuitif*, il faut retenir que les fonctions lipschtziennes sont les fonctions dont toutes les pentes sont bornées. Ce sont des fonctions beaucoup plus régulières que les fonctions juste "continues" et sont parfois qualifiées de "presque différentiables".

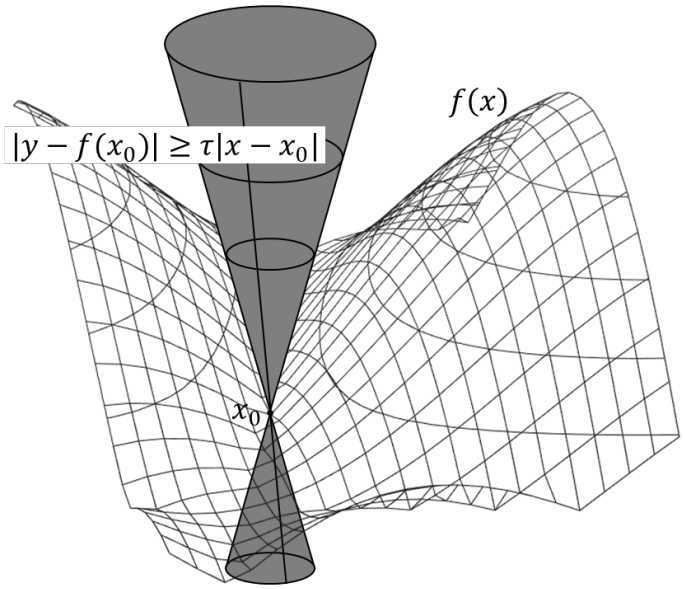


FIGURE 1 – Un cône d'exclusion au voisinage d'un point x_0 pour une fonction f lipschitzienne.

Exemples de fonctions globalement lipschitz (sur $\mathbb{R}$) :

$x \mapsto 2x + 1, \quad x \mapsto |x|, \quad x \mapsto \tanh(x), \quad x \mapsto \arccos(\cos(x)), \quad x \mapsto \sum_{n=0}^{+\infty} \frac{|\sin(nx)|}{n^3}.$

Exemples de fonctions localement mais pas globalement lipschitz :

$x \mapsto x^2, \quad x \mapsto e^x, \quad x \mapsto \sin(x^2), \quad x \mapsto |x^3 - x|, \quad x \mapsto x^3 \sin(1/x)$

Exemples de fonctions continues mais pas localement lipschitz :

$x \mapsto x \ln(|x|), \quad x \mapsto \sqrt[3]{x}, \quad x \mapsto x \sin(1/x), \quad x \mapsto \sqrt{|\sin(x)|}.$

PROPOSITION 0.2 Régularité des fonctions lipschitziennes

Soit $\mathcal{F} : \Omega \to \mathbb{R}^p$ (Ω ouvert de $\mathbb{R}^d$).

(i) Si $\mathcal{F}$ est $\mathcal{C}^1$, alors elle localement lipschitz. (†)

(ii) Si $\mathcal{F}$ est $\mathcal{C}^1$ à dérivée bornée, alors elle globalement lipschitz. Sa constante de Lipschitz vaut $\|\nabla \mathcal{F}\|_{L^\infty}$.

(iii) Si $\mathcal{F}$ est localement lipschitz, alors elle est continue.

Pour montrer qu'une fonction est localement lipschitz, il suffit souvent de montrer qu'elle est de régularité $\mathcal{C}^1$. Pour montrer qu'elle est globalement lipschitz, il suffit de calculer sa dérivée puis la majorer uniformément. Si cette approche ne fonctionne pas, on peut par exemple utiliser les propriétés complémentaires suivantes :

(iv) Si $\mathcal{F}$ est dérivable sauf en un nombre fini de points, avec une dérivée localement bornée, alors elle localement lipschitz.

(v) Si $\mathcal{F}$ est dérivable sauf en un nombre fini de points, avec une dérivée globalement bornée, alors elle globalement lipschitz.

Remarque : Tous les polynômes sont localement lipschitz. Les fonctions affines sont les seuls polynômes qui sont globalement lipschitz.

Proposition 0.3 Stabilité de l'ensemble des fonctions lipschitziennes

Soit $\mathcal{F} : \Omega \to \mathbb{R}^p$ (Ω ouvert de $\mathbb{R}^d$) et $\mathcal{G} : \Gamma \to \mathbb{R}^q$ (Γ ouvert de $\mathbb{R}^m$) deux fonctions globalement (resp. localement) lipschitz et soit $\lambda, \mu \in \mathbb{R}$.

(*i*) $\lambda\mathcal{F} + \mu\mathcal{G}$ est globalement (resp. localement) lipschitz sur $\Omega \cap \Gamma$ si $p = q$.

(*ii*) $\mathcal{F}^T\mathcal{G}$ est localement lipschitz (globalement pour $\mathcal{F}$ et $\mathcal{G}$ bornées) sur $\Omega \cap \Gamma$ si $p = q$.

(*iii*) $\mathcal{F} \circ \mathcal{G}$ est globalement (resp. localement) lipschitz sur Γ si $q = d$ et si $\mathcal{G}(\Gamma) \subset \Omega$.

(*iv*) $1/\mathcal{F} := (1/f_1, \ldots, 1/f_p)$ est localement lipschitz sur $\Omega \setminus \{x \in \Omega : \mathcal{F}(x) = 0\}$.

$\rightarrow$ Ces propriétés sont utiles pour montrer qu'une fonction est lipschitzienne en la décomposant en plusieurs fonctions lipschitziennes plus simples.

Dans le cadre de ce cours, nous utiliserons la version générale du théorème des accroissements finis dans sa version adaptée à la régularité lipschitzienne :

Théorème 0.2 Théorème des accroissements finis (†)

Soit $\mathcal{F} : \Omega \to \mathbb{R}^p$ (Ω ouvert de $\mathbb{R}^d$) une fonction lipschitzienne. Alors :

$$\lambda_{\mathcal{F}} = \sup_{\substack{X,Y \in \Omega \\ X \neq Y}} \frac{|\mathcal{F}(X) - \mathcal{F}(Y)|}{|X - Y|}. \tag{11}$$

Nous allons par la suite considérer des fonctions $\mathcal{F}$ qui dépendent d'un paramètre, en l'occurrence le temps $t \in [0, T[$ avec $T > 0$ (éventuellement égal à $+\infty$). Pour cela, nous allons généraliser la notion de fonction lipschizienne à des fonctions dépendant d'un paramètre réel.

Définition 0.13 Fonction partiellement lipschitzienne

Soit $\mathcal{F} : [0, T[\times\Omega \to \mathbb{R}^p$ (avec $T \in \mathbb{R}_+^* \cup \{+\infty\}$ et Ω ouvert de $\mathbb{R}^d$). On dit que $\mathcal{F}$ est *lipschitzienne pour sa deuxième variable* ssi elle est continue et si pour tout $t \in [0, T[$ fixé la fonction $\mathcal{F}(t, \cdot)$ est lipschitzienne avec des constantes de Lipschitz qui satisfont :

$$\sup_{t \in [0,T[} \lambda_{\mathcal{F}(t,\cdot)} < +\infty. \tag{12}$$

La fonction $\mathcal{F}$ est dite *localement lipschitzienne pour sa deuxième variable* si la propriété (12) est vérifiée par $\mathcal{F}_{|\Gamma}$ pour tout $\Gamma \subset\subset [0, T[\times\Omega$.

Dans l'énoncé ci-dessus, la notation $\mathcal{F}(t, \cdot)$ désigne la fonction $X \in \Omega \mapsto \mathcal{F}(t, X)$. Comme le nom de la variable est connu (en l'occurrence $X \in \Omega$ ici), alors on dit aussi de manière équivalente que $\mathcal{F}$ est *(localement) lipschitzienne par rapport à* X. Si une fonction est localement lipschitzienne alors elle est partiellement localement lipschitzienne. En particulier, les fonctions $\mathcal{C}^1$ sont localement partiellement lipschitziennes.

0.3 Comparaison asymptotique (†)

On rappelle ici les concepts essentiels pour comparer le comportement asymptotique des fonctions au voisinage d'un point $x_0 \in \mathbb{R}^d$. Ces notions se généralisent pour l'analyse à l'infini.

Définition 0.14 Fonctions négligeables et prépondérantes

Soit $f, g : \Omega \to \mathbb{R}$ (Ω ouvert de $\mathbb{R}^d$). On dit que f est *négligeable* devant g (ou bien g est *prépondérante* sur f) au voisinage d'un point $x_0 \in \Omega$ ssi

$$\frac{f(x)}{g(x)} \longrightarrow 0, \qquad \text{lorsque} \qquad x \to x_0.$$

On le note $f \ll g$ ou bien $f = o(g)$, tout en précisant le point x_0 étudié.

Exemple : $e^{-\frac{1}{x}} \ll x^2 \ll x \ll x\ln(x) \ll \sqrt{x} \ll \ln(x)^{-1}, \qquad \text{lorsque} \qquad x \to 0^+.$

Définition 0.15 Fonctions équivalentes

Soit $f, g : \Omega \to \mathbb{R}$ (Ω ouvert de $\mathbb{R}^d$). On dit que f est *équivalente à* g au voisinage d'un point $x_0 \in \Omega$ ssi

$$\frac{f(x)}{g(x)} \longrightarrow 1, \qquad \text{lorsque} \qquad x \to x_0.$$

On le note $f \sim g$ ou bien $f = g + o(g)$, tout en précisant le point x_0 étudié.

Exemple : $\sin(x) \sim x \quad \text{lorsque} \quad x \to 0.$

Définition 0.16 Fonctions dominantes et dominées

Soit $f, g : \Omega \to \mathbb{R}$ (Ω ouvert de $\mathbb{R}^d$). On dit que f est *dominée* par g (ou g est *dominante* sur f) au voisinage de $x_0 \in \Omega$ ssi il existe une constante $C > 0$ et un rayon $r > 0$ tels que

$$\forall\, x \in \Omega \cap \mathcal{B}(x_0, r), \qquad f(x) \;\le\; C\, g(x).$$

On le note $f \lesssim g$ ou bien $f = \mathcal{O}(g)$, tout en précisant le point x_0 étudié.

Exemple : $3x^2 \sin^2\left(\frac{1}{x}\right) \lesssim x^2 \quad \text{ou} \quad 3x^2 \sin^2\left(\frac{1}{x}\right) = \mathcal{O}(x^2), \quad \text{lorsque} \quad x \to 0.$

La notation "*petit o*" et "*grand* $\mathcal{O}$" sont très utiles pour écrire les développements limités :

$$\tanh(x) = z - \frac{z^3}{3} + \frac{2z^5}{15} - \frac{17z^7}{315} + \frac{62z^9}{2835} + \mathcal{O}\big(z^{11}\big).$$

Définition 0.17 Fonctions comparables

Soit $f, g : \Omega \to \mathbb{R}$ (Ω ouvert de $\mathbb{R}^d$). On dit que f et g sont *comparables* au voisinage d'un point $x_0 \in \Omega$ ssi il existe deux constantes $0 < c \le C$ et un rayon $r > 0$ tels que

$$\forall\, x \in \Omega \cap \mathcal{B}(x_0, r), \qquad c\, g(x) \;\le\; f(x) \;\le\; C\, g(x).$$

On le note $f \simeq g$, ou bien $f = \Theta(g)$.

Exemple : $3x^2\left(1 + \sin^2\left(\frac{1}{x}\right)\right) \simeq x^2 \quad \text{ou} \quad 3x^2\left(1 + \sin^2\left(\frac{1}{x}\right)\right) = \Theta(x^2), \quad \text{si} \quad x \to 0.$

Remarque : par convention la notation $\approx$ n'est pas utilisée pour les fonctions comparables. On réserve cette dernière notation pour des manipulations formelles et non rigoureuses en théorie de l'approximation pour faciliter la lecture et expliquer la démarche.

Il ne faut donc pas confondre "$\simeq$" (notation rigoureuse) et "$\approx$" (notation non-rigoureuse utilisée à des fins pédagogiques).

0.4 Autres résultats d'analyse à connaître (†)

Quelques résultats fondamentaux d'analyse qu'il faut absolument connaître :

Théorème 0.3 Théorème fondamental de l'analyse

Soit $f : I \to \mathbb{R}$ de classe $\mathcal{C}^1$ avec I un intervalle de $\mathbb{R}$.

$$\forall\, x, y \in I, \qquad f(x) = f(y) + \int_y^x f'(t)\,\mathrm{d}t.$$

Théorème 0.4 Théorème des valeurs intermédiaires

Soit $f : \Omega \to \mathbb{R}$ continue avec Ω un ouvert ***connexe*** de $\mathbb{R}^d$. Soit $X, Y \in \Omega$ tels que $f(X) \leq 0 \leq f(Y)$. Alors il existe $Z \in \Omega$ tel que $f(Z) = 0$.

Théorème 0.5 Théorème des croissances comparées

Pour tout $\alpha \geq 0$, on a $\quad x^\alpha \ll e^x, \quad$ lorsque $x \to +\infty$.

Pour tout $\alpha \geq 0$, on a $\quad x^\alpha \gg -\ln(x), \quad$ lorsque $x \to 0^+$.

Proposition 0.4 Inegalité de Hölder

Soit $X \in \mathbb{R}^d$ et soit $p, q \in [1, +\infty]$ tels que $\dfrac{1}{p} + \dfrac{1}{q} = 1$. On a

$$\left|\sum_{k=1}^{d} x_k\right| \;\leq\; \sqrt[p]{\sum_{k=1}^{d} |x_k|^p}\ \sqrt[q]{\sum_{k=1}^{d} |x_k|^q}\,. \tag{13}$$

Lorsque $p = q = 2$ on retrouve l'inégalité de Cauchy-Schwarz. Cette inégalité se généralise aux fonctions et au calcul d'intégrales grâce aux sommes de Riemann.

Proposition 0.5 Formule de Taylor d'ordre 2 en dimension supérieure

Soit $f : \Omega \to \mathbb{R}$ de classe $\mathcal{C}^2$ avec Ω un ouvert de $\mathbb{R}^d$. Au voisinage d'un point $X \in \Omega$ nous avons, pour tout H tel que $X + H \in \Omega$:

$$f(X+H) \;=\; f(X) \;+\; \nabla f(X)^T H \;+\; \frac{1}{2} H^T \nabla^2 f(X) H \;+\; o\big(|H|^2\big). \tag{14}$$

Il est important de connaître également les développements de Taylor d'ordre supérieurs pour les fonctions d'une seule variable.

Théorème 0.6 Théorème valeurs extrêmes de Weierstrass

Une fonction continue sur un compact est bornée et atteint ses bornes.

PROPOSITION 0.6 Différentiation des fonctions composées

Soit $\mathcal{G} : \Omega \to \mathbb{R}^p$ et $\mathcal{F} : \Gamma \to \mathbb{R}^q$ de classes $\mathcal{C}^1$ avec Ω ouvert de $\mathbb{R}^d$ et Γ ouvert de $\mathbb{R}^p$. Soit $X \in \Omega$ tel que $\mathcal{G}(X) \in \Gamma$. Alors la fonction composée $\mathcal{F} \circ \mathcal{G}$ est bien définie et $\mathcal{C}^1$ au voisinage de X et

$$\nabla\big(\mathcal{F} \circ \mathcal{G}\big)(X) \;=\; \nabla\mathcal{F}\big(\mathcal{G}(X)\big)^T \nabla\mathcal{G}(X). \tag{15}$$

PROPOSITION 0.7 Formule du développement limité de Taylor avec reste intégral

Soit $f : I \to \mathbb{R}$ de classe $\mathcal{C}^{n+1}$ avec I intervalle ouvert de $\mathbb{R}$. Soit $x \in I$ et $h \in \mathbb{R}$ "*assez petit*". On a :

$$f(x+h) \;=\; \sum_{k=0}^{n} \frac{\mathrm{d}^k f}{\mathrm{d}x^k}(x)\, \frac{h^k}{k!} \;+\; \int_x^{x+h} \frac{\mathrm{d}^{n+1} f}{\mathrm{d}x^{n+1}}(y) \frac{(x-y)^n}{n!}\, \mathrm{d}y \tag{16}$$

0.5 Rappels et compléments d'algèbre (†)

Dans le cadre de l'analyse des équations différentielles et des équations aux dérivées partielles, nous faisons appels à de nombreux outils issus d'algèbre linéaire et bilinéaire. Il est donc important de bien les connaître dans le cadre de ce cours et plus généralement tout au long de la formation d'ingénieur. Il est ainsi impératif d'être au point sur :

- Matrices carrées, rectangles, triangulaires, diagonales, tri-diagonales.
- Transposée d'une matrice, conjuguée, adjointe, inverse.
- Matrices équivalentes et matrices semblables.
- Polynômes annulateurs, polynôme caractéristique.
- Algorithme du pivot de Gauss, trigonalisation, diagonalisation.
- Le groupe linéaire et le groupe spécial linéaire.
- Matrices symétriques et anti-symétriques.
- Matrices orthogonales, unitaires, nilpotentes.
- Algorithme de Gram-Schmidt.

0.5.1 Formules de changement de bases

On considère un morphisme linéaire de $\mathbb{R}^d$ dans $\mathbb{R}^p$ qui se représente dans les bases canoniques par la matrice $A \in \mathcal{M}_{p,d}(\mathbb{R})$. On souhaite représenter l'endomorphisme dans une nouvelle base de départ $(X_1, \ldots, X_d)$ pour $\mathbb{R}^d$ et dans une nouvelle base d'arrivée $(Y_1, \ldots, Y_p)$ pour $\mathbb{R}^p$. On introduit les matrices de passages, qui sont la concaténation des coordonnées des vecteurs de la nouvelle base, écrits sous forme colonne :

$$P \;:=\; \Big[X_1\Big|X_2\Big|\cdots\Big|X_d\Big] \in \mathcal{M}_d(\mathbb{R}), \qquad \text{et} \qquad Q \;:=\; \Big[Y_1\Big|Y_2\Big|\cdots\Big|Y_p\Big] \in \mathcal{M}_p(\mathbb{R})$$

On représente le vecteur de départ $X \in \mathbb{R}^d$ dans la nouvelle base :

$$X \;=\; \sum_{i=1}^{d} u_i X_i \;=\; PU,$$

où le vecteur colonne $U := [u_1, \ldots, u_d]^T$ contient les nouvelles coordonnées de X dans la nouvelle base de l'espace de départ. De même, on représente les vecteurs de l'espace d'arrivée dans leur

nouvelle base :

$$Y \;=\; \sum_{j=1}^{p} v_j Y_j \;=\; QV,$$

où $v = [v_1, \ldots, v_p]^T$ est le vecteur colonne qui contient les nouvelles coordonnées de Y dans la nouvelle base de l'espace d'arrivée. On en déduit alors la matrice qui représente l'endomorphisme dans les nouvelles bases par :

$$Y = AX \qquad \Longleftrightarrow \qquad QV = APU$$

Dans les nouvelles bases, l'endomorphisme est donc représenté par la matrice $Q^{-1}AP$.

0.5.2 Réduction de Jordan et exponentielle de matrice

Dans le cadre de ce cours, nous utiliserons la réduction de Jordan pour les endomorphismes.

Définition 0.18 Bloc de Jordan

Pour $k \in \mathbb{N}^*$ et $\lambda \in \mathbb{C}$, on définit le *bloc de Jordan* $J_k(\lambda) \in \mathcal{M}_k(\mathbb{C})$ comme étant la matrice carrée de taille k suivante :

$$J_k(\lambda) = \begin{pmatrix} \lambda & 1 & & & \\ & \lambda & 1 & & (0) \\ & & \ddots & \ddots & \\ & & & \ddots & \ddots \\ & (0) & & & \lambda \;\; 1 \\ & & & & \lambda \end{pmatrix}. \tag{17}$$

Cette matrice n'est diagonalisable que si $k = 1$ et elle est nilpotente ssi $\lambda = 0$. Dans le cas nilpotent, le calcul des puissances de la matrice $J_k(0)$ est donné par la formule de la *diagonale décalée*. En effet, la diagonale de 1 est au centre lorsque l'exposant est nul, puis se décale d'un cran vers le coin supérieur droit à chaque fois que l'exposant augmente (jusqu'à ce que la matrice soit nulle) :

$$J_k(0) = \begin{pmatrix} 0 & 1 & & & \\ & 0 & 1 & & (0) \\ & & \ddots & \ddots & \\ & & & \ddots & \ddots \\ & (0) & & & 0 \;\; 1 \\ & & & & 0 \end{pmatrix}, \qquad J_k(0)^2 = \begin{pmatrix} 0 & 0 & 1 & & & (0) \\ & 0 & 0 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & \ddots & \ddots & 1 \\ & (0) & & & 0 & 0 \\ & & & & & 0 \end{pmatrix} \qquad \text{etc...} \tag{18}$$

$$J_k(0)^{k-1} = \begin{pmatrix} 0 & & \cdots & \cdots & 0 & 1 \\ & 0 & & & & 0 \\ & & \ddots & & & \vdots \\ & & & \ddots & & \vdots \\ & (0) & & & 0 & \\ & & & & & 0 \end{pmatrix}, \qquad \text{et} \qquad J_k(0)^k = 0. \tag{19}$$

Le calcul des puissances pour les autres blocs de Jordan s'obtient grâce à la formule du binôme de Newton :

$$J_k(\lambda)^n \;=\; \Big(\lambda I_k + J_k(0)\Big)^n \;=\; \sum_{i=0}^{n} \binom{n}{i} \lambda^i J_k(0)^{n-i}. \tag{20}$$

Le principal intérêt des matrices de Jordan réside dans le Théorème de Jordan qui affirme que toutes les matrices sont diagonalisables par blocs de Jordan :

THÉORÈME 0.7 Réduction de Jordan (†)

Toute matrice carrée $M \in \mathcal{M}_d(\mathbb{C})$ est semblable à une matrice diagonale par blocs et dont les blocs sont de Jordan :

$$M \sim \begin{pmatrix} J_{k_1}(\lambda_1) & & & \\ & J_{k_2}(\lambda_2) & & (0) \\ & & \ddots & \\ & (0) & & \ddots \\ & & & & J_{k_r}(\lambda_r) \end{pmatrix} \tag{21}$$

C'est-à-dire que, tout endomorphisme de $\mathbb{R}^d$ se laisse représenter dans une certaine base par une matrice de Jordan (diagonale par blocs de Jordan). Un tel endomorphisme est diagonalisable ssi ses blocs de Jordan sont tous de taille 1. L'obtention de la réduction de Jordan se fait par la même approche que les méthodes de diagonalisation avec quelques ajouts. Plus précisément, l'agorithme est le suivant (†) :

1. Calculer le polynôme caractéristique de la matrice M et trouver ses racines (qui sont les valeurs propres de M).

2. Pour chaque valeur propre λ, déterminer l'espace propre associé $E_\lambda := \ker(M - \lambda I_n)$.

3. Pour construire les blocs de Jordan, on insère d'abord dans notre nouvelle base un vecteur propre de la matrice. ie : on prend $v_1 \in E_\lambda$ tel que $Mv_1 = \lambda v_1$.

4. On cherche alors un vecteur non-nul $v_2 \in \mathbb{C}^d$ tel que $(A - \lambda I_n)v_2 = v_1$ et on le rajoute dans notre base de $\mathbb{C}^d$.

5. On procède itérativement on cherchant $(A - \lambda I_n)v_k = v_{k-1}$ jusqu'à ce que ce processus itératif ne donne plus de solutions.

6. Par construction, L'espace vectoriel engendré par les vecteurs $(v_1, \ldots, v_k)$ est stable par M. Dans cette base, l'endomorphisme canoniquement associé à M est représenté par le bloc de Jordan $J_k(\lambda)$.

7. Après avoir obtenu un bloc de Jordan, on revient à l'étape 2 et on recommence jusqu'à avoir obtenu tous les blocs de Jordan de la matrice M et la base associée.

L'un des principaux intérêts de la réduction de Jordan est qu'elle facilite grandement le calcul des puissances de cette matrice car il suffit de calculer les puissances des blocs de Jordan. En particulier, cela permet de calculer facilement les exponentielles de matrices :

Définition 0.19 Exponentielle de matrice (†)

Pour une matrice carrée $M \in \mathcal{M}_d(\mathbb{C})$ on définit l'exponentielle de la matrice M par

$$\exp(M) := \sum_{k=0}^{+\infty} \frac{M^k}{k!}.$$

Par ailleurs, si M est un bloc de Jordan $J(\lambda)$ alors pour tout $t \in \mathbb{R}$,

$$\exp\big(tJ(\lambda)\big) = \exp(t\lambda) \begin{pmatrix} 1 & t & \frac{t^2}{2} & \cdots & \frac{t^{p-1}}{(p-1)!} \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & t & \frac{t^2}{2} \\ \vdots & 0 & \ddots & \ddots & t \\ 0 & \cdots & \cdots & 0 & 1 \end{pmatrix}. \tag{22}$$

Ainsi, pour calculer l'exponentielle d'une matrice, on peut commencer par calculer sa réduction de Jordan et ensuite appliquer l'exponentielle à la réduite de Jordan (dans le cas diagonal, l'exponentielle s'applique bloc par bloc). Ces matrices exponentielles vont jouer un rôle important dans l'analyse des équations différentielles.

Remarque importante : puisque nous travaillons avec des matrices à coefficients réels, lorsque l'on a une valeur propre λ avec une partie imaginaire non-nulle alors $\overline{\lambda}$ est aussi une valeur propre. Ces deux valeurs propres complexes conjuguées auront les mêmes blocs de Jordan (il est donc important de toujours les étudier les espaces propres associés à λ et $\overline{\lambda}$ ensemble).

0.5.3 Rappels et compléments d'algèbre bilinéaire

Définition 0.20 Formes quadratiques (†)

Une application $\mathrm{Q} : \mathbb{R}^d \to \mathbb{R}$ est une *forme quadratique* ssi il existe une matrice A symétrique telle que

$$\mathrm{Q}(X) \;=\; X^T A X.$$

La *forme polaire* associée est la *forme bilinéaire* suivante :

$$\Phi \;:\; (X, Y) \;\longmapsto\; X^T A Y.$$

Il existe une unique matrice symétrique A qui représente la forme bilinéaire Φ. Le produit scalaire euclidien canonique est la forme bilinéaire représentée par la matrice $A = I_d$.

Proposition 0.8 Diagonalisation des matrices symétriques (†)

Les matrices symétriques sont *ortho-diagonalisables*. C'est-à-dire que pour A une matrice symétrique, il existe une matrice orthogonale O (par définition elle vérifie $O^T = O^{-1}$) et une matrice diagonale D telles que :

$$A \;=\; O^T D O.$$

En particulier, si on considère le changement de base $Y = OX$, alors toute forme quadratique se laisse représenter par une matrice diagonale si on travaille dans une base adaptée. En effet :

$$\Phi(X) \;=\; X^T A X \;=\; X^T O^T D O X \;=\; (OX)^T D (OX) \;=\; Y^T D Y,$$

où l'on a utilisé la propriété de contravariance de la transposition : $(AB)^T = B^T A^T$.

Dans le cadre de ce cours, nous allons exploiter les propriétés des formes quadratiques qui sont en lien avec l'analyse des fonctions de $\mathbb{R}^d$ dans $\mathbb{R}$. Pour cela nous introduisons la notion de *convexité* et de *coercivité*.

Définition 0.21 Fonctions convexes (†)

Soit $f : \Omega \to \mathbb{R}$ avec Ω ouvert de $\mathbb{R}^d$ de classe $\mathcal{C}^1$. On dit que la fonction f est *convexe* au point $X \in \Omega$ si il existe un rayon $r_X > 0$ tel que pour tout $Y \in \Omega \cap \mathcal{B}(X, r_X)$:

$$f(Y) \ \geq \ f(X) + \nabla f(X)^T (Y - X). \tag{23}$$

Si cette inégalité est stricte, on dit que la fonction est *strictement convexe* au point X. On dit que f est *fortement convexe* au point X s'il existe un $\alpha > 0$ tel que

$$f(Y) \ \geq \ f(X) + \nabla f(X)^T (Y - X) + \frac{\alpha}{2}|Y - X|^2. \tag{24}$$

Si le paramètre α et le rayon r peuvent être choisis indépendamment du point X, on parle de convexité forte uniforme.

D'un point de vue intuitif, ce nombre α vient mesurer la courbure locale du graphe de la fonction. La forte convexité implique la stricte convexité, qui implique la convexité.

Remarque : il est possible de définir la convexité en $X \in \Omega$ pour des fonctions moins régulières que $\mathcal{C}^1$ à l'aide de la caractérisation suivante :

$$\forall\, \lambda \in [0,1], \qquad f\Big(\lambda X + (1-\lambda)Y\Big) \ \leq \ \lambda f\big(X\big) + (1-\lambda) f\big(Y\big). \tag{25}$$

Dans le cadre de ce cours, on utilisera uniquement la définition 0.21.

Pour étudier les puits de potentiels, notion très importante pour décrire de nombreux systèmes physiques, on a souvent recours à la notion de *coercivité*. [1]

Définition 0.22 Fonctions coercives (†)

Soit $f : \Omega \to \mathbb{R}$ avec Ω ouvert de $\mathbb{R}^d$.

(*i*) Elle est *globalement coercive* (on dit parfois simplement "*coercive*") ssi ses ensembles de niveaux sont tous connexes et compacts.

(*ii*) Une fonction f est *coercive à l'infini* ssi ses ensembles de sous-niveaux $\{X \in \Omega : f(X) \leq \mu\}$ sont compacts pour tout $\mu \in \mathbb{R}$.

(*iii*) Elle est *coercive* en un point $X_0 \in \Omega$ si X_0 appartient à une composante connexe compacte de l'ensemble de sous-niveau $\{X \in \Omega : f(X) \leq f(X_0)\}$.

En plus des ensembles de *sous-niveaux*, on se sert également beaucoup des ensembles de niveaux de f comme étant $\{X \in \Omega : f(X) = \mu\}$ pour tout $\mu \in \mathbb{R}$ un paramètre, ainsi que les ensembles de sur-niveaux $\{X \in \Omega : f(X) \geq \mu\}$.

1. Cette notion est très utilisée en dimension infinie pour les EDP mais il est moins standard de l'utiliser en dimension finie. Pensez donc à redonner les définitions si vous utiliser ce terme dans un autre cadre que celui de ce cours.

PROPOSITION 0.9 Propriétés de la coercivité

Soit $f : \Omega \to \mathbb{R}$ avec Ω ouvert de $\mathbb{R}^d$.

- Si f est globalement coercive alors elle est coercive à l'infini.
- Si f est coercive à l'infini alors elle est coercive en tout point.
- Si $\Omega = \mathbb{R}^d$ alors f est coercive à l'infini si et seulement si

$$f(X) \longrightarrow +\infty \qquad \text{lorsque} \quad |X| \to +\infty. \tag{26}$$

- Si Ω est borné, alors f est coercive à l'infini si et seulement si

$$f(X) \longrightarrow +\infty \qquad \text{lorsque} \quad \text{dist}(X, \partial\Omega) \to 0, \tag{27}$$

  avec $\text{dist}(X, A) := \inf_{Y \in A} |X - Y|$.
- Si f est continue et coercive en un point alors elle admet un minimum local.
- Si f est continue coercive à l'infini alors elle admet un minimum global.
- Si f est coercive à l'infini et convexe alors elle est globalement coercive.
- Si f est globalement coercive et si elle admet un minimum local strict alors ce point est son unique minimum global.

L'image à avoir en tête pour les fonctions coercives à l'infini est celle d'un système de puits de potentiels avec un confinement infini. Si la fonction est globalement coercive, alors il n'y a qu'un seul puits de potentiel.

Exemples de fonctions coercives sur $\mathbb{R}$:

- $x \mapsto x^4$, • $x \mapsto \sqrt{|x|}$, $x \mapsto |x + \sin(x)|$, $x \mapsto \max\{x - 1; -x - 1; 0\}$.

Exemples de fonctions coercives à l'infini mais pas coercives sur $\mathbb{R}$:

- $x \mapsto x^4 - x^2$, • $x \mapsto x^2\Big(1 + \cos^2(x)\Big)$ • $x \mapsto \arccos(\cos(x)) + \dfrac{1}{2}|x|$.

Exemples de fonctions coercives en 0 mais pas à l'infini :

- $x \mapsto \sin(x)$, • $x \mapsto x^3 - x$ • $x \mapsto \dfrac{x^2}{1 + x^2}$.

Dans le cas des formes quadratiques, la coercivité est une propriété qui devient très contraignante. Plus précisément, nous avons la caractérisation suivante :

PROPOSITION 0.10 Caractérisation des formes quadratiques coercives

Soit $A \in \mathcal{M}_d(\mathbb{R})$ une matrice symétrique et soit $\mathrm{Q} : X \in \mathbb{R}^d \mapsto X^T A X$ la forme quadratique associée. Les propositions suivantes sont toutes équivalentes :

- Q est coercive.
- Q est coercive en un point.
- Q est fortement convexe.
- Q est strictement convexe en 0.
- Les ensembles de niveaux de Q sont des ellipses.
- Les valeurs propres de A sont toutes strictement positives.
- Pour tout $X \in \mathbb{R}^d$ non nul, $Q(X) > 0$.
- Il existe $\lambda_1 > 0$ tel que pour tout $X \in \mathbb{R}^d$ on a $Q(X) \geq \lambda_1 |X|^2$.

Dans le cas où la forme quadratique est coercive, on dit que la matrice A associée est *defini-positive*. La forme bilinéaire associée à une forme quadratique coercive est appelé *produit scalaire*. Les formes quadratiques coercives sont très pratiques pour démontrer qu'un point critique d'une fonction $f \in \mathcal{C}^2$ est un minimum local strict, à l'aide de la matrice hessienne de f et de la dernière propriété de la liste ci-dessus [2].

Théorème 0.8 Théorème de la forte convexité (†)

Soit $f : \mathbb{R}^d \to \mathbb{R}$ de classe $\mathcal{C}^2$.

(i) La fonction f est fortement convexe en $X \in \mathbb{R}^d$ si et seulement si $\nabla^2 f(X)$ est une matrice symétrique défini-positive.

(ii) Si de plus X est un point-critique, c'est-à-dire si $\nabla f(X) = 0$, alors il s'agit d'un minimum local strict.

Remarque 1 **:** La plus grande valeur possible pour le choix de λ_1 dans la proposition 0.8 est égale à la plus petite valeur propre de la matrice A.

Remarque 2 **:** Le paramètre de forte convexité $\alpha/2$ (voir Définition 0.21) pour la fonction quadratique Q peut être choisi librement dans l'intervalle $]0; \lambda_1[$.

Remarque 3 **:** Ce théorème est très utile pour démontrer qu'un point critique est un minimum local car cela permet de transformer un problème d'analyse en une étude algébrique de la matrice hessienne au point considéré. Il est important de noter que la réciproque du point (ii) est fausse ; dans certains cas il faut effectuer une étude plus approfondie pour démontrer qu'un point critique est minimum local. Plus généralement, on peut caractériser les points-critiques à l'aide de la hessienne comme suit :

- Les valeurs propres sont strictement positives $\Rightarrow$ minimum local strict.
- Les valeurs propres sont strictement négatives $\Rightarrow$ maximum local strict.
- Les valeurs propres sont non-nulles avec des signes différents $\Rightarrow$ point col.
- Une valeur propre est nulle : la matrice hessienne est dégénérée (il faut alors monter à l'ordre supérieur pour décrire le comportement local de la fonction).

2. Cette dernière propriété est tellement importante dans la description des formes quadratiques coercives et des minima locaux des fonctions que beaucoup d'auteurs la choisissent comme définition de la coercivité. Cette propriété est celle qui se généralise le plus naturellement à la dimension infinie.

0.6 Bilan du chapitre

0.6.1 Ce qu'il faut retenir et savoir-faire

Ce chapitre est surtout consacré à des rappels de base + quelques compléments. Ce sont des concepts et des outils indispensables pour l'analyse des équations différentielles ; et plus généralement pour la compréhension des mathématiques enseignées aux ingénieurs. Ce chapitre doit donc être compris et maîtrisé dans son ensemble. Les éléments les plus saillants à retenir sont les suivants :

(†) Espace vectoriels normés et topologie, fonctions $\mathcal{C}^0$, $\mathcal{C}^1$.
(†) Les fonctions lipschitziennes : retenir $\mathcal{C}^1 \Rightarrow$ lipschitz $\Rightarrow \mathcal{C}^0$.
(†) Comparaison asymptotique des fonctions.
(†) Analyse fonctionnelle élémentaire.
(†) Algèbre linéaire et bilinéaire.
(†) Algorithme de réduction de Jordan et exponentielle de matrice
(†) Matrices symétriques défini-positives et théorème de la forte convexité.

0.6.2 Exercices

Les exercices du chapitre 0 ***ne seront pas traités en TD*** mais il est conseillé de les faire afin de s'assurer de bien maîtriser les concepts de base de l'analyse et de l'algèbre. S'il y a le moindre blocage sur ces exercices fondamentaux, ne pas hésiter à se rapprocher des chargés de TD.

Exercice 0.1. Les fonctions suivantes se prolongent-elles par continuité en $x = 0$? le prolongement est-il $\mathcal{C}^1$? $\mathcal{C}^k$? $\mathcal{C}^\infty$?

$$x \mapsto \frac{\sin(x)}{x}, \qquad x \mapsto x^2 \sin\left(\frac{1}{x}\right), \qquad x \mapsto \frac{\tan^2(x)\sinh(x)}{|x|}, \qquad x \mapsto \exp\left(-\frac{1}{x^2}\right).$$

Exercice 0.2. Ces fonctions se prolongent-elles par continuité en $(x, y) = (0, 0)$? Le prolongement est-il $\mathcal{C}^1$?

$$(x,y) \mapsto \frac{xy}{x^2+y^2}, \qquad (x,y) \mapsto \frac{x^2y^2}{\sqrt{x^2+y^2}^3}, \qquad (x,y) \mapsto \frac{xy(x^2-y^2)}{x^2+y^2}.$$

Exercice 0.3. Montrer que : $e^x\left(2 - \frac{1}{x} - \sin(x)\right) \simeq e^x$ lorsque $x \to +\infty$.

Exercice 0.4. Montrer que, par croissance comparée :

$$e^{-\frac{1}{x}} \ll x^2 \ll x \ll x\ln(x) \ll \sqrt{x} \ll \frac{1}{\ln(x)}, \quad \text{lorsque } x \to 0^+.$$

Exercice 0.5. Montrer que $\displaystyle\sum_{k=1}^{n} \frac{k^3}{k^2+n^2} \sin^2\left(\frac{k\pi}{8} + \frac{1}{n}\right) \simeq n^2$ lorsque $n \to +\infty$.

Exercice 0.6. Trouver l'asymptote polynomiale en $+\infty$ de $x \longmapsto \dfrac{x^3}{x+1}$.

Exercice 0.7. Montrer que les fonctions suivantes sont globalement lipschitz (sur $\mathbb{R}$) :

$$x \longmapsto 2x+1, \qquad x \longmapsto |x|, \qquad x \longmapsto \tanh(x), \qquad x \longmapsto \max\{x, 0\}, \qquad x \longmapsto \sum_{n=0}^{+\infty} \frac{|\sin(nx)|}{n^3}.$$

Montrer que les fonctions suivantes sont localement mais pas globalement lipschitz :

$$x \longmapsto x^2, \qquad x \longmapsto e^x, \qquad x \longmapsto \sin(x^2), \qquad x \longmapsto |x^3 - x|, \qquad x \longmapsto x^2 \cos(1/x).$$

Montrer que les fonctions suivantes sont continues mais pas localement lipschitz :

$$x \longmapsto x \ln(|x|), \qquad x \longmapsto \sqrt[3]{x}, \qquad x \longmapsto x \sin(1/x), \qquad x \longmapsto \sqrt{|\cos(x)|}.$$

Exercice 0.8. Faire la réduction de Jordan des matrices suivantes :

$$\begin{pmatrix} 4 & 3 & -2 \\ -3 & -1 & 3 \\ 2 & 3 & 0 \end{pmatrix}, \qquad \begin{pmatrix} 4 & 0 & -1 \\ -1 & 1 & 3 \\ 0 & -1 & 4 \end{pmatrix} \qquad \text{et} \qquad \begin{pmatrix} 5 & 0 & 4 & -2 & -3 \\ -2 & 3 & -3 & 2 & 4 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 3 & 1 \\ 1 & 0 & 2 & -1 & 1 \end{pmatrix}.$$

Exercice 0.9. Dans la Définition 0.19 : démontrer que l'exponentielle de matrice est bien définie (la série converge) et démontrer la formule pour les exponentielles de blocs de Jordan (avec le paramètre $t \in \mathbb{R}$).

Exercice 0.10. Trouver et décrire les points critiques des fonctions suivantes (est-ce un minimum local ? un maximum local ? un point col ?) :

$$f_1(x,y) = x^2+xy+y^2, \qquad f_2(x,y) = x^3+y^2, \qquad f_3(x,y) = x^4+y^4, \qquad f_4(x,y) = \cos(x)\cos(y).$$

$$f_5(x,y) = x^2 + 4xy + y^2, \qquad f_6(x,y) = x^2 + xy^2, \qquad f_7(x,y) = x^4 - xy + y^4.$$

Parmi ces fonctions, lesquelles sont coercives à l'infini ?

* * *

Chapitre 1

Existence et unicité des solutions

Dans ce chapitre nous présentons la théorie générale des équations différentielles ordinaires (EDO) permettant d'étudier le caractère bien posé de ces équations. Nous présentons les principaux théorèmes permettant d'établir l'existence et l'unicité locale (sur un petit intervalle de temps $[0, T_0[$) ou bien globale (sur $[0, +\infty[$ ou sur $\mathbb{R}$ tout entier) des solutions d'une équation munie d'une donnée initiale à $t = 0$. On évoque également quelques principes de comparaison pour étudier le comportement des solutions au voisinage des temps pour lesquels existence ou unicité cessent d'être vérifiés.

Le chapitre se termine par un recueil d'exercices permettant de s'entraîner à l'étude des équations différentielles en appliquant les théorèmes fondamentaux de ce chapitre à des équations issus de modèles physiques ou biologiques.

1.1 Existence et unicité locale d'une solution

On appelle *équation différentielle* un équation dont l'inconnue est une fonction d'une seule variable réelle (assimilée à l'écoulement temps) et à valeurs vectorielles $X : [0, T[\to \mathbb{R}^d$. Dans le cadre de ce cours, nous allons nous concentrer sur l'étude des *équations différentielles ordinaires*, c'est-à-dire les équations qui se mettent sous la forme dite *de Cauchy*. Pour cela on considère une fonction

$$\begin{array}{rccc} \mathcal{F} \ : & [0,T[\times\Omega & \longrightarrow & \mathbb{R}^d \\ & (t,X) & \longmapsto & \mathcal{F}(t,X), \end{array} \tag{1.1}$$

avec Ω ouvert non-vide de $\mathbb{R}^d$ et $T \in \mathbb{R}_+^* \cup \{+\infty\}$. On considère également un $X_0 \in \Omega$.

Définition 1.1 Problème de Cauchy (‡)

On appelle *équation différentielle sous forme de Cauchy* toute équation différentielle d'inconnue $X : [0, T) \to \Omega$ sous la forme suivante :

$$\frac{dX}{\mathrm{d}t}(t) \ = \ \mathcal{F}\Big(t, X(t)\Big). \tag{1.2}$$

On appelle *problème de Cauchy* le problème consistant à trouver les solutions d'une équation différentielle sous forme de Cauchy étant donné la position au temps initial. Autrement dit, pour $\mathcal{F} : [0, T[\times\Omega \to \mathbb{R}^d$ et $X_0 \in \Omega$ fixés, on cherche les fonctions $X : [0, T) \to \Omega$ telles que

$$\frac{dX}{\mathrm{d}t}(t) \ = \ \mathcal{F}\Big(t, X(t)\Big), \qquad \text{et} \qquad X(0) = X_0 \in \Omega. \tag{1.3}$$

Dans ce qui suit, on dira que $X(t)$ est la *position* du système au temps t et que $dX/\mathrm{d}t$, également noté $\dot{X}$ est son *vecteur-vitesse.* La fonction $\mathcal{F}$ est appelé le *champ de vitesse* associé à l'équation. Dans le cas où $\mathcal{F}$ ne dépend pas de la variable temporelle, on dit que l'équation est *autonome.*

Très important : Il ne faut SURTOUT PAS confondre la fonction inconnue $t \mapsto X(t)$ avec la variable $X \in \Omega$ qui apparaît dans la définition de la fonction $\mathcal{F}$ à l'équation (1.1). En effet, nous allons étudier les propriété du champs de vitesses $\mathcal{F}$ indépendamment de l'équation différentielle qu'il engendre (1.2). Malgré cet abus de notations très pratique, il faut garder à l'esprit que ce sont 2 choses différentes. Le terme $\mathcal{F}\big(t, X(t)\big)$ doit donc se comprendre au sens des fonctions composées $\mathcal{F}(t,\cdot) \circ X(t)$.

Étant donné une équation différentielle munie d'une donnée initiale, il est naturel de penser qu'il existe une unique solution au problème de Cauchy. En effet, il suffit pour le point $X(t)$ de suivre le champ de vitesse $\mathcal{F}$ dès lors que celui-ci n'est pas trop irrégulier. Cette intuition a été formalisée dans le théorème d'existence et d'unicité suivant :

Théorème 1.1 Théorème de Cauchy-Lipschitz (‡)

On considère, pour $\mathcal{F} : [0,T[\times\Omega \to \mathbb{R}^d$ et $X_0 \in \Omega$ fixés le problème de Cauchy (1.3).

Si le champ de vitesse $\mathcal{F}$ est une fonction *continue* et *localement lipschitzienne pour la variable* X, alors il existe $T_0 \in]0,T]$ tel que le problème de Cauchy (1.3) admet une *unique* solution $X : [0,T_0[\to \Omega$.

Quand on a existence et unicité de la solution, on dit alors que le problème est *bien posé.* Le plus grand temps T_0 tel que le théorème 1.1-(i) soit vérifié s'appelle le *temps maximal d'existence* et on le note $T^\star$. Si on démontre que $T^\star$ vaut $+\infty$, on parle alors de *solution globale.* Sinon on parle seulement de *solution locale.*

La démonstration rigoureuse de ce théorème a nécessité plus d'un siècle (le XIXe siècle) avec des contributions de mathématiciens parmi les meilleurs de leur époque. Parmi-eux on retient souvent les noms principaux, à savoir Augustin-Louis Cauchy, Rudolf Lipschitz, Emile Picard, Stephan Banach, et Ernst Lindelöf. Cette théorème est célébré à la fois pour son caractère fondamental dans l'étude des équations différentielles, mais également de par l'étonnant détour théorique nécessaire pour obtenir sa démonstration. On utilise notamment le théorème (abstrait) d'analyse fonctionnelle suivant :

Théorème 1.2 Point-fixe de Picard-Banach

On considère l'espace vectoriel $E = \mathcal{C}^0\big([0,T);\Omega\big)$, avec Ω ouvert de $\mathbb{R}^d$, et on munit l'espace E de la norme L^∞. On considère une fonctionnelle $\mathbf{G} : E \to E$ et on suppose qu'elle est *contractante* :

$$\exists\, \delta < 1, \quad \forall\, \phi, \psi \in E, \qquad \big\|\mathbf{G}(\phi) - \mathbf{G}(\psi)\big\|_{L^\infty} \;\leq\; \delta\, \|\phi - \psi\|_{L^\infty}. \tag{1.4}$$

Dans ce cas, la fonctionnelle $\mathbf{G}$ admet un unique point-fixe :

$$\exists!\, \phi^\star \in E, \quad \phi^\star = \mathbf{G}(\phi^\star).$$

Les outils théoriques permettant de démontrer le théorème de Picard-Banach seront présentés lors du prochain semestre dans le cadre du cours consacré au calcul d'intégrales. Pour

l'instant nous admettons ce résultat qui va nous permettre de démontrer le théorème de Cauchy-Lipschitz.

Démonstration du théorème 1.1. *Etape 1 : Formule de Duhamel.* Si on considère l'équation différentielle et qu'on l'intègre au cours du temps entre 0 et $t \geq 0$, on obtient :

$$X(t) - X(0) = \int_0^t \mathcal{F}\Big(s, X(s)\Big)\,\mathrm{d}s. \tag{1.5}$$

On remplace à présent $X(0)$ par la donnée initiale du problème de Cauchy $X_0 \in \Omega$ et on aboutit à la formule de Duhamel :

$$X(t) = X_0 + \int_0^t \mathcal{F}\Big(s, X(s)\Big)\,\mathrm{d}s. \tag{1.6}$$

Réciproquement, une fonction continue $X : [0, T[\to \mathbb{R}^d$ qui vérifie la formule de Duhamel est également solution du problème de Cauchy associé à $\mathcal{F}$ avec la donnée initiale $X(0) = X_0$. En effet, il est immédiat de vérifier que si on prend $t = 0$ dans (1.6) alors $X(0) = X_0$. De plus, puisque $s \mapsto X(s)$ est continue, on a donc $s \mapsto \mathcal{F}\big(s, X(s)\big)$ continue comme composée de fonctions continues. Le théorème fondamental de l'analyse nous permets alors de conclure que

$$t \mapsto \int_0^t \mathcal{F}\Big(s, X(s)\Big)\,\mathrm{d}s \quad \text{est } \mathcal{C}^1, \quad \text{et sa dérivée vaut } t \mapsto \mathcal{F}\Big(t, X(t)\Big). \tag{1.7}$$

On peut ainsi dériver (1.6) et retrouver l'équation initiale. Ceci établit l'équivalence entre le problème de Cauchy et la formule de Duhamel.

Etape 2 : Construction de la fonctionnelle $\mathbf{G}$. Il n'est pas possible de travailler directement sur Ω tout entier car on va avoir besoin d'un argument nécessitant un compact. On travaille alors dans le compact $K := \overline{\mathcal{B}}(X_0, R)$ où le rayon $R > 0$ est choisi assez petit pour que $K \subset \Omega$ (un tel R existe car Ω un ouvert). Soit $T_0 > 0$ un paramètre (que l'on ajustera plus tard). On pose $E = \mathcal{C}^0([0, T_0]; K)$ et on définit la fonctionnelle $\mathbf{G}$ par

$$\forall\, X \in E, \qquad \mathbf{G}(X)(t) := X_0 + \int_0^t \mathcal{F}\Big(s, X(s)\Big)\,\mathrm{d}s \tag{1.8}$$

Puisque X et $\mathcal{F}$ sont continues, on en déduit que la fonction $\mathbf{G}(X)$ est continue à valeurs dans $\mathbb{R}^d$. Autrement dit, $\mathbf{G}$ est une fonction qui va de $E = \mathcal{C}^0([0, T_0]; K)$ dans $\mathcal{C}^0([0, T_0]; \mathbb{R}^d)$.

A présent, pour $X, Y \in E$ fixés on a (par l'inégalité triangulaire) :

$$\forall\, t \in [0, T_0], \qquad \Big|\mathbf{G}(X)(t) - \mathbf{G}(Y)(t)\Big| \;\leq\; \int_0^t \Big|\mathcal{F}\Big(s, X(s)\Big) - \mathcal{F}\Big(s, Y(s)\Big)\Big|\,\mathrm{d}s. \tag{1.9}$$

On rappelle alors que $\mathcal{F}$ par hypothèse est une fonction localement lipschitzienne pour la variable X. On pose alors

$$\lambda \;:=\; \sup_{t\in[0,T]} \sup_{\substack{Y,Z\in K\\ Y\neq Z}} \frac{|\mathcal{F}(t,X) - \mathcal{F}(t,Y)|}{|X - Y|} \quad < +\infty$$

Ce nombre λ est fini par définition des fonctions partiellement lipschitziennes et théorème des accroissements finis. On a donc $\big|\mathcal{F}\big(s, X(s)\big) - \mathcal{F}\big(s, Y(s)\big)\big|$ majoré par $\lambda|X(s) - Y(s)|$ pour tout $s \in [0, T_0]$. Ainsi (1.9) devient :

$$\forall\, t \in [0, T_0], \qquad \Big|\mathbf{G}(X)(t) - \mathbf{G}(Y)(t)\Big| \;\leq\; \lambda \int_0^t \big|X(s) - Y(s)\big|\,\mathrm{d}s. \tag{1.10}$$

En prenant à présent le suprémum pour la variable temporelle sous l'intégrale :

$$\forall\, t \in [0,T_0], \qquad \Big|\mathbf{G}(X)(t) - \mathbf{G}(Y)(t)\Big| \;\leq\; \lambda \int_0^t \big\|X-Y\big\|_{L^\infty} \mathrm{d}s = \lambda\, t\, \big\|X-Y\big\|_{L^\infty}. \tag{1.11}$$

Puis le suprémum à l'extérieur de l'intégrale :

$$\big\|\mathbf{G}(X) - \mathbf{G}(Y)\big\|_{L^\infty} \;\leq\; \lambda T_0\, \big\|X-Y\big\|_{L^\infty}. \tag{1.12}$$

Par conséquent, si $T_0 > 0$ est choisi assez petit (strictement inférieur à $1/\lambda$), alors $\mathbf{G}$ est contractante.

Etape 3 : Conclusion de la preuve par un argument de point-fixe. On souhaite utiliser le Théorème de point-fixe de Picard-Banach (Théorème 1.2) sur la fonctionnelle contractante $\mathbf{G}$. Le problème c'est que *a priori* la fonctionnelle $\mathbf{G}$ prends ses valeurs dans $\mathcal{C}^0([0,T_0],\mathbb{R}^d)$ et non pas dans $E = \mathcal{C}^0([0,T_0],K)$. Cependant, on remarque que (1.8) implique par l'inégalité triangulaire que

$$\forall\, t \in [0,T_0], \qquad \Big|\mathbf{G}(X)(t) - X_0\Big| \;\leq\; \int_0^t \Big|\mathcal{F}\Big(s, X(s)\Big)\Big| \,\mathrm{d}s \tag{1.13}$$

Puisque on travaille avec des fonctions $t \mapsto X(t)$ à valeurs dans le compact K, on en déduit, par le théorème des valeurs extrêmes de Weierstrass, que cette fonction est bornée. De même, puisque $[0,T_0]\times K$ est un compact, on a que $\mathcal{F}$ est bornée sur $[0,T_0]\times K$. On peut donc effectuer la majoration suivante :

$$\forall\, t \in [0,T_0], \qquad \Big|\mathbf{G}(X)(t) - X_0\Big| \;\leq\; \int_0^t \big\|\mathcal{F}\big\|_{L^\infty([0,T_0]\times K)} \,\mathrm{d}s \;\leq\; T_0\, \big\|\mathcal{F}\big\|_{L^\infty([0,T_0]\times K)} \tag{1.14}$$

Par conséquent, si on choisit T_0 assez petit, on a $T_0\, \big\|\mathcal{F}\big\|_{L^\infty} \leq R$ et donc pour tout $t \in [0,T_0]$ on a $\mathbf{G}(X)(t) \in K = \overline{\mathcal{B}}(X_0,R)$.

Ainsi, $\mathbf{G}$ est bien une fonction de E dans E contractante. On peut alors appliquer le Théorème de point-fixe de Picard-Banach (Théorème 1.2) et en déduire l'existence et l'unicité de $X^\star \in E$ tel que

$$X^\star = \mathbf{G}(X^\star). \tag{1.15}$$

Ceci équivaut à dire que $t \in [0,T_0] \;\mapsto X^\star(t)$ est l'unique fonction continue qui vérifie la formule de Duhamel (1.6). Cette formule étant équivalente au problème de Cauchy étudié, on a donc démontré l'existence et l'unicité des solutions au problème de Cauchy sur $[0,T_0]$. □

Remarque 1 **:** On démontre le théorème dans le cadre où la solution est définie sur un intervalle de temps de la forme $[0,T[$. Si on souhaite traiter un intervalle de temps plus général, de type $[T_1,T_2[$, avec une donnée initiale donnée au temps T_1, il suffit d'appliquer le théorème de Cauchy-Lipschitz au problème translatée, en étudiant $t \mapsto X(t-T_1)$. On mets donc le temps initial à 0 pour simplifier (en règle générale).

Remarque 2 **:** Il est également possible d'étudier ce qui se passe pour les temps négatif $]-T,0]$ avec une donnée initiale au temps 0. Il suffit pour cela d'appliquer le théorème de Cauchy-Lipschitz au problème symétrisée en temps en étudiant $t \mapsto X(-t)$. En conséquence, en plus du temps maximal d'existence sur les temps positifs $T^\star \in \mathbb{R}_+ \cup \{+\infty\}$, il y a aussi un temps maximal d'existence dans les temps négatifs $T_\star \in \mathbb{R}_- \cup \{-\infty\}$. La solution $t \mapsto X(t)$ définie sur l'intervalle de temps $]T_\star, T^\star[$ est souvent appelée *solution maximale.*

Remarque 3 **:** Dans le cas des équations différentielles faisant intervenir des dérivées d'ordre plus élevé, il est toujours possible de se ramener à un système d'équations avec seulement

des dérivées d'ordre 1 en faisant intervenir un plus grand nombre de variables. Par exemple si on travaille avec

$$\ddot{x}(t) = a\dot{x}(t) + bx(t),$$

on peut introduire une nouvelle variable $v(t) = \dot{x}(t)$ afin de faire diminuer l'ordre de la dérivation tout en faisant grossir la taille du système. En effet, l'équation ci-dessus est équivalente à

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} x \\ v \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ a & b \end{pmatrix}\begin{pmatrix} x \\ v \end{pmatrix}, \qquad \text{et} \qquad \begin{pmatrix} x \\ v \end{pmatrix}(0) = \begin{pmatrix} x_0 \\ v_0 \end{pmatrix}. \tag{1.16}$$

De manière générale (†) : il faut s'entrainer à reformuler les systèmes d'EDO sous la forme d'un problème de Cauchy !

1.2 Durée de vie d'une solution et Lemme de Grönwall

1.2.1 Recollement de deux solutions

Il est possible de recoller deux solutions d'une même équation différentielle posées sur des intervalles de temps contiguës. La fonction ainsi recollée devient solution de l'équation sur la réunion des deux intervalles de temps. Ce type de manipulations permet notamment d'étudier de temps de vie des solutions de l'EDO :

Théorème 1.3 Fin d'existence d'une solution

On considère, $\mathcal{F} : [0, T[\times\Omega \to \mathbb{R}^d$ et $X_0 \in \mathbb{R}^d$ fixés (avec Ω ouvert de $\mathbb{R}^d$ et $T \in \mathbb{R}_+^* \cup \{+\infty\}$) le problème de Cauchy suivant :

$$\frac{\mathrm{d}X}{\mathrm{d}t}(t) \;=\; \mathcal{F}\Big(t, X(t)\Big), \qquad \text{et} \qquad X(0) = X_0 \in \Omega. \tag{1.17}$$

On suppose que $T^\star$, le temps maximal d'existence de la solution $t \mapsto X(t)$, est strictement inférieur à T. Dans ce cas, pour tout $K \subseteq \Omega$ compact, il existe un temps $t_K \in [0, T^\star[$ tel que

$$\forall t \in]t_K, T^\star[, \qquad X(t) \in \Omega \setminus K.$$

Démonstration. *Etape 1 :* L'idée de la démonstration consiste à travailler sur un intervalle de temps $[0, T_1[$ sur lequel la conclusion du théorème est fausse et on démontre alors que $T_1 < T^\star$. Par contraposition, ceci démontrera le théorème. L'argument consiste à faire des "*recollements de solutions*" au temps T_1. On suppose donc qu'il existe un compact $K \subset \Omega$ et une suite de temps $(t_n)_{n\in\mathbb{N}}$ telle que $t_n \to T_1^-$ et $X(t_n) \in K$. Puisque K est un compact, on peut supposer (quitte à faire une extraction) que la suite de temps est telle que $X(t_n) \longrightarrow X_1$ pour un certain $X_1 \in \mathbb{R}^d$. On considère alors

$$\frac{\mathrm{d}Y}{\mathrm{d}t}(t) = \mathcal{F}\Big(t, Y(t)\Big), \qquad \text{et} \qquad Y(t = T_1) = X_1. \tag{1.18}$$

Par Cauchy-Lipschitz, la solution Y est bien définie de manière unique sur un intervalle $[T_1, T[$ pour un certain $T > T_1$. On pose alors $Z : [0, T[\to \Omega$ telle que $Z(t) := X(t)$ si $t < T_1$ et $Z(t) = Y(t)$ si $t \geq T_1$. On va démontrer que $t \mapsto Z(t)$ est $\mathcal{C}^1$ et vérifie :

$$\forall\, t \in [0, T[, \qquad \frac{\mathrm{d}Z}{\mathrm{d}t}(t) = \mathcal{F}\Big(t, Z(t)\Big).$$

Autrement dit, si on recolle la solution sur $[0, T_1[$ avec la solution sur $[T_1, T[$, on obtient la solution sur $[0, T[$. Un tel résultat impliquerait que T_1 n'est pas le temps maximal d'existence de la solution.

Etape 2 : Il est aisé de montrer que Z est $\mathcal{C}^1$ et vérifie la bonne équation pour les temps $t \in [0, T[$ dès lors que $t \neq T_1$. Le cas $t = T_1$ nécessite un argument pour montrer que le raccord est bien $\mathcal{C}^1$ et vérifie l'équation ci-dessus. On étudie d'abord la limite à droite (pour $t > T_1$; c'est-à-dire là où $Z(t) = Y(t)$).

— On sait que Y est continue, de sorte que $Y(t)$ converge lorsque $t \to T_1^+$ (elle converge vers X_1)
— Comme $\mathcal{F}$ est lipschitzienne, elle est donc continue et donc $t \mapsto \mathcal{F}\big(t, Y(t)\big)$ est aussi continue.
— Ainsi, cette fonction converge vers $\mathcal{F}\big(T_1, X_1\big)$.
— Il s'en suit, par l'égalité (1.18), que $\dot{Y}$ converge vers $\mathcal{F}\big(T_1, X_1\big)$ lorsque $t \to T_1^+$.

Concernant la limite à gauche (pour $t < T_1$; c'est-à-dire là où $Z(t) = X(t)$), il n'est pas possible de refaire directement le même raisonnement car on se sait pas si $X(t)$ converge vers X_1 lorsque $t \to T_1^-$. En effet, cette convergence n'est vraie que le long de la suite (t_n) (on ne sait même pas si $t \mapsto X(t)$ va rester dans le compact K entre t_n et t_{n+1}).

Etape 3. Soit $\varepsilon > 0$ fixé tel que $\overline{\mathcal{B}}(X_1, 2\varepsilon) \subseteq \Omega$. Pour N assez grand, on a $X(t_N) \in \overline{\mathcal{B}}(X_1, \varepsilon)$ (car $X(t_N) \to X_1$) et par continuité on sait que $X(t) \in \overline{\mathcal{B}}(X_1, 2\varepsilon)$ sur un petit intervalle de la forme $[t_N, t_N + \delta[$ avec $\delta > 0$. On sait que $\mathcal{F}$ est globalement lipschitz sur le compact $\overline{\mathcal{B}}(X_1, 2\varepsilon)$ et donc bornée (par le théorème des valeurs extrêmes de Weierstrass). Ainsi, avec l'inégalité triangulaire, on obtient pour tout $t \in [t_N, t_N + \delta[$,

$$\begin{aligned} \big|X(t) - X(t_N)\big| &= \left| \int_{t_N}^{t} \frac{\mathrm{d}X}{\mathrm{d}t}(s)\,\mathrm{d}s \right| = \left| \int_{t_N}^{t} \mathcal{F}\Big(s, X(s)\Big)\,\mathrm{d}s \right| \\ &\leq \int_{t_N}^{t} \left| \mathcal{F}\Big(s, X(s)\Big) \right| \mathrm{d}s \leq \int_{t_N}^{t} \big\|\mathcal{F}\big\|_{L^\infty(\overline{\mathcal{B}}(X_1, 2\varepsilon))} \mathrm{d}s \end{aligned} \tag{1.19}$$

En sortant la norme L^∞ de l'intégrale (car c'est une constante) on aboutit à :

$$\big|X(t) - X(t_N)\big| \leq \big\|\mathcal{F}\big\|_{L^\infty(\overline{\mathcal{B}}(X_1, 2\varepsilon))} \int_{t_N}^{t} ds = (t - t_N)\|\mathcal{F}\big\|_{L^\infty(\overline{\mathcal{B}}(X_1, 2\varepsilon))} \leq \delta \,\big\|\mathcal{F}\big\|_{L^\infty(\overline{\mathcal{B}}(X_1, 2\varepsilon))}. \tag{1.20}$$

On choisit à présent δ le plus grand possible. Dans le cas où on a $t \mapsto X(t)$ qui reste dans $\overline{\mathcal{B}}(X_1, 2\varepsilon)$ jusqu'au temps T_1 (cas 1) on peut conclure que $t \mapsto X(t)$ est borné sur l'intervalle $[t_N, T_1]$. Si au contraire (cas 2) on peut choisir $\delta > 0$ de telle sorte que $t \mapsto X(t)$ sorte de la boule pour la première fois au temps $t_N + \delta$ (ce qui veux dire $\big|X(t_N + \delta) - X(t_N)\big| = 2\varepsilon$), l'inégalité ci-dessus implique

$$2\varepsilon \leq \delta \|\mathcal{F}\big\|_{L^\infty(\overline{\mathcal{B}}(X_1, 2\varepsilon))}. \tag{1.21}$$

Donc si on choisit N assez grand de sorte que $T_1 - t_N > 2\varepsilon / \|\mathcal{F}\big\|_{L^\infty(\overline{\mathcal{B}}(X_1, 2\varepsilon))}$, on en déduit que le cas 2. ne se produit pas.

Etape 4. Le bilan de l'étape 3. est que $t \mapsto X(t)$ reste borné pour $t \in [t_N, T_1]$ pour $N \in \mathbb{N}$ assez grand.

— Comme $\mathcal{F}$ est continue, on en déduit que $t \mapsto \mathcal{F}\big(t, Y(t)\big)$ est également bornée.
— On utilise à présent l'équation pour obtenir que $\dot{X}$ est bornée sur $t \in [t_N, T_1]$
— Par conséquent on a $t \mapsto X(t)$ qui se prolonge par continuité lorsque $t \to T_1^-$.

On peut alors refaire le raisonnement de l'étape 2 et en déduire que $t \mapsto Z(t)$ admet un prolongement $\mathcal{C}^1$ au temps $t = T_1$ et que $\dot{Z}(T_1) = \mathcal{F}\big(T_1, Z(T_1)\big)$. □

Dans la pratique on utilise plutôt le corollaire suivant :

Corollaire 1.1 Théorème de prolongement **(†)**

On considère, $\mathcal{F} : [0, T[\times\Omega \to \mathbb{R}^d$ et $X_0 \in \mathbb{R}^d$ fixés (avec Ω ouvert de $\mathbb{R}^d$ et $T \in]0, +\infty]$), le problème de Cauchy suivant :

$$\frac{\mathrm{d}X}{\mathrm{d}t}(t) \;=\; \mathcal{F}\Big(t, X(t)\Big), \qquad \text{et} \qquad X(0) = X_0 \in \Omega. \tag{1.22}$$

On suppose qu'il existe un compact $K \subset \Omega$ et un $T_0 < T$ tels que $X(t) \in K$ pour tout temps $t \in [0, T_0[$. Alors dans ce cas $T_0 < T^*$.

1.2.2 Lemme de Grönwall

Le théorème de fin d'existence des solutions (Théorème 1.3) ou son corollaire, à savoir le théorème de prolongement, nous donnent une caractérisation du temps de fin d'existence T^*. Le théorème de Cauchy-Lipschitz nous donne l'existence locale ; il est naturel de se demander si $T^\star = +\infty$, c'est-à-dire si on a l'existence *l'existence globale* de la solution. Malheureusement, on peut exhiber de nombreux exemples de problèmes de Cauchy pour lesquels l'existence globale n'est pas toujours vraie. Un exemple simple pour le voir est le suivant :

$$\dot{x} = 1 + x^2, \qquad \text{avec} \qquad x(0) = 0. \tag{1.23}$$

La solution de cette équation est la bien-connue fonction "tangente" et le temps maximal d'existence de la solution est $T^\star = \pi/2$. L'équation ci-dessus cesse d'être bien posé au temps $\pi/2$ et on ne peut pas prolonger la solution de manière $\mathcal{C}^1$. Un autre exemple pour lequel l'existence globale n'est pas assuré est le suivant :

$$\dot{x} = -\frac{1}{2x}, \qquad \text{avec} \qquad x(0) = 1. \tag{1.24}$$

Cette fois la solution est $t \mapsto \sqrt{1-t}$ et sa dérivée diverge lorsqu'on s'approche du temps $T^\star = 1$. Dans ce deuxième exemple, l'ouvert sur lequel est bien posé le problème est $\Omega = \mathbb{R}^*$. On peut vérifier directement sur ces deux exemples la validité du théorème de fin d'existence (Théorème 1.3) : les solutions sortent bien définitivement de tous les compacts (ceux inclus dans $\mathbb{R}$ pour le premier exemple et ceux dans $\mathbb{R}^*$ pour le second).

Afin d'étudier le *temps d'existence* de la solution d'une équation différentielle, on utilise généralement le lemme suivant :

Lemme 1.1 Lemme de Grönwall (version intégrale) **(†)**

Si $\psi \geq 0$ et ϕ sont deux fonctions continues réelles sur un intervalle $[a, b]$ qui vérifient

$$\forall\, t \in [a, b], \qquad \phi(t) \leq K + \int_a^t \psi(s)\phi(s)\,\mathrm{d}s, \tag{1.25}$$

pour une certaine constante $K \in \mathbb{R}$, alors

$$\forall\, t \in [a, b], \qquad \phi(t) \leq K\, \exp\left(\int_a^t \psi(s)\,\mathrm{d}s\right). \tag{1.26}$$

Démonstration. (†) On pose

$$f(t) = \frac{K + \int_a^t \psi(s)\phi(s)\,\mathrm{d}s}{\exp\big(\int_a^t \psi(s)\,\mathrm{d}s\big)} \tag{1.27}$$

On calcule sa dérivée

$$\frac{\mathrm{d}f}{\mathrm{d}t} = \psi(t)\,\frac{\phi(t) - K - \int_a^t \psi(s)\phi(s)\,\mathrm{d}s}{\exp\big(\int_a^t \psi(s)\,\mathrm{d}s\big)}. \tag{1.28}$$

En utilisant l'hypothèse, on remarque que le numérateur de cette fraction est négatif. Comme ψ est positive, on en déduit que $df/\mathrm{d}t \leq 0$ et donc $f(t)$ est décroissante. En utilisant à nouveau l'hypothèse du lemme, le caractère décroissant de f donne :

$$\frac{\phi(t)}{\exp\big(\int_a^t \psi(s)\,\mathrm{d}s\big)} \leq f(t) \leq f(0) = K. \tag{1.29}$$

□

Remarque **:** Si on veut généraliser en autorisant la constante K à varier au cours du temps $(t \mapsto K(t))$, l'implication donnée par le lemme de Grönwall devient :

$$\forall t \geq t_0 \qquad \phi(t) \leq K(t) + \int_{t_0}^t \psi(s)\phi(s)\,\mathrm{d}s$$

implique

$$\forall t \geq t_0 \qquad \phi(t) \leq K(t) + \int_{t_0}^t K(s)\psi(s)\exp\left(\int_s^t \psi(\tau)\,\mathrm{d}\tau\right)\,\mathrm{d}s.$$

Le lemme de Grönwall admet également une version différentielle :

Lemme 1.2 Lemme de Grönwall (version différentielle) (†)

Si ψ et ϕ sont deux fonctions dérivables réelles sur un intervalle $[a, b]$ qui vérifient pour tout temps

$$\frac{\mathrm{d}\phi}{\mathrm{d}t}(t) \leq \psi(t)\,\phi(t),$$

alors on a :

$$\forall\, t \in [a, b], \qquad \phi(t) \leq \phi(a)\exp\left(\int_a^t \psi(s)\,\mathrm{d}s\right).$$

Si on suppose que ϕ est positive alors ce lemme est plus faible que le lemme de Grönwall sous forme intégrale (car on peut intégrer les inégalités mais pas les dériver). L'intérêt de la version différentielle est qu'il n'est pas nécessaire de faire d'hypothèse de signe sur ψ.

La démonstration est analogue ; il faut commencer par dériver $t \mapsto \dfrac{\phi(t)}{\exp\big(\int_a^t \psi(s)\,\mathrm{d}s\big)}$ □

Le principal intérêt du lemme de Grönwall réside dans le fait qu'il nous permet d'étudier le comportement d'une solution au voisinage de la fin de son intervalle d'existence $T^\star$. A titre d'illustration, nous allons démontrer le théorème suivant :

Théorème 1.4 Théorème de Cauchy-Lipschitz global

On considère le problème de Cauchy (1.3) avec $T = +\infty$ et $\Omega = \mathbb{R}^d$.

Si le champ de vitesse $\mathcal{F}$ est une fonction ***globalement*** *lipschitzienne pour la variable* X, alors le problème de Cauchy (1.3) admet une unique solution **globale**.

Démonstration. D'après le théorème de Cauchy-Lipschitz "*local*" (Théorème 1.1), le problème de Cauchy étudié admet une unique solution locale $t \mapsto X(t) \in \mathbb{R}^d$. On considère à présent un temps $T_0 > 0$ tel que $T_0 < +\infty$ et tel que la solution soit bien définie sur $[0, T_0[$. Par le théorème fondamental de l'analyse, nous avons pour tout $t \in [0, T_0[$ (en utilisant l'équation vérifiée par X) :

$$X(t) - X_0 \;=\; \int_0^t \frac{\mathrm{d}X}{dt}(s)\,\mathrm{d}s \;=\; \int_0^t \mathcal{F}\Big(s, X(s)\Big)\,\mathrm{d}s. \tag{1.30}$$

Par conséquent, d'après l'inégalité triangulaire,

$$\begin{aligned} \big|X(t) - X_0\big| \;&=\; \left|\int_0^t \left[\mathcal{F}\Big(s, X(s)\Big) - \mathcal{F}\Big(s, X_0\Big) + \mathcal{F}\Big(s, X_0\Big)\right]\mathrm{d}s\right| \\ &\leq\; \int_0^t \left|\mathcal{F}\Big(s, X(s)\Big) - \mathcal{F}\Big(s, X_0\Big)\right|\mathrm{d}s + \int_0^t \left|\mathcal{F}\Big(s, X_0\Big)\right|\mathrm{d}s. \end{aligned} \tag{1.31}$$

D'une part nous avons, par le théorème des valeurs extrêmes de Weierstrass (en prenant le suprémum sous l'intégrale) :

$$\int_0^t \left|\mathcal{F}\Big(s, X_0\Big)\right|\mathrm{d}s \;\leq\; t \sup_{s\in[0,T_0]} \left|\mathcal{F}\Big(s, X_0\Big)\right|\mathrm{d}s \;\leq\; T_0 \left\|\mathcal{F}(\cdot, X_0)\right\|_{L^\infty([0,T_0])}. \tag{1.32}$$

D'autre part, comme $\mathcal{F}$ est globalement lipschitzienne, on a une constante λ telle que

$$\int_0^t \left|\mathcal{F}\Big(s, X(s)\Big) - \mathcal{F}\Big(s, X_0\Big)\right|\mathrm{d}s \;\leq\; \int_0^t \lambda\big|X(s) - X_0\big|\mathrm{d}s \tag{1.33}$$

Ces deux estimées impliquent que

$$\big|X(t) - X_0\big| \;\leq\; T_0 \left\|\mathcal{F}(\cdot, X_0)\right\|_{L^\infty([0,T_0])} + \int_0^t \lambda\big|X(s) - X_0\big|\mathrm{d}s. \tag{1.34}$$

On peut alors appliquer le Lemme de Grönwall version intégrale (Lemme 1.1) à la fonction $\phi : t \mapsto |X(t) - X_0|$ avec $K = T_0 \left\|\mathcal{F}(\cdot, X_0)\right\|_{L^\infty}$ et $\psi : t \mapsto \lambda$ et ainsi obtenir :

$$\big|X(t) - X_0\big| \;\leq\; T_0 \left\|\mathcal{F}(\cdot, X_0)\right\|_{L^\infty} \exp\big(\lambda t\big). \tag{1.35}$$

Par conséquent, pour tout $t \in [0, T_0[$, on a $X(t)$ dans la boule fermée de centre X_0 et de rayon $T_0 \left\|\mathcal{F}(\cdot, X_0)\right\|_{L^\infty} \exp\big(\lambda T_0\big)$. Or, cette boule est un fermé borné de dimension finie (donc elle est compacte). On peut donc appliquer le théorème de prolongement et conclure que $T_0 < T^\star$. Comme T_0 a été choisi de manière quelconque dans l'intervalle $]0, T^\star] \setminus \{+\infty\}$, on en déduit que $T^\star = +\infty$. On a donc démontré l'existence globale. □

Remarque **:** De manière générale, une méthode possible pour montrer l'existence globale est la suivante :

1) On étudie les propriétés asymptotiques du champs de vitesse $\mathcal{F}$ aux voisinages de $\partial\Omega$.
2) On exploite l'équation pour étudier des propriétés sur la croissance asymptotique de la solution et faire le lien entre $X(t)$ et sa dérivée (éventuellement avec une formulation intégrale comme dans la preuve ci-dessus).
3) Le lemme de Grönwall nous permet de montrer que la solution reste dans un compact quitte à effectuer une localisation en temps (sur un intervalle $[0, T_0[$ avec T_0 quelconque).
4) Le théorème de prolongement nous permet de conclure à l'existence globale.

1.3 Résolution d'équations et analyse asymptotique

Dans cette section on présente un certain nombre de techniques pour résoudre explicitement des équations différentielles ordinaires dans certains cas particuliers. On peut ensuite utiliser ces solutions particulières pour les comparer avec des équations plus compliquées afin de comprendre leur comportement asymptotique de leurs solutions.

1.3.1 Equations différentielles linéaires à coefficients constants

Lemme 1.3 Equations différentielles linéaires à coefficients constants **(†)**

Soit $X_0 \in \mathbb{R}^d$ et soit $A \in \mathcal{M}_d(\mathbb{R})$ une matrice à coefficients constants. Soit $b : [0, T[\to \mathbb{R}^d$ continu. Alors le problème de Cauchy

$$\frac{dX}{\mathrm{d}t} = AX + b(t), \qquad \text{et} \qquad X(0) = X_0 \tag{1.36}$$

admet comme unique solution : $X(t) = e^{tA}X_0 + \displaystyle\int_0^t e^{(t-s)A}\, b(s)\, \mathrm{d}s$.

La démonstration de ce lemme est laissé en exercice au lecteur. Ainsi, l'étude des systèmes d'équations différentielles linéaires se fait à l'aide du concept d'exponentielle de matrice introduit à la définition 0.19. Une étude plus systématique du cas particulier $b \equiv 0$ sera faite au chapitre 2.

Un exemple important : l'oscillateur harmonique amorti. On étudie une équation sur $x : [0, T[\to \mathbb{R}$ de la forme

$$\ddot{x} + \alpha\, \dot{x} + \beta\, x = 0, \qquad \text{et} \qquad x(0) = x_0, \quad \dot{x}(0) = v_0, \tag{1.37}$$

avec $a \geq 0$ et $b > 0$. On commence par mettre ce système sous la forme (1.36) en introduisant la variable auxiliaire $v(t) := \dot{x}(t)$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} x \\ v \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -\beta & -\alpha \end{pmatrix}\begin{pmatrix} x \\ v \end{pmatrix}, \qquad \text{et} \qquad \begin{pmatrix} x \\ v \end{pmatrix}(0) = \begin{pmatrix} x_0 \\ v_0 \end{pmatrix}. \tag{1.38}$$

Pour étudier l'exponentielle de la matrice ci-dessus (noté M), on utilise la réduction de Jordan (théorème 0.7). Le calcul du polynôme caractéristique nous donne 3 cas.

Cas 1 : Si les deux valeurs propres, notées λ et μ, sont réelles distinctes alors la matrice est diagonalisable. Puisqu'il s'agit d'une matrice de taille 2, nous avons $\mathrm{tr}(A) = \lambda + \mu$ et $\det(A) = \lambda\mu$. Comme la trace est négative et le déterminant positif, on en déduit que $\lambda, \mu < 0$. Le calcul de l'exponentielle de la matrice A nous dit que l'ensemble des solutions à l'équation différentielle est l'espace vectoriel de dimension 2 engendré par les vecteurs

$$t \longmapsto e^{\lambda t} \qquad \text{et} \qquad t \longmapsto e^{\mu t}. \tag{1.39}$$

Les solutions sont donc de la forme $x(t) = Ae^{\lambda t} + Be^{\mu t}$ avec A et B deux constantes déterminées grâce aux conditions initiales. Les solutions convergent vers 0 exponentiellement en décroissant sans faire d'oscillations.

Cas 2 : Si en revanche on a des valeurs propres complexes conjuguées $\lambda = a + \mathrm{i}b$ et $\mu = a - \mathrm{i}b$ alors par les même arguments que précédemment, on a $a \leq 0$. Le calcul de l'exponentielle de

matrice implique que l'ensemble des solutions à l'équation différentielle est l'espace vectoriel de dimension 2 engendré par les vecteurs

$$t \longmapsto e^{at}\sin(bt) \qquad \text{et} \qquad t \longmapsto e^{at}\cos(bt), \tag{1.40}$$

où l'on a utilisé les formules d'Euler - De Moivre :

$$\cos(\theta) = \frac{e^{\mathrm{i}\theta}+e^{-\mathrm{i}\theta}}{2}, \qquad \text{et} \qquad \sin(\theta) = \frac{e^{\mathrm{i}\theta}-e^{-\mathrm{i}\theta}}{2\mathrm{i}}. \tag{1.41}$$

Les solutions sont donc de la forme $x(t) = e^{at}(A\cos(bt)+B\sin(bt))$ avec A et B deux constantes déterminées grâce aux conditions initiales. Les solutions convergent vers 0 exponentiellement en faisant une infinité d'oscillations (sauf dans le cas particulier $a = 0$, c'est-à-dire si $\alpha = 0$, où l'on a des oscillations sans amortissement).

Cas 3 : Le dernier cas, appelé cas critique, correspond au cas où les deux valeurs propres sont réelles et égales, $\lambda = \mu < 0$. Dans ce cas, la matrice n'est pas diagonalisable (car sinon elle serait égale à λI_d) ; elle est semblable à la matrice Jordan $J_2(\lambda)$. Le calcul de l'exponentielle de cette matrice nous donne que l'ensemble des solutions à l'équation différentielle est l'espace vectoriel de dimension 2 engendré par les vecteurs

$$t \longmapsto e^{\lambda t} \qquad \text{et} \qquad t \longmapsto t\,e^{\lambda t}. \tag{1.42}$$

Les solutions sont donc de la forme $x(t) = e^{\lambda t}(A + Bt)$ avec A et B deux constantes qui sont déterminées à l'aide des données initiales. Les solutions convergent vers 0 exponentiellement vite avec une unique oscillation. On peut démontrer que, β étant fixé, la solution donnée par le 3^e cas est celle qui converge vers 0 le plus vite lorsqu'on fait varier α. Si on note $\alpha^\star$ le alpha critique correspondant, on peut vérifier que si $\alpha < \alpha^\star$ alors on est dans le cas 2 et si $\alpha > \alpha^\star$ alors on est dans le cas 1. Le cas critique correspond au cas d'apparition des oscillations.

Remarque : La formule donnée par le lemme 1.3 devient fausse si on autorise les coefficients de la matrice A à dépendre du temps. Dans ce cas la résolution explicite n'est pas toujours possible. Néanmoins, il est possible de faire une étude à l'aide du déterminant de la matrice wronskienne (hors programme).

1.3.2 Équations différentielles séparables

Dans le cas particulier des équations scalaires, c'est-à-dire dont l'inconnue est une fonction à valeur dans $\mathbb{R}$, il y a un cas pour lequel l'étude est beaucoup plus simple : ce sont les *équations séparables*. On dit qu'une équation d'inconnue $t \mapsto x(t) \in \mathbb{R}$ est séparable si elle se met sous la forme suivante :

$$\frac{\mathrm{d}x}{\mathrm{d}t} = b(t)\,g\big(x(t)\big) \qquad \text{et} \qquad x(0) = x_0, \tag{1.43}$$

avec $b : [0, T[\to \mathbb{R}$ et $g : \mathbb{R} \to \mathbb{R}$. Dans ce cas, la méthode pour étudier cette équation différentielle s'appelle la *méthode de séparation des variables* et elle consite à effectuer les opérations suivantes (†) :

1. On divise à gauche et à droite par $g(x)$ et on intègre en temps :

$$\int_0^t \frac{\dot{x}(s)}{g(x(s))}\,\mathrm{d}s = \int_0^t b(s)\,\mathrm{d}s.$$

2. On fait alors le changement de variable $u = x(s)$ et on aboutit à

$$\int_{x(0)}^{x(t)} \frac{\mathrm{d}u}{g(u)} = \int_0^t b(s)\,\mathrm{d}s. \tag{1.44}$$

3. En posant $\phi(y) := \displaystyle\int_0^y \frac{\mathrm{d}s}{g(s)}$, On peut intégrer et obtenir

$$\phi\big(x(t)\big) - \phi\big(x_0\big) = \int_0^t b(s)\,\mathrm{d}s. \tag{1.45}$$

4. On inverse la fonction ϕ pour obtenir une formule sur $x(t)$.

Dans beaucoup de cas, cette méthode nécessite des adaptations en fonction de l'équation étudiée. **Il faut donc savoir mettre en place cette méthode au cas-par-cas.**

Cas particulier important : Dans le cas où $\dot{x} = ax^\alpha$ avec $a, \alpha \in \mathbb{R}$ et $x(0) = x_0 > 0$, la méthode de séparation des variables donne (exercice) :

$$\begin{aligned} &x(t) = x_0 \exp\big(at\big) && \text{si} \quad \alpha = 1, \\ &x(t) = \Big(x_0^{1-\alpha} - a(1-\alpha)t\Big)^{\frac{1}{1-\alpha}} && \text{sinon.} \end{aligned} \tag{1.46}$$

On remarque que cette équation est globalement bien posée (sur $\mathbb{R}$) seulement dans les cas $\alpha = 0$ et $\alpha = 1$. Les 3 cas possibles de perte du caractère bien posé sont présents ici :

— Si $\alpha > 1$, on a $x(t) \to \pm\infty$ (en fonction du signe de a) lorsque $t \to T^\star$.

— Si $\alpha < 0$, on travaille sur $\Omega = \mathbb{R}^*$ et on a $x(t)$ qui converge vers 0 en temps fini. En revanche $\dot{x}(t) \to \pm\infty$.

— Si $0 < \alpha < 1$ alors $x(t)$ et $\dot{x}(t)$ convergent mais on perd l'unicité de la solution (on peut exhiber deux solutions différentes qui ont la même valeur au moment où $X(t) = 0$).

1.3.3 Perte de l'unicité dans les équations différentielles ordinaires

Si on considère le problème de Cauchy initial (1.3), nous avons une propriété d'unicité pour l'ensemble des solutions pour une équation définie sur un ouvert Ω avec une donnée initiale fixée pourvu que le champs de vitesse soit lipschitzien en X (Théorème de Cauchy-Lipschitz). Comme on peut appliquer le théorème de Cauchy-Lipschitz pour tous les instants de l'intervalle de temps, on en déduit que les solutions sont distinctes pour tout temps :

Lemme 1.4 Propriété de séparation des solutions

Soit $X_0, Y_0 \in \Omega$ et soit $\mathcal{F}$ une fonction qui satisfait les hypothèses du théorème de Cauchy-Lipschitz (Théorème 1.1). On note respectivement X et Y les solutions au problème de Cauchy avec comme donnée initiale X_0 et Y_0 respectivement.

On suppose qu'il existe un temps t_0 pour lequel $X(t_0)$ et $Y(t_0)$ soient bien définis et tel que $X(t_0) = Y(t_0)$. Alors $X \equiv Y$.

La notation "$\equiv$" utilisée ci-dessus est une notation standard pour signifier que les fonctions coïncident en tout point. Il faut impérativement savoir refaire le raisonnement qui permet de démontrer un tel résultat :

Démonstration. On suppose par l'absurde qu'il existe un temps $t_0 > 0$ tel que $X(t_0) = Y(t_0)$ mais $X(t) \neq Y(t)$ sur un petit intervalle $[t_0 - \delta, t_0[$. On considère alors le problème de Cauchy rétrograde (c'est-à-dire avec le temps qui va "en sens inverse") et qui commence au temps t_0. Autrement dit, on remplace la variable t par $t_0 - t$. On obtient le problème de Cauchy suivant :

$$\frac{\mathrm{d}Z}{\mathrm{d}t} = -\mathcal{F}\Big(t_0 - t, Z(t)\Big), \qquad \text{et} \qquad Z(0) = Y(t_0) = X(t_0) \in \Omega. \tag{1.47}$$

Par le théorème de Cauchy-Lipschitz, ce problème de Cauchy admet une unique solution sur un petit intervalle de temps $[0, T(X_0)[$. Or, les fonctions $t \mapsto X(t_0 - t)$ et $t \mapsto Y(t_0 - t)$ sont solutions de ce problème de Cauchy (par hypothèse). On en déduit alors que

$$\forall\, t \in [0, T[, \qquad X(t_0 - t) = Y(t_0 - t). \tag{1.48}$$

Ceci est contradictoire avec l'hypothèse initiale. □

On peut reformuler ce lemme en disant que si deux solutions à une même équation différentielle diffèrent en un temps donné alors elles diffèrent pour tout temps. Par exemple, pour les équations scalaires (pour les EDO à valeur en $\mathbb{R}^d$ avec $d = 1$), on a la propriété dite des ***solutions étagées*** :

$$X(0) < Y(0) \qquad \Longrightarrow \qquad X(t) < Y(t), \quad \forall\, t. \tag{1.49}$$

Remarque : Il existe des situations où la fonction $\mathcal{F}$ cesse d'être lipschitzienne par rapport à X en un certain temps t_0 et pourtant l'existence de solutions peut rester valide. Dans ce genre de configurations, on peut alors avoir une perte l'unicité au temps t_0. Cela se caractérise par l'existence de 2 solutions distinctes X et Y qui vont coïncider au temps t_0. Autrement dit, la singularité dans la fonction $\mathcal{F}$ au temps t_0 rend possible de fait d'avoir $X(t) \neq Y(t)$ pour tout $t < t_0$ mais $X(t_0) = Y(t_0)$ avec X et Y solutions de la même équation $\dot{X} = \mathcal{F}(t, X)$.

1.3.4 Analyse asymptotique par principes de comparaison

Dans de nombreux cas il n'est pas possible d'obtenir une formule explicite pour les solution d'une équation différentielle donnée. Pourtant, nous pouvons néanmoins étudier les propriétés qualitatives des solutions dans le cas 1D grâce à deux méthodes différentes dont nous présentons ici les idées générales sur un exemple :

$$\frac{\mathrm{d}x}{\mathrm{d}t} = x(t)\, \arctan\big(x(t)\big)\, \sin(t), \qquad \text{et} \qquad x(0) = x_0. \tag{1.50}$$

Principe de comparaison des solutions **:** cette méthode consiste à comparer deux solutions différentes d'une même équation mais avec des données initiales différentes. Dans notre cas, on peut démontrer que si $x_0 \neq 0$ alors la solution de (1.50) ne s'annule jamais.

1. On constate tout d'abord que la fonction nulle est solution de l'équation avec $x_0 = 0$.
2. On considère x une solution avec $x_0 \neq 0$ et on suppose qu'en un temps $t_0 > 0$ on a $x(t_0) = 0$.
3. On regarde alors l'équation rétrograde (en replaçant t par $t_0 - t$). On a donc trouvé 2 fonctions solutions, la fonction nulle et $x(t_0 - t)$, pour l'équation rétrograde avec la même donnée initiale 0.
4. Ceci est contradictoire avec le théorème de Cauchy-Lipschitz.

Principe de comparaison des équations **:** Cette fois-ci, on va comparer la solution à l'équation avec des solutions d'autre équations similaires plus simples (en général on conserve la donnée initiale).

1. On sait que arc-tangente prend ses valeurs entre $-\pi/2$ et $\pi/2$ et donc : $\dfrac{\mathrm{d}x}{\mathrm{d}t} \leq \dfrac{\pi}{2} x(t)$.
2. Si on suppose par exemple que $x_0 > 0$ alors $x(t) > 0$ et donc $\quad \dfrac{\dot{x}(t)}{x(t)} \leq \dfrac{\pi}{2}$.
3. On intègre cette inégalité en temps en on en déduit que : $\quad x(t) \leq x_0 \exp\Big(\dfrac{\pi}{2}\, t\Big)$.

Remarque : La solution est ainsi localement bornée et donc globalement bien posée par la caractérisation du temps de fin d'existence (Théorème 1.3).

1.4 Bilan du Chapitre et exercices

1.4.1 Ce qu'il faut retenir et savoir-faire

Ce chapitre est consacré aux théorèmes d'existence et d'unicité des solutions aux équations différentielles mises sous la forme d'un problème de Cauchy. Il se termine par une analyse exacte ou asymptotique dans des cas particuliers. Les principaux éléments à retenir sont les suivants :

(‡) Problème de Cauchy et théorème de Cauchy-Lipschitz.
- Savoir reformuler une EDO sous la forme d'un problème de Cauchy.

(†) Caractérisation de la fin d'existence des solutions (théorème 1.3).
(†) Lemme de Grönwall (les deux versions) et sa démonstration.
(†) Résolution des équations différentielles linéaires à coefficients constants.
- Méthode de séparation des variables.
- Méthode de variation de la constante
- Les deux principes de comparaison.

1.4.2 Exercices

Les exercices ci-dessous proposent d'étudier quelques équations différentielles à l'aide des outils présentés dans ce chapitre 1. Il est recommandé de les traiter dans l'ordre. Les exercices les plus importants sont identifiés avec le symbole ($\star$).

Exercice 1.1 (Exercice introductif)**.** ($\star$) On considère le problème suivant dont l'inconnue est une fonction $x : [0, T) \to \mathbb{R}$ de classe $\mathcal{C}^1$ vérifiant les conditions suivantes

$$\dot{x}(t) = -x(t) + t, \qquad \text{et} \qquad x(0) = 0.$$

1) Montrer que ce problème se met sous forme de Cauchy :

$$\dot{x}(t) = \mathcal{F}\Big(t, x(t)\Big), \qquad \text{et} \qquad x(0) = 0,$$

où $\mathcal{F} : [0, T) \times \mathbb{R} \to \mathbb{R}$ est une fonction que l'on précisera et où $x_0 \in \mathbb{R}$ (à préciser également).

2) Montrer que la fonction $\mathcal{F}$ est de classe $\mathcal{C}^1$. En déduire qu'elle est lipschitzienne pour sa 2e variable.

3) A l'aide du Théorème de Cauchy-Lipschitz, montrer qu'il existe une unique solution locale à cette équation.

4) Remarquer à présent que cette équation est linéaire à coefficients constants et résoudre explicitement cette équation à l'aide d'une résolution exponentielle (voir Lemme 1.3).

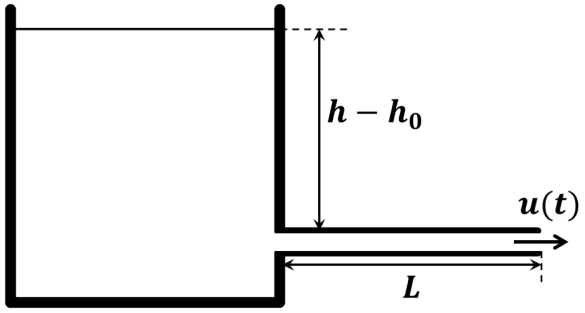
Exercice 1.2 (Utilisation du théorème de Cauchy-Lipschitz)**.** ($\star$) Montrer l'existence et l'unicité locale d'une solution pour les équations suivantes (avec une donnée initiale quelconque) :

- $\dot{x} = \sin(x - t) - x - t.$
- $\dot{x} = x^2 + y^2, \quad \text{et} \quad \dot{y} = y^2 - x^2.$
- $\dot{x} = x - 2y + z + t, \quad \dot{y} = x + y - 3z + t, \quad \text{et} \quad \dot{z} = x + y + z - t.$
- $\ddot{\theta} + \sin(\theta) = 0.$

Exercice 1.3 (La sortie du tuyau)**.** ($\star$) On considère un grand réservoir de hauteur totale h et à sa base se trouve un tuyau de longueur L relié au réservoir par une ouverture à la hauteur h_0. Si on suppose que h et L sont grands par rapport au diamètre du tuyau et que le fluide est initialement au repos, on peut simplifier les équations de Euler de la mécanique des fluides (l'approximation est valable sur un intervalle de temps pas trop grand de sorte que h puisse être considéré comme constant) et alors obtenir :

$$L\frac{du}{\mathrm{d}t} = (h - h_0)g - \frac{u^2}{2}, \tag{1.51}$$

où g et l'accélération de la pesanteur et où $u : [0, T[\to \mathbb{R}$ la vitesse de fluide à la sortie du tuyau est l'inconnue du problème. Au temps initial cette vitesse est nulle.



1) Réécrire ce système sous la forme d'un problème de Cauchy :

$$\dot{u}(t) = \mathcal{F}\big(t, u(t)\big), \qquad \text{et} \qquad u(0) = 0,$$

où $\mathcal{F}$ est une fonction que l'on précisera.

2) A l'aide du théorème de Cauchy-Lipschitz, montrer que ce problème est localement bien posé (existence d'une unique solution locale en temps).

3) Montrer qu'il existe une donné initiale positive (que l'on précisera) telle que la solution à l'équation soit constante. On note cette constante u_∞.

4) Montrer que une autre solution de l'EDO ne peut jamais atteindre cette valeur u_∞.

5) Montrer que la solution associée à la donnée initiale $u(0) = 0$ est croissante. En déduire à l'aide du théorème de prolongement, que ce problème de Cauchy est globalement bien posé (existence et unicité pour $t \in [0, +\infty[$).

6) Montrer que cette solution est concave. Montrer alors que $u(t)$ et $\dot{u}(t)$ convergent lorsque $t \to +\infty$.

7) Montrer que la limite de $\dot{u}(t)$ est 0 puis que la limite de $u(t)$ est u_∞.

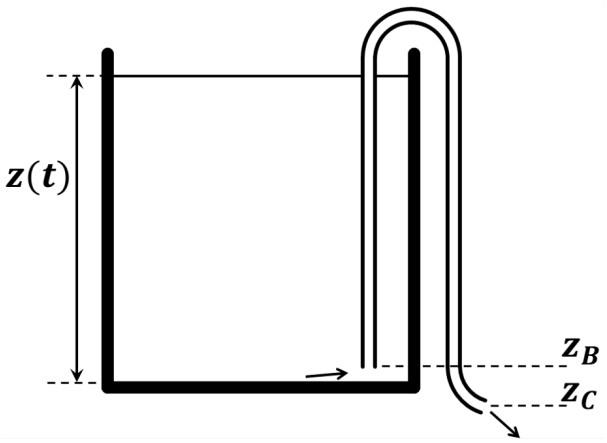
8) A l'aide de la méthode de séparation des variables, résoudre cette équation explicitement et montrer que $|u_\infty - u(t)|$ décroît exponentiellement.

9) Calculer le temps t_ε à partir duquel $u(t)$ est à distance $\varepsilon > 0$ de la limite u_∞ (avec $\varepsilon > 0$ fixé quelconque).

Exercice 1.4 (Vidange d'un réservoir par un siphon)**.** ($\star$) On considère un réservoir cylindrique de section S muni d'un siphon de section s. Initialement le réservoir est rempli jusqu'à une hauteur H. A partir des équations de Euler, on peut dériver une équation simplifiée pour l'évolution de la hauteur de l'eau $z(t)$ dans le réservoir cylindrique :

$$\frac{1}{g}\left(\frac{S}{s}\dot{z}(t)\right)^2 = z(t) - z_C,$$

avec z_C la hauteur de l'extrémité du siphon située à l'extérieur du réservoir.



1) Sachant que $\dot{z}$ est négatif, écrire ce système sous la forme d'un problème de Cauchy défini sur $[z_c, +\infty[$ et montrer qu'il existe une unique solution locale sur l'ouvert $\Omega :=]z_c, +\infty[$.

2) On se place dans l'hypothèse $z_B \leq z_C$, où z_B est la la hauteur de l'extrémité du siphon située à l'intérieur du réservoir. A l'aide de la méthode de séparation des variables, montrer que le système atteint la position $z = z_C$ en temps fini et qu'il cesse d'être bien posé à ce moment-là (montrer par exemple la perte d'unicité en comparant avec des solutions constantes).

3) On suppose désormais que $z_B \geq z_C$. On dit que le réservoir est vide lorsque $z = z_B$. Calculer le temps t_B pour lequel le siphon est vidé.

Exercice 1.5 (L'équation logistique)**.** ($\star$) L'équation logistique est un modèle simple empirique d'évolution de la population d'une espèce animale ou végétale (par exemple la prolifération d'un ensemble de bactéries). L'équation logistique est donnée par

$$\dot{x}(t) = ax - bx^2.$$

Ici $x(t)$ représente le nombre de bactéries au temps t. Le terme ax représente la croissance du nombre de bactéries par division cellulaire et le terme $-bx^2$ représente la compétition pour l'accès aux ressources (lorsque ce terme devient dominant on parle de "surpopulation").

1) Montrer que, si on se donne une donnée initiale x_0 positive, cette équation admet une unique solution locale.

2) Calculer les solutions stationnaires. En déduire le caractère globalement bien posé.

3) Calculer la solution exacte à l'aide de la méthode de séparation des variables.

4) Calculer la limite en temps long de la solution et montrer qu'elle converge vers cette limite à vitesse exponentielle.

Exercice 1.6 (Équations séparables et principes de comparaison)**.** ($\star$)

1) On souhaite utiliser la méthode des équations séparables et les principes de comparaison pour montrer que l'équation suivante n'est pas globalement bien posée :

$$\dot{x} = x^2 t + 1, \qquad \text{et} \qquad x(0) = x_0 > 0. \tag{1.52}$$

1.1) Résoudre explicitement par la méthode de séparation l'équation suivante :

$$\dot{y} = y^2 t, \qquad \text{et} \qquad y(0) = x_0 > 0. \tag{1.53}$$

1.2) Calculer $T_y^\star \in]0, +\infty]$ le temps de fin d'existence à cette deuxième équation.

1.3) Montrer que (1.52) admet une unique solution locale $t \mapsto x(t)$ et que celle-ci vérifie pour tout temps : $x(t) \geq y(t)$.

1.4) Montrer que le temps de fin d'existence pour cette deuxième équation, noté $T_x^\star$ vérifie $T_x^\star \leq T_y^\star < +\infty$ et que

$$x(t) \longrightarrow +\infty, \qquad \text{lorsque} \quad t \to T_x^\star.$$

Exercice 1.7 (Le système de Lotka-Volterra)**.** ($\star$) Le système de Lotka-Volterra s'écrit de la façon suivante :

$$\dot{x} = x - xy, \qquad \text{et} \qquad \dot{y} = -y + xy. \tag{1.54}$$

Ce système modélise l'évolution au cours du temps d'une population d'un ensemble de proies (par exemple les sardines dans la mer) représentée par la variable x, et d'une population de prédateurs (par exemple des requins) représentée par y. Il y a tout d'abord un terme d'auto-engendrement : il est positif pour les proies (qui se reproduisent rapidement et se nourrissent de plancton supposé très abondant) et négatif pour les prédateurs qui, pour pouvoir prospérer, doivent trouver à se nourrir. L'autre terme est donc l'interaction entre les deux espèces : cette interaction est négative pour les proies et positive pour les prédateurs.

1) Étant donnée une donnée initiale fixée quelconque $x_0, y_0 \geq 0$, Montrer qu'il existe une unique solution locale à ce problème de Cauchy.

2) Calculer les solutions stationnaires.

3) Montrer que pour tout t on a, $x(t) > 0$ et $y(t) > 0$ dès lors que $x_0, y_0 > 0$.

4) on pose $\psi(t) = \ln x(t) + \ln y(t) - x(t) - y(t)$. Montrer que la fonction ψ est constante.

5) On souhaite montrer le caractère globalement bien-posé de cette équation.
5.1) Étudier la fonction $u \longmapsto \ln(u) - u$.
5.2) En déduire que $\ln(x(t)) - x(t) \geq \psi(0) + 1$. Faire de même pour y.
5.3) Montrer alors que les solutions x et y sont bornées.
5.4) En déduire que le système est globalement bien posé.

6) On divise le premier quadrant $(\mathbb{R}_+^*)^2$ en quatre zones :

$$A = \{(u; v) : 0 < u < 1; 0 < v < 1\}, \qquad B = \{(u; v) : u > 1; 0 < v < 1\},$$

$$C = \{(u; v) : u > 1; v > 1\}, \qquad D = \{(u; v) : 0 < u < 1; v > 1\}.$$

Montrer que si $(x_0; y_0) \neq (1; 1)$, alors la solution passe successivement de A à B, puis à C puis enfin à D avant de revenir dans A.

7) En utilisant la quantité conservée ψ, montrer que la solution est périodique.

8) Montrer que la courbe du plan formée par la trajectoire $t \mapsto (x(t), y(t))$ est convexe. Pour cela montrer que le produit vectoriel du vecteur $(\dot{x}, \dot{y})$ par le vecteur $(\ddot{x}, \ddot{y})$ ne s'annule jamais (faire un dessin).

9) Tracer approximativement les deux fonctions $t \mapsto x(t), y(t)$ et tracer dans le plan les courbes de niveaux de ψ (c'est-à-dire les courbes d'équation $\psi =$ constante).

Exercice 1.8 (Utilisation du lemme de Grönwall)**.** ($\star$) On propose dans cet exercice quelques exemples d'équations pour lesquelles on démontre le caractère globalement bien posé à l'aide du théorème de prolongement et du lemme de Grönwall.

1) Montrer que le problème de Cauchy sur $\mathbb{R}$ (avec $\lambda \in \mathbb{R}$ un paramètre fixé quelconque) admet une unique solution globale :

$$\frac{\mathrm{d}x}{\mathrm{d}t} = x\big(\lambda + \sin(x^2)\big), \qquad \text{et} \qquad x(0) = x_0 \in \mathbb{R},$$

2) Même question avec le problème de Cauchy ($\lambda, \mu \in \mathbb{R}$) :

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \lambda\, x \sin^2\left(\frac{1}{x}\right) + \mu x \cos^2\left(\frac{1}{x}\right), \qquad \text{et} \qquad x(0) = x_0 \in \mathbb{R}^*,$$

3) Même question avec le système de 2 équations suivant (équipé de conditions initiales $x(0) = x_0 \in \mathbb{R}$ et $y(0) = y_0 \in \mathbb{R}$) :

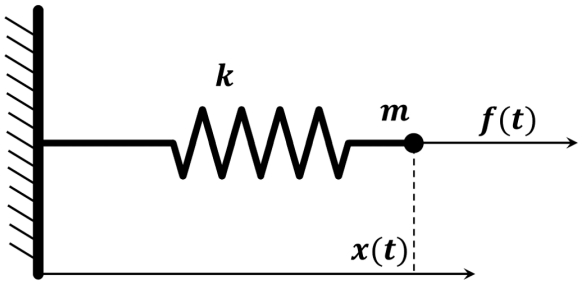
$$\dot{x} = x - y - x^3, \qquad \text{et} \qquad \dot{y} = x + y - y^3.$$

* * *

Exercice 1.9 (L'oscillateur harmonique)**.** L'oscilateur harmonique est le système mécanique le plus simple et modélise l'évolution 1D d'une masse ponctuelle accrochée à un ressort. Ce système est décrit par l'équation d'évolution suivante :

$$m\ddot{x}(t) + \lambda \dot{x} + kx(t) = g(t), \tag{1.55}$$

avec m la masse de l'objet, k la raideur du ressort, λ le coefficient de frottement, $g : [0, +\infty[\to \mathbb{R}$ la force extérieure et $x : [0, +\infty[\to \mathbb{R}$ la position de l'objet (l'inconnue du problème).



1) En introduisant la vitesse $v(t) := \dot{x}(t)$ comme variable auxiliaire du système, montrer que l'équation étudiée est équivalente à

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} x \\ v \end{pmatrix}(t) = A\begin{pmatrix} x \\ v \end{pmatrix}(t) + \begin{pmatrix} f(t) \\ 0 \end{pmatrix},$$

où A est une matrice constante dont on précisera les coefficients.

2) On suppose que g est lipschitzienne. Montrer que si on pose $X(t) := \big(x(t),\, v(t)\big)^T$, alors on peut réécrire cette équation sous la forme

$$\dot{X} = \mathcal{F}\Big(t, X(t)\Big),$$

où $\mathcal{F} : [0, +\infty[\times\mathbb{R}^2$ est une fonction à préciser.

3) A l'aide du théorème de Cauchy-Lipschitz, montrer que ce système admet une unique solution globale étant donné la position initiale $x(0) = x_0$ et la vitesse initiale $v(0) = v_0$ où $x_0, v_0 \in \mathbb{R}$ sont deux paramètres fixés.

4) Calculer χ_A le polynôme caractéristique de la matrice A.

5) Dans le cas où ce polynôme admet 2 racines réelles distinctes $\lambda, \mu < 0$ la matrice est diagonalisable dans $\mathbb{R}$. Montrer alors, en utilisant les exponentielles de matrices dans le cas où $g \equiv 0$ (lemme 1.3), que la solution s'écrit sous la forme :

$$x(t) = ae^{\lambda t} + be^{\mu t}, \qquad \text{et} \qquad v(t) = ce^{\lambda t} + de^{\mu t},$$

avec a, b, c, d des constantes qui dépendent de la diagonalisation de A et des données initiales du problème.

6) Dans le cas où χ_A admet une racine double $\lambda < 0$, la matrice A n'est pas diagonalisable. Montrer que la solution s'écrit :

$$x(t) = ae^{\lambda t} + bte^{\lambda t}, \qquad \text{et} \qquad v(t) = ce^{\lambda t} + dte^{\lambda t},$$

avec a, b, c, d des constantes qui dépendent de la réduction de Jordan de A et des données initiales du problème.

7) Dans le cas où χ_A n'admet pas de racines réelles alors il admet 2 racines complexes conjuguées $\lambda + i\omega$ et $\lambda - i\omega$ avec $\lambda < 0$. la matrice A est alors diagonalisable dans $\mathbb{C}$. Montrer que la solution s'écrit :

$$x(t) = ae^{\lambda t}\cos(\omega t) + be^{\lambda t}\sin(\omega t), \qquad \text{et} \qquad v(t) = ce^{\lambda t}\cos(\omega t) + de^{\lambda t}\sin(\omega t),$$

avec a, b, c, d des constantes qui dépendent de la diagonalisation de A et des données initiales du problème.

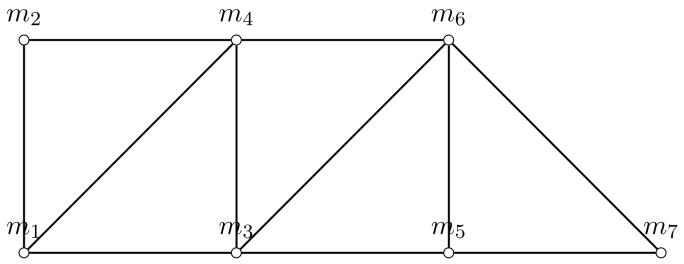
8) A l'aide de la formule exacte pour les équations linéaires (lemme 1.3), résoudre les cas $f(t) = t$ et $f(t) = \cos(\omega_1 t)$.

Exercice 1.10 (Treillis de poutres élastiques)**.** On considère un système de poutres de masses négligeables qui sont connectées les unes aux autres avec des liaisons de type "rotules". On place une masse ponctuelle m_i au niveau de chacune des N connexions de poutres. On note par $x_i \in \mathbb{R}^3$ la position de la i^e masse ponctuelle. On note $\mathcal{N}_i$ l'ensemble des masses ponctuelles voisines de la i^e masse, c'est à dire qu'il existe une poutre qui relie ces deux masses. On note k_{ij} la raideur élastique de cette poutre et ℓ_{ij} sa longueur au repos. Finalement, on note par $\mathcal{A}$ l'ensemble des masses qui sont immobiles (fixées à un mur par exemple).

L'équation de la dynamique de ce système de poutres s'écrit pour tout $i \notin \mathcal{A}$:

$$m_i\,\ddot{x}_i = -\sum_{j\in\mathcal{N}_i} k_{ij}\Big(\ell_{ij} - |x_i - x_j|\Big)e_{ij}, \tag{1.56}$$

où $e_{ij} := \dfrac{x_j - x_i}{|x_j - x_i|}$ est le vecteur unitaire qui pointe en direction de x_j depuis la position x_i. Dans le cas où $i \in \mathcal{A}$, on fixe l'accélération à 0.



1) Montrer que ce système d'équations différentielles admets une unique solution globale étant donné les conditions initiales suivantes

$$x_i(0) = x_{i,0}, \qquad \text{et} \qquad \dot{x}_i(0) = \begin{cases} 0 & \text{si } i \in \mathcal{A}, \\ v_{i,0} & \text{sinon,} \end{cases}$$

où $x_{i,0}$ et $v_{i,0}$ sont des vecteurs de $\mathbb{R}^3$ fixés quelconques.

2) Que se passe-t-il si on rajoute au sommet de chaque poutre une force extérieure qui varie continûment au cours du temps ?

Exercice 1.11. Trouver toutes les solutions globales $\mathcal{C}^1(\mathbb{R}, \mathbb{R})$ pour l'équation $t^2\,\dot{x} = x$. Pour quelles données initiales $x(0) = x_0 \in \mathbb{R}$ a-t-on existence ? unicité ?

Exercice 1.12. Trouver les données initiales pour lesquelles l'équation $\dot{x} = x^2 \ln\left(\frac{x^2+1}{2}\right)$ est globalement bien posée sur $\mathbb{R}$. Même question avec $\dot{x} = x + (x - e^t)(x - e^{-t})$.

Exercice 1.13. On considère le problème de Cauchy suivant :

$$\dot{x} \;=\; t^3 x - 2\sin(t)\ln\left(x^4+1\right) + 1, \qquad \text{et} \qquad x(0) = 0 \tag{1.57}$$

1) Montrer que ce problème admet une unique solution locale.

2) Montrer que cette solution est positive pour tout temps.

3) Montrer que si $x(t) \geq 1$ sur un intervalle de temps $[t_0, t_1)$ alors

$$\forall\, t \in [t_0, t_1), \qquad x(t) \;\leq\; x(t_0)\,\exp\left(\frac{t^4}{4} + t\right). \tag{1.58}$$

En déduire l'existence globale de la solution.

4) Montrer que $x(t) \longrightarrow +\infty$ lorsque $t \to +\infty$.

5) Établir le développement asymptotique suivant : $\ln\big(x(t)\big) \simeq t^4$ lorsque $t \to +\infty$.

Exercice 1.14. Donner la CNS sur $\alpha \in \mathbb{R}$ pour avoir $\ddot{x} = x^\alpha$ globalement bien posé sur tous les temps positifs sachant que $x(0)$ et $\dot{x}(0)$ sont strictement positives (*indication* : multiplier par $\dot{x}$ et intégrer en temps). Étudier le comportement asymptotique au voisinage du temps de fin d'existence de la solution (qu'il soit fini ou infini).

Exercice 1.15 (Équations de la gravitation universelle de Newton)**.** On considère N masses ponctuelles $m_1, \dots, m_N$ situées au temps initial aux positions respectives $X_1, \dots, X_N \in \mathbb{R}^3$ et allant à la vitesse initiale $V_1, \dots, V_N$. Les équations de la gravitation universelle de Newton sont données par

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} X_i \;=\; \mathcal{G} \sum_{\substack{j=1 \\ j \neq i}}^{N} \frac{m_i\, m_j}{|X_i - X_j|^2}\, e_{ij}, \tag{1.59}$$

où e_{ij} est le vecteur unitaire allant vers X_j depuis la position X_i et $\mathcal{G}$ est la constante de Newton.

1) Montrer que ce problème est localement bien posé sur l'ouvert

$$\Omega \;:=\; \Big\{ (X_1, \dots, X_N, V_1, \dots, V_N) \in (\mathbb{R}^3)^N \times (\mathbb{R}^3)^N \;:\; \forall\, i \neq j, \quad X_i \neq X_j \Big\}. \tag{1.60}$$

2) Proposer un contre-exemple à l'existence globale.

Exercice 1.16 (Lemme de Grönwall amélioré)**.**

1) Calculer la dérivée de $s \mapsto \ln(\ln(s))$ et résoudre le problème de Cauchy

$$\frac{\mathrm{d}}{\mathrm{d}t} x = x \ln(x), \qquad \text{et} \qquad x(0) = x_0 > 0.$$

2) Montrer le lemme de Grönwall amélioré (lemme de Grönwall-Osgood) :

$$\frac{\mathrm{d}\phi}{\mathrm{d}t}(t) \;\leq\; \psi(t)\, \phi(t) \ln\big(\phi(t)\big) \quad \Longrightarrow \quad \phi(t) \leq \phi(t_0) \exp\left(\exp\left(\int_a^t \psi(s)\, \mathrm{d}s \right) \right).$$

3) On considère à présent un problème de Cauchy $\dot{X} = \mathcal{F}(t, X)$ avec $\mathcal{F}$ défini sur $\mathbb{R} \times \mathbb{R}^d$ de classe $\mathcal{C}^1$ et avec la donnée initiale $X(0) = X_0$. Montrer que si la solution vérifie

$$\left| \frac{dX}{\mathrm{d}t} \right| \lesssim |X| \Big(1 + \big| \ln(X) \big| \Big),$$

alors la solution est globale en temps.

4) Calculer la dérivée de $x \mapsto \ln(\ln(\ln(x)))$ et de $x \mapsto \ln(\ln(\ln(\ln(x))))$. Proposer une généralisation des résultats démontré ci-dessus.

Exercice 1.17 (Collisions de points-vortex)**.** (illustrations de C. Marchioro et M. Pulvirenti). Le système de points-vortex est un système d'équations différentielles issues de la mécanique des fluides incompressibles du plan (équations de Euler 2D ou équations atmosphériques en régime fortement stratifié). Ce système rend compte de la dynamique des centres des tourbillons présents dans un fluide planaire en faisant l'approximation que la vorticité du fluide est très concentrée. Chaque vortex est représenté par sa position dans le plan $z_j(t) \in \mathbb{C}$ pour $j = 1, \dots, N$ et son intensité $a_j \in \mathbb{R}^*$ (celle-ci est constante au cours du temps). L'évolution de la position du vortex $z_j \in \mathbb{C}$ au cours du temps est donné par (avec la notation $i^2 = -1$.) :

$$\frac{\mathrm{d}}{\mathrm{d}t} z_j = \sum_{k=1}^{N} a_j \frac{i(z_j - z_k)}{|z_j - z_k|^2}, \tag{1.61}$$

1) Existence et unicité des solutions :

1.1) Monter l'existence et l'unicité locale d'une solution à ce système sur l'ouvert

$$\Omega := \bigcap_{j\neq k} \Big\{ Z = (z_1, \dots, z_N) \in \mathbb{C}^N \Big/ \, z_j \neq z_k \Big\}.$$

1.2) Montrer que le temps maximal d'existence $T^\star$ est fini ssi on a une collision de vortex :

$$\liminf_{t\to T^-} \ \min_{j\neq k} \ \big| z_j(t) - z_k(t) \big| \ = \ 0.$$

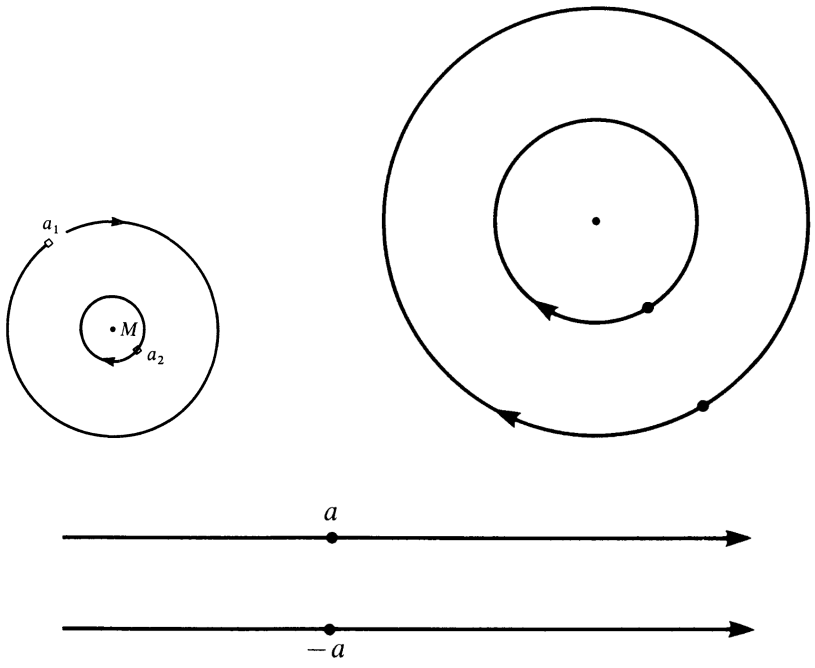
1.3) Montrer que les quantités suivantes sont constantes au cours du temps :

$$B(Z) := \sum_{j=1}^{N} a_j \, z_j, \qquad \text{et} \qquad L(Z) := \sum_{j=1}^{N} \sum_{\substack{k=1 \\ k\neq j}}^{N} a_j \, a_k \, \big| z_j - z_k \big|^2.$$

1.4) Montrer que le barycentre $M(Z) := B(Z) / \sum_{j=1}^{N} a_j$ est constant au cours du temps.

1.5) Dans le cas $N = 2$ montrer que la préservation de $L(Z)$ implique la préservation de la distance entre les 2 vortex (et donc l'existence globale).

1.6) En déduire la solution exacte du système pour $N = 2$. Il faut pour cela distinguer les cas selon les valeurs de a_1 et a_2 (les 3 cas possibles correspondent aux 3 illustrations ci-dessous).



2) Pour $N = 2$, on a montré que le système était globalement bien posé. On va montrer que pour $N \geq 3$ il existe des données initiales qui aboutissent à des collisions en temps fini. On va plus précisément s'intéresser aux collisions auto-similaires. Dans un premier temps on va extraire les conditions nécessaires associées à ces solutions particulières pour mieux comprendre la collision. On appelle "*collision auto-similaire*" une solution du problème de points-vortex de la forme

$$z_j : t \in [0, T[, \longmapsto z_j(0) \, r(t) \, e^{i\theta(t)}, \tag{1.62}$$

pour tout $j = 1, \dots, N$, où les fonctions r et θ sont continues et vérifient

$$r(0) = 1, \quad r(T) = 0, \quad \text{et} \quad \theta(0) = 0.$$

2.1) Expliquer l'appellation "collision auto-similaire" à partir d'une telle définition.

2.2) Montrer que la condition d'auto-similarité implique que pour tout $j \neq k$

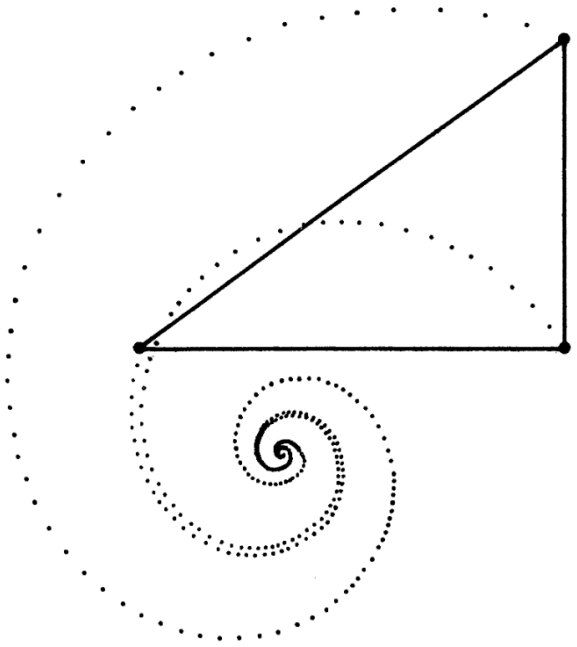
$$\frac{\mathrm{d}}{\mathrm{d}t}\big|z_j(t) - z_k(t)\big|^2 = 2\,\big|z_j(0) - z_k(0)\big|^2\,\dot{r}(t)\,r(t).$$

2.3) En utilisant l'équation des points-vortex et le résultat de la question précédente, en déduire que nécessairement $t \mapsto r(t)$ vérifie $\dot{r}(t) = C/r$ où C est une constante qui dépend des données initiales du problème (à expliciter).

2.4) En utilisant la séparation des variables, montrer que nécessairement $r(t) = \sqrt{\dfrac{T-t}{T}}$.

2.5) Montrer également qu'il existe une constante D telle que $\theta(t) = -DT\ln\left(\dfrac{T-t}{T}\right)$.

2.6) Pour le système de points-vortex, on parle souvent de "*collision en spirale*" (voir illustration ci-dessous). Justifier cette terminologie au regard des résultats des deux questions précédentes.



3) On va maintenant construire la collision auto-similaire (solution du système point-vortex) qui est représentée sur l'illustration ci-dessus. On va donc travailler avec $N = 3$. Plus précisément, on pose $a_2 = a_3 = 1$ et $a_1 = a \in \mathbb{R}^*$ un paramètre fixé plus tard. On pose $A = |x_2 - x_3|$, $B = |x_3 - x_1|$ and $C = |x_1 - x_2|$. Soit $\lambda \in]0,1[$. On va choisir les valeurs de $x_1(0)$, $x_2(0)$ et $x_3(0)$ tel que ces trois points forment un triangle orthogonal direct avec $A(0) = 1$, $B(0) = \lambda$ et $C(0) = \sqrt{\lambda^2+1}$ (par Pythagore, ceci définit bien un triangle rectangle). Nous voulons maintenant trouver une valeur pour λ et a telles que $B(t) = \lambda A(t)$ et $C(t) = \sqrt{\lambda^2+1}A(t)$ soient vraies pour tout temps t, et telles que $A(t) \to 0$ lorsque $t \to T$ pour un certain $T > 0$.

1) Pour $N = 3$, les équations peuvent être réécrites en utilisant $\triangle$, l'aire du triangle direct (x_1, x_2, x_3). Rappel sur le calcul de l'aire d'un triangle :

$$(x_2 - x_3)\cdot(x_2 - x_1)^\perp = (x_3 - x_1)\cdot(x_3 - x_2)^\perp = (x_1 - x_2)\cdot(x_1 - x_3)^\perp = -2\triangle. \qquad (1.63)$$

En utilisant l'équation des points-vortex, montrer que

$$\frac{\mathrm{d}}{\mathrm{d}t}A^2 = 4a\triangle\left(\frac{1}{B^2}-\frac{1}{C^2}\right), \qquad \frac{\mathrm{d}}{\mathrm{d}t}B^2 = 4\triangle\left(\frac{1}{C^2}-\frac{1}{A^2}\right) \qquad \text{et} \qquad \frac{\mathrm{d}}{\mathrm{d}t}C^2 = 4\triangle\left(\frac{1}{A^2}-\frac{1}{B^2}\right). \tag{1.64}$$

2) Pour avoir le caractère auto-similaire, il faut propager la propriété $B(t) = \lambda A(t)$ et $C(t) = \sqrt{\lambda^2+1}\,A(t)$. En utilisant la question précédent, montrer que ceci équivaut à

$$4\triangle\left(\frac{1}{C^2}-\frac{1}{A^2}\right) = 4a\triangle\lambda\left(\frac{1}{B^2}-\frac{1}{C^2}\right), \qquad \text{et} \qquad 4\triangle\left(\frac{1}{A^2}-\frac{1}{B^2}\right) = 4a\triangle\sqrt{\lambda^2+1}\left(\frac{1}{B^2}-\frac{1}{C^2}\right).$$

On montrera par exemple que c'est équivalent à la préservation au cours du temps du rapport A/B et du rapport A/C.

3) En utilisant le fait que $B(t) = \lambda A(t)$ et $C(t) = \sqrt{\lambda^2+1}\,A(t)$, montrer qu'une telle condition équivaut à

$$\left(\frac{1}{\lambda^2+1}-1\right) = a\lambda\left(\frac{1}{\lambda^2}-\frac{1}{\lambda^2+1}\right), \qquad \text{et} \qquad \left(1-\frac{1}{\lambda^2}\right) = a\sqrt{\lambda^2+1}\left(\frac{1}{\lambda^2}-\frac{1}{\lambda^2+1}\right).$$

4) Résoudre ce système de 2 équations à 2 inconnues λ et a et conclure à l'existence d'une collision auto-similaire pour $N = 3$ point-vortex.

*
* *

Chapitre 2

Analyse asymptotique et stabilité des solutions

Dans le chapitre précédent, nous avons proposé quelques outils simples pour faire des analyses asymptotiques du comportement des solutions, notamment à l'aide de principes de comparaisons ou du lemme de Grönwall.

Au cours de ce chapitre nous allons poursuivre cette étude du comportement asymptotique des solutions avec des outils plus puissants, notamment la théorie développée par Alexander Lyapunov et d'autres mathématiciens au début du XX^e siècle. Le grand intérêt de tous ces outils concerne l'analyse de la stabilité des solutions aux équations. C'est-à-dire : étant donné deux valeurs initiales différentes mais proches l'une de l'autre, peut-on savoir si elles vont rester proches au cours du temps ou non ?

Pour des raisons pédagogiques, nous étudierons la théorie de la stabilité des EDO dans le cadre des EDO autonomes (c'est-à-dire $\mathcal{F}$ ne dépend pas du temps). De nombreux concepts et théorèmes s'étendent au cas non-autonome.

2.1 Quantités conservées et dissipées

2.1.1 Définitions et premières propriétés

Pour l'étude de la stabilité et du comportement asymptotique des solutions, une étude incontournable est celle des quantités conservées par l'équation. L'étude des systèmes physiques abonde d'exemples montrant l'utilité des quantités telles que l'énergie mécanique, le moment cinétique, etc... On va étudier ici une formulation générale de ce type d'outils.

On considère, pour $\mathcal{F} : \Omega \to \mathbb{R}^d$ et $X_0 \in \mathbb{R}^d$ fixés (avec Ω ouvert de $\mathbb{R}^d$) le problème de Cauchy suivant :

$$\frac{dX}{\mathrm{d}t}(t) \;=\; \mathcal{F}\Big(X(t)\Big), \qquad \text{et} \qquad X(0) = X_0. \tag{2.1}$$

Afin d'alléger les notations mais surtout pour exprimer la dépendance de la solution $X(t)$ par rapport à la donnée initiale, on utilisera abondamment la notation "*flot*" :

Définition 2.1 Nouvelle notation : le flot (†)

Étant donné une fonction $\mathcal{F}$ fixée quelconque (autonome en temps), on note $\mathscr{S}^t X$ la solution au temps t du problème de Cauchy (2.1) avec la donnée initiale $X \in \Omega$. Autrement-dit : $\mathscr{S}^t X_0 := X(t)$.

Si la fonction $\mathcal{F}$ n'est pas autonome en temps, on précise alors le temps initial t_0 à partir duquel on travaille (si $t_0 \neq 0$) pour pouvoir définir le flot $\mathscr{S}^t_{t_0}$.

Remarque : Dans le cas où le problème de Cauchy (2.1) est globalement bien posé pour tout temps $t \in \mathbb{R}$, le théorème de Cauchy-Lipschitz nous assure que le flot $\mathscr{S}^t_{t_0} : \Omega \to \mathbb{R}^d$ est une application bien définie et injective. On peut aisément démontrer que le flot admet une structure de semi-groupe engendré par le semi-morphisme associé à la relation : $\mathscr{S}^{t_2}_{t_1} \circ \mathscr{S}^{t_1}_{t_0} = \mathscr{S}^{t_2}_{t_0}$. L'élément neutre de ce semi-groupe est $\mathcal{I}_\Omega$, la fonction identité sur Ω. On peut également démontrer (admis) que pour tout t fixé, le flot $\mathscr{S}^t_{t_0} : X \in \Omega \longmapsto \mathscr{S}^t_{t_0} X$ est une fonction $\mathcal{C}^k$ à réciproque $\mathcal{C}^k$ (on dit alors que $\mathscr{S}^t$ est un $\mathcal{C}^k$-*difféomorphisme*) pourvu que le champs de vitesse $\mathcal{F}$ soit assez régulier. Dans le cadre de ce chapitre où $\mathcal{F}$ est autonome en temps, le flot admet une structure de groupe (dont le neutre est la fonction identité) engendré par le morphisme $\mathscr{S}^t \circ \mathscr{S}^\tau = \mathscr{S}^{t+\tau}$.

Remarque 2 : Si on a seulement existence locale, alors le flot n'est pas nécessairement bien-défini au voisinage du bord de Ω puisque le théorème de fin d'existence nous dit que l'équation cesse d'être bien posé lorsque la trajectoire du point étudié atteint $\partial\Omega$. Il est en revanche bien défini au voisinage de tout point intérieur de $X_0 \in \Omega$ sur un petit intervalle de temps (qui va dépendre a priori du choix de X_0).

Définition 2.2 Quantités conservées et dissipées (†)

Soit $\mathcal{G} : \Omega \to \mathbb{R}$ de classe $\mathcal{C}^1$ et soit $\mathscr{S}^t$ le flot de l'équation (2.1).

(i) On dit que $\mathcal{G}$ est une *quantité conservée* ssi pour tout $X \in \Omega$,

$$\forall\, t \in [0, T^\star[, \qquad \frac{\mathrm{d}}{\mathrm{d}t}\mathcal{G}\big(\mathscr{S}^t X\big) = 0. \tag{2.2}$$

(ii) On dit que $\mathcal{G}$ est une *quantité dissipée* ssi pour tout $X \in \Omega$,

$$\forall\, t \in [0, T^\star[, \qquad \frac{\mathrm{d}}{\mathrm{d}t}\mathcal{G}\big(\mathscr{S}^t X\big) \leq 0. \tag{2.3}$$

Si l'inégalité ci-dessus est stricte, on parle alors de *dissipation stricte*.

Les quantités conservées ou dissipées donnent de nombreuses informations précieuses sur la dynamique. En effet, si on démontre qu'une quantité $\mathcal{G}$ est conservée, cela nous dit que le point $X(t)$ va rester à l'intérieur d'une même composante connexe d'un ensemble de niveau de $\mathcal{G}$. Autrement dit, la dynamique se fait sur l'une des composantes connexes de l'hyper-surface d'équation $\mathcal{G}(X) = \textit{constante}$ (où la constante est déterminée à l'aide de la donnée initiale).

Si en revanche cette quantité est dissipée alors on en déduit que $X(t)$ va rester à l'intérieur d'une composante connexe de l'ensemble de sous-niveau $\{X \in \Omega : \mathcal{G}(X) \leq \textit{constante}\}$. Dans le cas où $\mathcal{G}$ admet des propriété de coercivité (voir définition 0.22), on peut montrer la propriété suivante :

Lemme 2.1 Dissipation de quantités coercives

Soit $\mathcal{G} : \Omega \to \mathbb{R}$ de classe $\mathcal{C}^1$ une quantité conservée ou dissipée par (2.1). Si $\mathcal{G}$ est coercive en $X \in \Omega$ alors la solution $t \mapsto \mathscr{S}^t X$ est globalement bien défini.

Démonstration. Si $\mathcal{G}$ est une quantité conservée ou dissipée, alors $t \mapsto \mathcal{G}(\mathscr{S}^t X)$ est décroissante par définition. Donc, pour tout $t \geq 0$ on a $\mathscr{S}^t X$ qui appartient à $\{Y \in \Omega : \mathcal{G}(Y) \leq \mathcal{G}(X)\}$.

Plus précisément, en vertu du Lemme 0.1, on a $\mathscr{S}^t X$ qui appartient à la même composante connexe que X. Or cette composante connexe est compacte par hypothèse de coercivité de $\mathcal{G}$ en X. Donc, par théorème de prolongement (Corollaire 1.1), on en déduit que la solution est globale. □

2.1.2 Le cadre des équations linéaires

Dans le cadre des équations différentielles linéaires à coefficients constants sans second membre, la résolution explicite à l'aide des exponentielles de matrices (lemme 1.3) permet d'obtenir aisément des quantités conservées et ainsi appréhender la dynamique obtenue. On considère, pour A une matrice constante de taille d le problème de Cauchy

$$\frac{dX}{\mathrm{d}t}(t) \;=\; AX(t), \qquad \text{et} \qquad X(0) = X_0. \tag{2.4}$$

Cas de deux valeurs propres réelles

Dans le cadre de ce paragraphe, on suppose que A admet 2 valeurs propres réelles λ et μ et on note $X_\lambda \in \mathbb{R}^d$ et $X_\mu \in \mathbb{R}^d$ respectivement un vecteur propre associé. La résolution explicite de l'équation différentielle par les exponentielles de matrices implique que

$$\mathscr{S}^t X_\lambda \;=\; e^{\lambda t} X_\lambda, \qquad \text{et} \qquad \mathscr{S}^t X_\mu \;=\; e^{\mu t} X_\mu. \tag{2.5}$$

De même, si on considère la projection de la dynamique générale sur les sous-espaces propres associés, nous avons

$$(\mathscr{S}^t X)^T X_\lambda \;=\; e^{\lambda t} X^T X_\lambda, \qquad \text{et} \qquad (\mathscr{S}^t X)^T X_\mu \;=\; e^{\mu t} X^T X_\mu. \tag{2.6}$$

Par conséquent, nous avons l'égalité suivante :

$$\frac{|(\mathscr{S}^t X)^T X_\mu|^\lambda}{|(\mathscr{S}^t X)^T X_\lambda|^\mu} \;=\; \frac{|e^{\mu t} X^T X_\mu|^\lambda}{|e^{\lambda t} X^T X_\lambda|^\mu} \;=\; \frac{|X^T X_\mu|^\lambda}{|X^T X_\lambda|^\mu}. \tag{2.7}$$

On en déduit donc la quantité conservée suivante :

$$\frac{\mathrm{d}}{\mathrm{d}t}\, \frac{|(\mathscr{S}^t X)^T X_\mu|^\lambda}{|(\mathscr{S}^t X)^T X_\lambda|^\mu} \;=\; 0. \tag{2.8}$$

Lorsque λ et μ sont de même signes, alors la quantité conservée nous donne des équations de *pseudo-paraboles*. Le système évolue alors à l'intérieur de ses pseudo-paraboles en convergeant vers 0 si $\lambda < 0$ et $\mu < 0$ et au contraire en divergeant si $\lambda > 0$ et $\mu > 0$. Lorsque ces valeurs propres sont de signe opposés, alors les équations sur cette quantité conservée donne des *pseudo-hyperboles*. La dynamique est toujours divergente le long de ces pseudo-hyperboles (voir Section 2.4.2 pour plus de détails).

Cas de deux valeurs propres imaginaires conjuguées

Si en revanche nous avons deux valeurs propres imaginaires $\mathrm{i}\omega$ et $-\mathrm{i}\omega = \overline{\mathrm{i}\omega}$ équipées de vecteur propres respectivement $X_{\mathrm{i}\omega} \in \mathbb{C}^d$ et $X_{\overline{\mathrm{i}\omega}} \in \mathbb{C}^d$ alors nous avons

$$\mathscr{S}^t X_{\mathrm{i}\omega} \;=\; e^{+\mathrm{i}\omega t} X_{\mathrm{i}\omega}, \qquad \text{et} \qquad \mathscr{S}^t X_{\overline{\mathrm{i}\omega}} \;=\; e^{-\mathrm{i}\omega t} X_{\overline{\mathrm{i}\omega}}. \tag{2.9}$$

Il est alors possible de réécrire ces solutions comme une combinaison linéaire de fonctions $t \mapsto \sin(\omega t)$ et $t \mapsto \cos(\omega t)$ en utilisant la formule d'Euler - De Moivre (†) donnée par :

$$\forall\, \theta \in \mathbb{R}, \qquad \cos(\theta) \;=\; \frac{e^{\mathrm{i}\theta} + e^{-\mathrm{i}\theta}}{2} \qquad \text{et} \qquad \sin(\theta) \;=\; \frac{e^{\mathrm{i}\theta} - e^{-\mathrm{i}\theta}}{2\mathrm{i}}. \tag{2.10}$$

Cette transformation d'Euler - De Moivre nous donne deux vecteurs $Y_\omega \in \mathbb{R}^d$ et $Z_\omega \in \mathbb{R}^d$ tels que

$$\mathscr{S}^t Y_\omega \;=\; \cos(\omega t) Y_\omega + \sin(\omega t) Z_\omega, \qquad \text{et} \qquad \mathscr{S}^t Z_\omega = -\sin(\omega t) Y_\omega + \cos(\omega t) Z_\omega. \tag{2.11}$$

Autrement dit : la dynamique dans le plan engendré par Y_ω et Z_ω se fait le long d'ensemble définis par des équations quadratiques définissant des ellipses. Il est possible d'obtenir les équations de ces ellipses par une étude algébrique directe présentée Section 2.4.2 ou en utilisant la proposition ci-dessous :

PROPOSITION 2.1 Quantités conservées quadratiques

Soit M une matrice symétrique de taille d. Alors ces propositions sont équivalentes :

- La quantité $X \longmapsto X^T M X$ est conservée par l'équation.
- La matrice MA est antisymétrique : $\quad (MA)^T + MA = 0$.

La démonstration de cette proposition s'obtient directement en dérivant $X^T M X$ au cours du temps et en utilisant la symétrie de M :

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} X^T M X &= \dot{X}^T M X + X^T M \dot{X} = (AX)^T M X + X^T M A X = X^T A^T M X + X^T M A X \\ &= X^T \big((MA)^T + (MA) \big) X \end{aligned} \tag{2.12}$$

Le principal intérêt de cette proposition est que la recherche d'une quantité conservée (c'est-à-dire la recherche de M) se réduit à la résolution d'un système linéaire : $(MA)^T + MA = 0$. On peut même préciser :

PROPOSITION 2.2 Critère de Lyapunov

(*i*) La quantité $\frac{1}{2} X^T M X$ est croissante au cours du temps pour toute donnée initiale $X_0 \neq 0$ ssi la matrice symétrique $(MA)^T + MA$ est une matrice positive.

(*ii*) Cette quantité est strictement croissante pour toute donnée initiale $X_0 \neq 0$ ssi $(MA)^T + MA$ est defini-positive.

Autres cas

Dans le cas où nous avons deux valeurs propres complexes conjuguées $\lambda \pm \mathrm{i}\omega$, la dynamique dans le plan engendré par les vecteurs propres est celle de spirales logarithmiques elliptiques (convergentes ou divergentes selon le signe de λ). Ces spirales sont les ensembles qui, à l'aide des coordonnées polaires elliptiques, s'écrivent :

$$\left\{ (x,y) \in \mathbb{R}^2 \;:\; r^2 := \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{2xy}{c^2}, \quad r \;=\; C\, e^{\frac{\theta}{\theta_0}} \right\}, \tag{2.13}$$

où θ désigne l'argument pour la réécriture en équation polaire. Les paramètres a, b, c permettant d'écrire le rayon elliptique $r > 0$, ainsi que les paramètres C et θ_0 sont donnés par le changement de base associé aux vecteurs propres et par les données initiales. L'équation de l'ellipse qui engendre la spirale logarithmique est donnée par le critère de Lyapunov ci-dessus. Le paramètre de divergence de la spirale est donné par λ. Pour plus de détails, voir Section 2.4.2.

Dans le cas non-diagonalisable, les équations pour les quantités conservées sont plus difficiles à écrire en général (hors programme).

2.2 Le cadre Hamiltonien

2.2.1 Définition de la dynamique hamiltonienne

Le cadre hamiltonien est un cadre dans lequel il est aisé d'obtenir des quantités conservées. C'est notamment le cadre de la mécanique classique dans le régime conservatif (c'est-à-dire non-dissipatif) et c'est dans ce cadre que l'on obtient les lois de conservation standard de la mécanique classique. On va écrire ici quelques éléments essentiels.

Définition 2.3 Système hamiltonien (†)

On travaille en dimension paire $d = 2k$. Soit $\mathcal{H} : \Omega \subseteq \mathbb{R}^d \to \mathbb{R}$ de classe $\mathcal{C}^1$. On dit que l'équation (1.3) est *hamiltonienne*, et on dit que $\mathcal{H}$ est son Hamiltonien ssi, quitte à faire un changement de base, on a :

$$\forall\, t \in [0, T^\star[, \qquad \frac{\mathrm{d}}{\mathrm{d}t} X = - \begin{pmatrix} 0 & -I_k \\ I_k & 0 \end{pmatrix} \nabla \mathcal{H}\big(X(t)\big), \tag{2.14}$$

où I_k désigne la matrice identité de taille k.

La notion de *Hamiltonien* est une version plus générale et plus "mathématique" de la notion d'énergie bien connue en physique. On définit la matrice J_k (simplement notée J s'il n'y a pas d'ambiguïté) :

$$J_k := \begin{pmatrix} 0 & -I_k \\ I_k & 0 \end{pmatrix}.$$

Cette matrice J a 2 propriétés algébriques très importantes :

- La matrice J est *anti-symmétrique* : $J^T = -J$.
- Elle est dans le *groupe spécial orthogonal* : $J^T = J^{-1}$ et $\det(J) = 1$.

Une conséquence directe est son caractère *anti-involutif* : $J^{-1} = -J$. Son polynôme annulateur minimal est $X^2 + 1$ et ses valeurs propres sont $\pm i$. Lorsque $k = 1$, on retrouve la célèbre matrice de rotation d'angle $\pi/2$ dans le plan $\mathbb{R}^2$ (c'est-à-dire la multiplication par $+i$ dans $\mathbb{C}$). Pour $k \geq 2$, la matrice J_k se comprend comme la généralisation multi-dimensionnelle de la rotation d'angle $\pi/2$. Plus précisément, on peut aisément réorganiser les vecteurs de la base canonique (changement de base orthogonale) en montrer que J_k est ortho-semblable à

$$\Gamma_k := \begin{pmatrix} J_1 & & & \\ & J_1 & & (0) \\ & & \ddots & \\ & (0) & & J_1 \end{pmatrix}.$$

Il suffit pour cela, si on part de la base canonique $(e_1, \ldots, e_k, e_{k+1}, \ldots, e_{2k})$ de réécrire la matrice J_k dans la base $(e_1, e_{k+1}, e_2, e_{k+2}, \ldots, e_k, e_{2k})$. Il s'agit d'une diagonalisation par blocs de rotation.

2.2.2 Lien avec la mécanique classique en régime conservatif

La mécanique classique pour la dynamique du point (ou d'un ensemble de points) se laisse réécrire sous la forme d'une dynamique hamiltonienne lorsque le système est conservatif. on

considère une équation de la mécanique classique pour une inconnue $t \mapsto X(t) \in \mathbb{R}^k$ avec une force conservative, c'est-à-dire qui dérive d'un potentiel $\mathcal{P} : \mathbb{R}^k \to \mathbb{R}$ de classe $\mathcal{C}^1$:

$$\ddot{X} = -\nabla \mathcal{P}(X). \tag{2.15}$$

Pour faire apparaître le hamiltonien, on commence par reformuler cette équation pour la mettre sous la forme de Cauchy. On introduit pour cela le vecteur vitesse $V(t) := \dot{X}(t)$. On a

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} X \\ V \end{pmatrix} \;=\; \begin{pmatrix} V \\ -\nabla \mathcal{P}(X) \end{pmatrix} \;=\; -J \nabla \mathcal{H}\big(X, V\big), \tag{2.16}$$

avec comme choix du hamiltonien : la célèbre énergie mécanique ! Elle est définie par :

$$E_M = \mathcal{H}(X, V) \;:=\; \frac{|V|^2}{2} + \mathcal{P}(X). \tag{2.17}$$

Lorsque le Hamiltonien se met sous la forme d'une énergie mécanique comme ci-dessus (quitte à faire un changement de base) on dit qu'il s'agit d'une ***équation de la mécanique hamiltonienne***. L'espace $\mathbb{R}^k \times \mathbb{R}^k$ dans lequel évolue le vecteur $(X, V)^T$ est souvent appelé ***l'espace des phases***.

Un exemple très important d'équation de la mécanique hamiltonienne est le cas particulier des équations de la ***mécanique hamiltoniennes linéaires***. Dans le cadre des équations linéaires, c'est-à-dire lorsque le potentiel $\mathcal{P}$ est quadratique, la dynamique s'étudie directement en utilisant les exponentielles de matrices (voir la définition 0.19). Lorsque $\mathcal{P}$ est une fonction quadratique défini-positive, on parle alors d'un ***oscillateur harmonique généralisé.***

2.2.3 Propriétés des équations hamiltoniennes

Le principal intérêt des systèmes hamiltoniens réside dans le fait qu'on peut aisément obtenir des quantités conservées.

Théorème 2.1 Préservation du hamiltonien

Si une équation différentielle est hamiltonienne alors son hamiltonien est une quantité conservée par la dynamique.

Démonstration. En utilisant la dérivation des fonctions composées :

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{H}\big(X(t)\big) = \nabla \mathcal{H}\big(X(t)\big) \cdot \frac{\mathrm{d}X}{\mathrm{d}t} = -\nabla \mathcal{H}\big(X(t)\big)^T \begin{pmatrix} 0 & -I_k \\ I_k & 0 \end{pmatrix} \nabla \mathcal{H}\big(X(t)\big) = 0,$$

où pour la dernière inégalité on a utilisé le fait général suivant : $Y^T J Y = 0$ pour tout vecteur Y dès que la matrice J est antisymétrique. □

Remarque : Dans le cadre des systèmes dynamiques *conservatifs*, c'est-à-dire de la forme $\dot{X} = -M \nabla \mathcal{H}(X)$ avec M une matrice anti-symétrique quelconque, nous avons également préservation de la quantité $\mathcal{H}$ par la même démonstration.

THÉORÈME 2.2 Invariance par translation

On suppose qu'il existe $X_0 \in \mathbb{R}^d$ tel que pour tout $X \in \Omega$ et pour tout $\lambda \in \mathbb{R}$ on ait $X + \lambda X_0 \in \Omega$ et

$$\mathcal{H}(X) = \mathcal{H}(X + \lambda X_0).$$

Dans ce cas, la dynamique hamiltonienne associée à $\mathcal{H}$ stabilise tous les hyperplans affines orthogonaux au vecteur JX_0. Ceci revient à la propriété de conservation suivante :

$$\frac{\mathrm{d}}{\mathrm{d}t}(JX_0)^T X(t) = 0. \tag{2.18}$$

Démonstration. Puisque $\mathcal{H}$ est invariant par translation dans la direction X_0, on a la dérivée directionnelle de $\mathcal{H}$ dans la direction X_0 qui est nulle :

$$0 = \frac{\mathcal{H}(X + \lambda X_0) - \mathcal{H}(X)}{\lambda} \quad \longrightarrow \quad \nabla\mathcal{H}(X)^T X_0, \quad \text{lorsque } \lambda \to 0^+.$$

Par conséquent, puisque $(JX_0)^T = X_0^T J^T$ et $J^T = J^{-1}$:

$$\frac{\mathrm{d}}{\mathrm{d}t}(JX_0)^T X(t) = (JX_0)^T \dot{X}(t) = -(JX_0)^T J \nabla\mathcal{H}(X) = -\nabla\mathcal{H}(X)^T X_0 = 0.$$

□

On va à présent énoncer le théorème de conservation des quantités quadratiques associées aux propriétés d'invariance du hamiltonien par rotation. Les idées sont très proches de la démonstration précédente mais leur mise en oeuvre nécessite quelques outils techniques. Pour $\Theta = (\theta_1, \ldots, \theta_k) \in \mathbb{R}^k$ on définit la matrice de rotation généralisée de taille $2k$ par la formule suivante :

$$\overline{R}_\Theta := \begin{pmatrix} R_{\theta_1} & & & (0) \\ & R_{\theta_2} & & \\ & & \ddots & \\ (0) & & & R_{\theta_k} \end{pmatrix}$$

avec la notation R_θ qui désigne la matrice 2×2 associée à la rotation du plan d'angle θ. On rappelle que les blocs de rotation d'angle θ s'écrivent

$$R_\theta := \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \tag{2.19}$$

On peut vérifier par exemple que $R_{\frac{\pi}{2}} = J_1$. Nous aurons également besoin des *matrices de cisaillement* (la généralisation des matrices anti-symétriques particulières Γ_k), définies par :

$$\Gamma_\Theta := \begin{pmatrix} \theta_1 J_1 & & & (0) \\ & \theta_2 J_1 & & \\ & & \ddots & \\ (0) & & & \theta_k J_1 \end{pmatrix}.$$

THÉORÈME 2.3 Invariance par rotation

Soit $\Theta \in \mathbb{R}^k$ et soit une matrice O spéciale orthogonale. On considère la matrice $M_\Theta := O\overline{R}_\Theta O^T$ (qui est elle-même une matrice spéciale orthogonale). On suppose que pour tout $X \in \Omega$ et pour tout $s \in \mathbb{R}$ on a $M_{s\Theta}X \in \Omega$ et

$$\mathcal{H}(X) = \mathcal{H}(M_{s\Theta}X).$$

Dans ce cas, l'équation hamiltonienne associée à $\mathcal{H}$ stabilise toutes les coniques centrées engendrées par la matrice $A_\Theta := JO\Gamma_\Theta O^T$. Autrement dit, on a :

$$\frac{\mathrm{d}}{\mathrm{d}t}X^T(t)A_\Theta X(t) = 0. \tag{2.20}$$

On peut remarquer que la matrice $O\Gamma_\Theta O^T$ est anti-symétrique, et donc que A_Θ est symétrique. On peut reformuler ce théorème en disant que la dynamique préserve la forme quadratique canoniquement associée à la matrice symétrique A_Θ. Concernant la matrice O : elle joue le rôle d'une matrice de changement de base. On peut refaire cette démonstration en prenant $O = I_{2d}$ et en se plaçant directement dans une base adaptée aux rotations $\overline{R}_{s\Theta}$.

Démonstration. Étape 1. On commence par reformuler la propriété d'invariance sur $\mathcal{H}$ par une condition sur son gradient. Par un développement limité au voisinage de $s = 0$:

$$0 = \frac{\mathcal{H}(M_{s\Theta}X) - \mathcal{H}(X)}{s} = \frac{\mathrm{d}}{\mathrm{d}s}\mathcal{H}(M_{s\Theta}X)\Big|_{s=0} + \mathcal{O}(s).$$

On calcule :

$$\frac{\mathrm{d}}{\mathrm{d}s}\mathcal{H}(M_{s\Theta}X) = \nabla\mathcal{H}(X)^T O\left(\frac{\mathrm{d}}{\mathrm{d}s}\overline{R}_{s\Theta}\right)O^T X.$$

On calcule la dérivation de la matrice $R_{s\Theta}$ blocs par blocs, sachant que

$$\frac{\mathrm{d}}{\mathrm{d}s}R_{s\theta} = \frac{\mathrm{d}}{ds}\begin{pmatrix}\cos(s\theta) & -\sin(s\theta)\\ \sin(s\theta) & \cos(s\theta)\end{pmatrix} = \begin{pmatrix}-\theta\sin(s\theta) & -\theta\cos(s\theta)\\ \theta\cos(s\theta) & \theta\sin(s\theta)\end{pmatrix}$$

On vérifie que en $s = 0$ on a $\mathrm{d}R_{s\theta}/\mathrm{d}s = \theta J_1$, de sorte que

$$\frac{\mathrm{d}}{\mathrm{d}s}\overline{R}_{s\Theta}\Big|_{\theta=0} = \Gamma_\Theta.$$

Si on injecte cette identité dans la première équation, on en déduit que

$$\nabla\mathcal{H}(X)^T O\Gamma_\Theta O^T X = 0.$$

Étape 2. A présent on démontre la propriété de conservation. En utilisant le fait que A_Θ est symétrique :

$$\frac{\mathrm{d}}{\mathrm{d}t}X^T A_\Theta X \;=\; \dot{X}^T A_\Theta X + X^T A_\Theta \dot{X} \;=\; 2\dot{X}^T A_\Theta X \;=\; -2\big(J\nabla\mathcal{H}(X)\big)^T A_\Theta X. \tag{2.21}$$

En remplaçant A_Θ par son expression et en utilisant $J^T J = I_{2d}$, on obtient

$$\frac{\mathrm{d}}{\mathrm{d}t}X^T A_\Theta X \;=\; -2\nabla\mathcal{H}(X)^T J^T JO\Gamma_\Theta O^T X \;=\; -2\nabla\mathcal{H}(X)^T O\Gamma_\Theta O^T X = 0.$$

□

L'important n'est pas vraiment de retenir les énoncés de ces deux théorèmes et leur démonstration mais plutôt l'état d'esprit avec lequel on traite ces questions :

- On identifie un invariant dans le hamiltonien $\mathcal{H}$.
- On en déduit une condition sur $\nabla\mathcal{H}$.
- On en déduit une loi de conservation.

Cette idée a été plus tard généralisée et rendue rigoureuse par le théorème de Noether, dont un énoncé simplifié est le suivant :

Theorem 2.2.1 (Noether)**.**
À toute transformation infinitésimale qui laisse invariante le hamiltonien correspond une quantité qui se conserve.

C'est avec ce théorème que l'on démontre toutes les lois de conservation de la mécanique classique (conservation de l'énergie, de la quantité de mouvement, du moment cinétique, de la charge électrique, du flux magnétique, etc...) mais aussi de la mécanique quantique ou relativiste.

Remarque importante : il n'est pas nécessaire qu'un système soit hamiltonien pour avoir des quantités conservées !

2.2.4 Systèmes dissipatifs

Les systèmes dissipatifs sont les systèmes qui perdent de l'énergie au cours du temps. On les définit de manière formelle comme suit :

Définition 2.4 Systèmes dissipatifs (†)

Un système dynamique $t \mapsto X(t)$ est dit *dissipatif* lorsqu'il existe un hamiltonien $\mathcal{H} : \Omega \to \mathbb{R}$ et un terme dissipatif $\mathcal{D} : \Omega \to \mathbb{R}^d$ tels que

$$\frac{\mathrm{d}}{\mathrm{d}t}X = -J\nabla\mathcal{H}(X) - \mathcal{D}(X), \tag{2.22}$$

où le terme dissipatif vérifie la condition suivante (*condition de dissipation*) :

$$\forall\, X \in \Omega, \qquad \nabla\mathcal{H}(X)^T\mathcal{D}(X) \;\geq 0.$$

On peut vérifier que si $\mathcal{D}$ est nul alors la dynamique est hamiltonienne. La notion de système dissipatif est donc pertinente dans les régions où l'inégalité ci-dessus est stricte. Les points X pour lesquels l'inégalité ci-dessus est stricte sont les ***points dissipatifs*** ou de ***zones de dissipation***. Le terme dissipatif est le terme dans l'équation qui engendre un phénomène de dissipation d'énergie. En effet, il est direct de montrer que :

Proposition 2.3 Dissipation du hamiltonien

Le hamiltonien décroît au cours du temps et sa vitesse de décroissance est donnée par

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{H}\big(X(t)\big) = -\nabla\mathcal{H}\big(X(t)\big)^T\mathcal{D}\big(X(t)\big) \;\leq 0.$$

Un exemple important de système dissipatif est celui des frottements linéaires à coefficients constants en mécanique classique. Il s'agit des systèmes de la forme (avec $X \in \mathbb{R}^d$) :

$$\ddot{X} = -\nabla \mathcal{P}(X) - D\dot{X}$$

où D est une matrice diagonale positive. A l'aide de l'énergie mécanique $E_M = |V|^2/2 + \mathcal{P}(X)$, on réécrit ce système sous forme hamiltonienne + un terme dissipatif :

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} X \\ V \end{pmatrix} \;=\; -J\nabla E_m - \begin{pmatrix} 0 \\ DV \end{pmatrix}.$$

Par un calcul direct, on peut vérifier que

$$\nabla E_m^T \begin{pmatrix} 0 \\ DV \end{pmatrix} = V^T DV \geq 0.$$

On remarque enfin que les zones de dissipations sont les zones pour lequelles le terme ci-dessus est strictement positif. Si la matrice D est défini-positive, alors les zones de dissipation correspondent à la condition $V \neq 0$.

Définition 2.5 Terme dissipatif pour la mécanique hamiltonienne

On considère un terme dissipatif $\mathcal{D}(X, V)$ associé à un hamiltonien mécanique $\mathcal{H}(X, V)$, on dit que le dissipateur est ***mécanique*** lorsqu'il est nul ssi la vitesse est nulle :

$$\forall\, X \in \Omega, \qquad V \neq 0 \quad \Longleftrightarrow \quad \nabla \mathcal{H}(X, V)^T \mathcal{D}(X, V) > 0.$$

On parle alors d'un ***système mécanique dissipatif.***

2.3 Différentes notions de stabilité

L'étude de la stabilité consiste à étudier le comportement des solutions d'une EDO au voisinage d'une solution. L'étude de la stabilité est fondamentale pour la construction de schémas numériques pour la résolution des EDO à l'aide d'un ordinateur. C'est également un concept qui joue un rôle important en physique pour l'étude de l'existence physique d'états stationnaires. Cependant, différentes notions de stabilité existent pour décrire des phénomènes de stabilité proches mais ayant certaines différences importantes. Pour la suite on considère le problème de Cauchy

$$\frac{\mathrm{d}}{\mathrm{d}t}X(t) \;=\; \mathcal{F}\Big(t, X(t)\Big), \qquad \text{et} \qquad X(0) = X_0 \in \Omega, \tag{2.23}$$

2.3.1 Stabilité et stabilité asymptotique

La première notion de *stabilité*, la plus naturelle également, consiste à regarder l'ensemble des données initiales proches de la donnée X_0 et à comparer le comportement relatif des solutions en temps long. Autrement dit, si on considère deux données initiales X_0 et X_1 "proches", c'est-à-dire à distance inférieure à un petit paramète $\delta > 0$, les solutions associées $\mathscr{S}^t X_0$ et $\mathscr{S}^t X_1$ vont-elles rester proches au cours du temps ou bien diverger ?

DÉFINITION 2.6 Stabilité et stabilité asymptotique (‡)

(*i*) La solution $t \mapsto \mathscr{S}^t X_0$ est dite ***stable*** ssi pour tout $\varepsilon > 0$, il existe un $\delta > 0$ tel que

$$|X_1 - X_0| \leq \delta \qquad \Longrightarrow \qquad \sup_{t \geq 0} \left|\mathscr{S}^t X_1 - \mathscr{S}^t X_0\right| \leq \varepsilon.$$

(*ii*) La solution $t \mapsto \mathscr{S}^t X_0$ est dite ***asymptotiquement stable*** ssi il existe $\delta > 0$ tel que

$$|X_1 - X_0| \leq \delta \qquad \Longrightarrow \qquad \left|\mathscr{S}^t X_1 - \mathscr{S}^t X_0\right| \longrightarrow 0, \quad \text{lorsque } t \to +\infty.$$

Remarque : La notion de *stabilité asymptotique* implique la notion de *stabilité* mais la réciproque est fausse puisque le ε ne dépend pas du temps.

DÉFINITION 2.7 Etat stable (‡)

On dit que $X_0 \in \Omega$ est un ***état stable*** (resp. un ***état asymptotiquement stable***) ssi

- Il s'agit d'un état stationnaire. Autrement dit : $\mathscr{S}^t X_0 = X_0$ pour tout temps $t \geq 0$.
- La solution associée est stable (resp. asymptotiquement stable).

Remarque : Si on cherche à étudier la stabilité d'une solution instationnaire $t \mapsto \mathscr{S}^t X_0$, c'est-à-dire si on souhaite étudier $t \mapsto |\mathscr{S}^t X_1 - \mathscr{S}^t X_0|$ pour tout les X_1 "assez proches" de X_0, on peut toujours se ramener au cas stationnaire. Il suffit pour cela de considérer l'équation différentielle satisfaite par la différence des deux solutions.

On définit également la notion de solutions instables comme étant la négation de la notion de stabilité :

DÉFINITION 2.8 instabilité (†)

La solution $t \mapsto \mathscr{S}^t X_0$ est dite ***instable*** ssi elle n'est pas stable. Si X_0 est un état stationnaire alors dans ce cas on dit que c'est un ***état instable.***

2.3.2 Stabilité de Lyapunov et théorème de Lyapunov

La notion de stabilité introduite par Aleksandr Lyapunov consiste à exploiter les propriétés des quantités conservées ou dissipées par le flot de l'équation pour décrire le comportement de la solution.

DÉFINITION 2.9 Stabilité au sens de Lyapunov (‡)

On considère un problème de Cauchy (2.23) associé à une fonction $\mathcal{F}$ autonome en temps et X_0 un état stationnaire. On dit que X_0 est ***stable au sens de Lyapunov*** ssi il existe $\mathcal{L} : \Omega \to \mathbb{R}$ et un $\delta > 0$ tels que :

- On a $\mathcal{L}(X) > \mathcal{L}(X_0)$ pour tout X dans $\mathcal{B}(X_0, \delta) \setminus \{X_0\}$.
- Pour tout $X \in \mathcal{B}(X_0, \delta) \setminus \{X_0\}$, et pour tout $t \geq 0$, nous avons $\dfrac{\mathrm{d}}{\mathrm{d}t}\mathcal{L}(\mathscr{S}^t X) \leq 0$.

Si cette dernière inégalité est stricte, on dit que X_0 est un état ***asymptotiquement stable au sens de Lyapunov.***

La première condition signifie que X_0 est un minimum local strict de $\mathcal{L}$ et la deuxième condition signifie que $\mathcal{L}$ est une *quantité dissipée* par la dynamique au voisinage de X_0. La fonction $\mathcal{L}$ est appelée ***fonction de Lyapunov.*** On a également une notion d'instabilité au sens de Lyapunov :

Définition 2.10 Instabilité au sens de Lyapunov (†)

On dit que X_0 est ***instable au sens de Lyapunov*** ssi il existe $\mathcal{L} : \Omega \to \mathbb{R}$ et une courbe lipschitz $s \in [0,1] \mapsto X_s \in \mathcal{B}(0,\delta)$ pour $\delta > 0$ qui satisfont les propriétés :

- On a $\mathcal{L}(X_s) \leq \mathcal{L}(X_0)$ pour tout $s \in]0,\delta]$.
- Pour tout $X \in \mathcal{B}(X_0,\delta) \setminus \{X_0\}$, et pour tout $t \geq 0$, nous avons $\dfrac{\mathrm{d}}{\mathrm{d}t}\mathcal{L}(\mathscr{S}^t X) \leq 0$.

Il est important de remarquer que la deuxième condition, c'est-à-dire ***la décroissance de $\mathcal{L}$ au cours du temps***, peut se réécrire sous la forme d'une condition de signe sur produit scalaire entre le champs de vitesse $\mathcal{F}$ et le gradient de la fonction de Lyapunov. En effet, par la dérivation des fonctions composées :

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{L}\big(X(t)\big) \;=\; \nabla\mathcal{L}\big(X(t)\big)\cdot\frac{\mathrm{d}X}{\mathrm{d}t} \;=\; \nabla\mathcal{L}\big(X(t)\big)^T\mathcal{F}\big(X(t)\big) \;\leq\; 0. \tag{2.24}$$

Dans le case stable, on peut regarder $\mathcal{L}$ comme étant la représentation d'une énergie dissipée au sein d'un puits de potentiel. Dans le cas instable, la courbe $s \mapsto X_s$ est une courbe le long de laquelle l'énergie décroît et donc au voisinage de laquelle le système va s'éloigner indéfiniment de l'état stationnaire X_0.

Théorème 2.4 Le théorème de stabilité de Lyapunov-Persidsky (‡)

On considère un problème de Cauchy associé à un champs de vitesse $\mathcal{F}$ supposée autonome en temps et soit X_0 un état stationnaire pour cette équation. On a :

- X_0 est stable au sens de Lyapunov $\iff$ X_0 est stable.
- X_0 asymptotiquement Lyapunov stable $\iff$ X_0 asymptotiquement stable.

Remarque : Dans le cadre de ce cours, on étudie seulement des fonctions de Lyapunov indépendantes du temps (cas autonome). Cette théorie se généralise de manière naturelle au cas ou $\mathcal{F}$ et $\mathcal{L}$ dépendent du temps. Remarquons enfin qu'il existe d'autres notions de stabilité (orbitale, structurelle, etc...) mais qui ne sont pas abordés dans ce cours. Il est donc recommandé de préciser qu'on parle de la ***stabilité de Lyapunov.***

Le sens $\Rightarrow$ a été démontré par Lyapunov dans le cadre de ses travaux fondateurs sur la stabilité de Lyapunov. Persidsky démontre la réciproque. Son nom étant moins connu, on appelle souvent ce théorème du seul nom de Lyapunov. Dans le cas instable, nous avons :

Théorème 2.5 Le théorème d'instabilité de Lyapunov-Chetaev (†)

On considère un problème de Cauchy associé à une fonction $\mathcal{F}$ autonome en temps et X_0 un état stationnaire. On a :

$$X_0 \text{ Lyapunov instable} \quad\Longrightarrow\quad X_0 \text{ instable.}$$

Attention ! Contrairement au théorème de stabilité, la réciproque est fausse ici ! En effet, il existe des instabilités qui ne suivent pas l'intuition d'une "courbe d'instabilité" telle que introduite dans la définition de l'instabilité de Lyapunov. Cependant, il s'agit de situations mathématiques considérées comme pathologiques et rarement rencontrées en pratique pour l'analyse des équations issues de la physique.

2.3.3 Application à la mécanique hamiltonienne et dissipative

La théorie de Lyapunov est particulièrement utile pour étudier la stabilité des point-critiques pour les potentiels d'énergie en mécanique hamiltonienne ou dissipative. Pour cela on s'appuie sur la reformulation hamiltonienne (ou dissipative) dans l'espace des phases $Z = (X, V)^T$:

$$\frac{\mathrm{d}Z}{\mathrm{d}t} \;=\; -J\nabla\mathcal{H}(Z) - \mathcal{D}, \tag{2.25}$$

où le hamiltonien désigne l'énergie mécanique $\mathcal{H}(X, V) := |V|^2/2 + \mathcal{P}(X)$ et $\mathcal{D}$ un dissipateur mécanique.

PROPOSITION 2.4 Stabilité pour les équations de la mécanique

On considère un problème de Cauchy de la mécanique hamiltonienne ou dissipative d'inconnue $(X, V) \in \mathbb{R}^d \times \mathbb{R}^d$. On note $\mathcal{P}$ le potentiel mécanique, $\mathcal{H}$ le hamiltonien mécanique et $\mathcal{D}$ un terme dissipatif (éventuellement nul). Soit X_0 un point-critique du potentiel $\mathcal{P}$.

(*i*) Si X_0 est un minimum local strict du potentiel $\mathcal{P}$, alors $(X_0, 0)$ est un *état stable* quel que soit le dissipateur $\mathcal{D}$.

(*ii*) Si X_0 est un minimum local strict de $\mathcal{P}$ et si $\mathcal{D}$ est un dissipateur mécanique, alors $(X_0, 0)$ est un état *asymptotiquement stable*.

(*iii*) Si X_0 est un maximum local de $\mathcal{P}$ le long d'une courbe lipschitzienne d'équation $s \in [0, 1] \mapsto X_0 + \Xi(s)$ (avec $\Xi(0) = 0$) alors $(X_0, 0)$ est un état *instable*.

Démonstration. Comme conséquence de la formulation mécanique donnée à l'équation (2.16), le fait que X_0 soit un point-critique de $\mathcal{P}$ implique que $(X_0, 0)$ soit une solution stationnaire. En effet, on remarque que $\nabla\mathcal{P}(X_0) = 0$ implique $\nabla\mathcal{H}(X_0, 0) = 0$. Par ailleurs, comme le terme dissipatif $\mathcal{D}$ est un terme dissipatif mécanique (Définition 2.5), on a aussi $\mathcal{D}(X_0, 0) = 0$ de sorte que $(X_0, 0)$ est bien un état stationnaire. Pour cette démonstration on notera ∇_x le gradient par rapport à X et ∇_v celui par rapport à V (et aussi $\nabla_{x,v}$ celui par rapport aux 2 variables vectorielles).

Preuve du point (i). On choisit le hamiltonien comme fonction de Lyapunov [1] : $\mathcal{L} := \mathcal{H}$. Comme X_0 est un minimum strict du potentiel, on a $\mathcal{P}(X) > \mathcal{P}(X_0)$ pour tout X dans la boule épointée $\mathcal{B}(X_0, \delta) \setminus \{X_0\}$ pour un certain $\delta > 0$ assez petit. Donc pour tout $X \neq X_0$ dans cette boule et pour tout $V \in \mathbb{R}^d$:

$$\mathcal{L}(X, V) = \frac{|V|^2}{2} + \mathcal{P}(X) \;>\; \frac{|V|^2}{2} + \mathcal{P}(X_0) \;\geq \mathcal{P}(X_0) \;=\; \mathcal{L}(X_0, 0).$$

Si en revanche, on a $X = X_0$ mais $V \neq 0$, on vérifie que

$$\mathcal{L}(X, V) = \frac{|V|^2}{2} + \mathcal{P}(X_0) \;>\; 0 + \mathcal{P}(X_0) \;=\; \mathcal{L}(X_0, 0).$$

1. Quand on cherche à étudier la stabilité d'un étant stationnaire, il FAUT penser à utiliser le hamiltonien comme fonction de Lyapunov ; c'est souvent suffisant pour conclure, ou en tout cas permet de beaucoup avancer.

Ainsi dans tous les cas lorsque (X, V) est proche mais distinct du point stationnaire $(X_0, 0)$ alors $\mathcal{L}(X, V) > \mathcal{L}(X_0, 0)$. On a donc vérifié le premier point de la définition de la stabilité de Lyapunov (Définition 2.9).

Par ailleurs, comme $\mathcal{D}$ est dissipateur pour ce hamiltonien, on a, pour $X \in \mathcal{B}(X_0, \delta) \setminus \{X_0\}$:

$$\begin{aligned} \nabla_{X,V}\, \mathcal{L}(X, V) \cdot \mathcal{F}(X, V) &= \nabla_{X,V}\, \mathcal{H}(X, V) \cdot \big(- J\, \nabla_{X,V}\, \mathcal{H}(X, V) - \mathcal{D}(X, V)\big) \\ &= - \nabla_{X,V}\, \mathcal{H}(X, V) \cdot \mathcal{D}(X, V) \ \leq 0. \end{aligned} \tag{2.26}$$

où l'on a utilisé $\nabla_{X,V}\, \mathcal{L} = \nabla_{X,V}\, \mathcal{H}$ ainsi que la propriété $Y^T J Y = 0$ car J est anti-symétrique. Cette condition de signe sur le produit scalaire est équivalente à la dissipation de $\mathcal{L}$ en vertu de l'égalité (2.24). On a donc vérifié le second point de la définition de la stabilité de Lyapunov (Définition 2.9). On en conclut donc par théorème que $(X_0, 0)$ est stable au sens de Lyapunov.

Preuve du point (ii). Si le voisinage de $(X_0, 0)$ était une zone de dissipation pour toutes les vitesses on pourrait conclure directement au caractère asymptotiquement stable avec l'inégalité précédente (devenue stricte). Cependant, cette propriété n'est vraie que si la vitesse est non nulle car $\mathcal{D}$ est un terme dissipatif mécanique. Pour arriver à la conclusion, il faut nécessairement modifier la fonction de Lyapunov. Nous allons faire la preuve dans le cas particulier de la dissipation linéaire uniforme :

$$\mathcal{D} := \begin{pmatrix} 0 \\ DV \end{pmatrix}, \qquad \text{avec } D \text{ une matrice } d \times d \text{ diagonale et définie positive.}$$

La preuve dans le cas général est beaucoup plus technique et difficile à lire mais pas très intéressante. Toutes les idées importantes sont contenues dans la démonstration du cas d'un dissipateur linéaire diagonal suivante. Nous commençons par introduire une fonction de Lyapunov modifiée (avec un paramètre $\mu > 0$) :

$$\mathcal{L}_\mu(X, V) := \mathcal{L}(X, V) + \mu V^T\, \nabla_x\, \mathcal{P}(X) + \frac{1}{\mu} D\big(\mathcal{P}(X) - \mathcal{P}(X_0)\big). \tag{2.27}$$

On remarque que nous avons toujours $\mathcal{L}(X_0, 0) = \mathcal{P}(X_0)$. nous avons aussi pour $X \in \mathcal{B}(X_0, \delta) \setminus \{X_0\}$ (comme D est diagonale défini-positive et X_0 un minimum local strict) :

$$\frac{1}{\mu} D\big(\mathcal{P}(X) - \mathcal{P}(X_0)\big) > 0.$$

Donc,

$$\mathcal{L}_\mu(X, V) > \frac{|V|^2}{2} + \mathcal{P}(X) - \mu V^T \nabla \mathcal{P}(X)$$

Par ailleurs, en utilisant l'inégalité de Young ($2|ab| \leq (a^2 + b^2)$) pour borner le produit scalaire $V^T \nabla \mathcal{P}$ terme à terme par $|V|^2/2 + |\nabla \mathcal{P}|^2/2$, on aboutit à

$$\frac{|V|^2}{2} + \mathcal{P}(X) - \mu V^T \nabla \mathcal{P}(X) \geq \frac{2 - \mu}{4} |V|^2 + \mathcal{P}(X) - \frac{\mu}{2} |\nabla \mathcal{P}(X)|^2, \tag{2.28}$$

On remarque à présent que, puisque X_0 est un minimum local, il n'y a pas de terme d'ordre 1 dans le développement limité (en notant $H := X - X_0$) :

$$\mathcal{P}(X) = \mathcal{P}(X_0) + H^T \nabla^2 \mathcal{P}(X_0) H + \mathcal{O}\big(|H|^3\big),$$

Comme le gradient du hamiltonien $\mathcal{H}$ est localement lipschitzien, il s'en suit que le gradient du gradient, c'est-à-dire la matrice hessienne $\nabla^2 \mathcal{P}$, est borné sur $\mathcal{B}(X_0, \delta)$. Donc

$$\big|\mathcal{P}(X) - \mathcal{P}(X_0)\big| \ \lesssim \ \big\|\nabla^2 \mathcal{P}\big\|_{L^\infty} |H|^2.$$

Cette dernière inégalité implique, en refaisant un développement limité,

$$|\nabla\mathcal{P}(X)|^2 \;=\; \frac{|\mathcal{P}(X)-\mathcal{P}(X_0)|^2}{H^2}+\mathcal{O}(H) \;\lesssim\; \|\nabla^2\mathcal{P}\|_{L^\infty}|\mathcal{P}(X)-\mathcal{P}(X_0)|$$

Par conséquent, si μ est assez petit :

$$\begin{aligned}\mathcal{P}(X)-\mathcal{P}(X_0)-\frac{\mu}{2}|\nabla\mathcal{P}(X)|^2 \;&\gtrsim\; \mathcal{P}(X)-\mathcal{P}(X_0)-\frac{\mu}{2}\|\nabla^2\mathcal{P}\|_{L^\infty}|\mathcal{P}(X)-\mathcal{P}(X_0)|\\ &\gtrsim\; \mathcal{P}(X)-\mathcal{P}(X_0)>0,\end{aligned} \tag{2.29}$$

où pour la dernière inégalité nous avons utilisé le fait que X_0 est un minimum strict. La constante multiplicative qui est cachée dans le dernier symbole $\gtrsim$ ci-dessus va dépendre de la valeur de $\|\nabla^2\mathcal{P}\|_{L^\infty}$. De même, le bon choix pour μ "assez petit" va dépendre de $\|\nabla^2\mathcal{P}\|_{L^\infty}$. A présent, si on injecte ce résultat dans l'équation (2.28) et les précédentes, on aboutit à :

$$\mathcal{L}_\mu(X,V) \;>\; \frac{2-\mu}{4}|V|^2.$$

Si on demande en plus à ce que $\mu\leq 1$, cette inégalité implique,

$$\forall\, X\in\mathcal{B}(X_0,\delta)\setminus\{X_0\},\qquad \mathcal{L}_\mu(X,V) \;>\; \frac{|V|^2}{4}.$$

A partir de ce résultat, on peut aisément en déduire que la fonction $\mathcal{L}$ satisfait le premier point de la définition des fonctions de Lyapunov.

Pour prouver le second point, on commence par calculer le gradient de $\mathcal{L}_\mu$:

$$\nabla_{\!x}\,\mathcal{L}_\mu(X,V) \;=\; \nabla_{\!x}\,\mathcal{P}(X)+\mu\,\nabla_{\!x}{}^2\,\mathcal{P}(X)V+\frac{1}{\mu}D\,\nabla_{\!x}\,\mathcal{P}(X),$$

$$\nabla_{\!v}\,\mathcal{L}_\mu(X,V) \;=\; V+\mu\,\nabla_{\!x}\,\mathcal{P}.$$

On calcule à présent le produit scalaire avec le champ de vitesse $\mathcal{F}=-J\,\nabla_{\!X,V}\,\mathcal{H}-\mathcal{D}$:

$$\begin{aligned}&\nabla_{\!X,V}\,\mathcal{L}_\mu\cdot\Big(-J\,\nabla_{\!X,V}\,\mathcal{H}-\mathcal{D}\Big)\\ &=-V^TDV+\mu V^T\,\nabla_{\!x}{}^2\,\mathcal{P}V+\mu V^T\left(\frac{1}{\mu}D\,\nabla_{\!x}\,\mathcal{P}\right)-\mu\big|\nabla_{\!x}\,\mathcal{P}\big|^2-V^TD\,\nabla_{\!x}\,\mathcal{P}\\ &=-V^TDV+\mu V^T\,\nabla_{\!x}{}^2\,\mathcal{P}V-\mu\big|\nabla_{\!x}\,\mathcal{P}\big|^2,\end{aligned}$$

où pour la deuxième égalité on a remarqué que les termes en positions 3 et 5 de l'équation précédente se simplifient. On remarque à présent que, si on note $\lambda_1(D)>0$ la plus petite valeur propre de la matrice diagonale D, on a

$$\big|V^TDV\big|\geq\lambda_1(D)|V|^2.$$

Par ailleurs, on va utiliser de nouveau le fait que la matrice hessienne de $\mathcal{P}$ est bornée sur la boule $\mathcal{B}(X_0,\delta)$ pour écrire

$$|V^T\,\nabla_{\!x}{}^2\,\mathcal{P}(X)V| \;\lesssim\; \|\nabla_{\!x}{}^2\,\mathcal{P}\|_{L^\infty}|V|^2.$$

Par conséquent, quitte à changer la valeur de μ, il existe une constante $c>0$ telle que

$$-V^TDV+\mu V^T\,\nabla_{\!x}{}^2\,\mathcal{P}(X)V \;\leq\; -c|V|^2.$$

En effet, il suffit de choisir μ assez petit par rapport à la valeur de $\lambda_1(D)$ et $\|\nabla_x{}^2\mathcal{P}\|_{L^\infty}$. Si on injecte cette estimation dans le calcul du produit scalaire, on a

$$\nabla_{x,v}\mathcal{L}_\mu\cdot\mathcal{F} \;\leq\; -c|V|^2-\mu|\mathcal{P}(X)-\mathcal{P}(X_0)|^2.$$

Puisque X_0 est un minimum strict de $\mathcal{P}$ on en déduit que $|\mathcal{P}(X)-\mathcal{P}(X_0)|^2>0$ sur $\mathcal{B}(X_0,\delta)\setminus\{X_0\}$ et donc :

$$\nabla_{x,v}\mathcal{L}_\mu(X,V)\cdot\mathcal{F}(X,V) \;<\; 0,$$

dès que $(X,V)\neq(X_0,0)$. Par théorème de Lyapunov, on en conclut que $(X_0,0)$ est un état stationnaire asymptotiquement stable.

Preuve du point (iii). Si on travaille de nouveau avec $\mathcal{L}:=\mathcal{H}$ comme fonction de Lyapunov, on constate que le long de la courbe lipschitzienne $s\in[0,1]\mapsto\big(X_0+\Xi(s),0\big)$ nous avons :

$$\mathcal{L}\big(X_0+\Xi(s),0\big) \;=\; \frac{|0|^2}{2}+\mathcal{P}\big(X_0+\Xi(s)\big)-\mathcal{P}\big(X_0\big) \;\leq\; 0.$$

On conclut par le théorème d'instabilité de Lyapunov-Chetaev. □

Remarque : Si un dissipateur est tel que l'inégalité de la Proposition 2.3 est stricte sur un voisinage de X_0, par abus de langage on dit qu'il s'agit d'un *dissipateur coercif*. On peut alors étudier la stabilité asymptotique directement en prenant comme fonction de Lyapunov $\mathcal{L}:=\mathcal{H}$ le hamiltonien.

Pour les systèmes mécaniques dissipatifs (Définition 2.5), le dissipateur s'annule sur tout l'espace vectoriel engendré par l'équation $V=0$. Pourtant, on a bien une propriété de stabilité asymptotique pour les minimal locaux du stricts du potentiel, ainsi que nous l'avons démontré à la proposition 2.4-(ii). Pour faire cette démonstration, nous n'avons pas utilisé directement le hamiltonien comme fonction de Lyapunov mais une perturbation du hamiltonien (2.27). Lorsque le dissipateur n'est pas coercif mais qu'il implique malgré tout la stabilité asymptotique, à l'image la proposition 2.4-(ii), on dit qu'il est *hypo-coercif.*

2.4 Stabilité des équations linéaires et stabilité du linéarisé

Une partie importante de l'analyse de la stabilité des système concerne l'analyse de la stabilité des systèmes linéaires. Lorsque le système est non-linéaire, on peut le linéariser au voisinage de l'état stationnaire et obtenir des résultats de stabilité non-linéaire à partir de l'étude de la stabilité linéaire.

2.4.1 Stabilité de l'exponentielle d'un bloc de Jordan

On rappelle que la résolution d'un système linéaire se fait à l'aide de la réduction de Jordan (définition 0.18) et des exponentielles de matrices (définition 0.19). On note $\mathfrak{R}(z)$ la partie réelle d'un nombre complexe $z\in\mathbb{C}$ et $\mathfrak{I}(z)$ sa partie imaginaire.

Proposition 2.5 Analyse asymptotique des systèmes linéaires de Jordan **(†)**

Soit $J_k(\lambda)$ une matrice de Jordan (voir Définition 0.18) avec $k\in\mathbb{N}^*$ et $\lambda\in\mathbb{C}$.

- Si $\mathfrak{R}(\lambda)<0$, alors $\big\|\exp(tJ_k(\lambda))\big\|\longrightarrow 0$, lorsque $t\to+\infty$.
- Si $\mathfrak{R}(\lambda)>0$, alors $\big\|\exp(tJ_k(\lambda))\big\|\longrightarrow+\infty$, lorsque $t\to+\infty$.
- Si $\mathfrak{R}(\lambda)=0$, et $k>1$ alors $\big\|\exp(tJ_k(\lambda))\big\|\longrightarrow+\infty$.
- Si $\mathfrak{R}(\lambda)=0$, et $k=1$ alors $\|\exp(tJ_k(\lambda))\|\leq 1$, $\forall t$.

On rappelle que les ensembles de matrices sont aussi des espaces vectoriels de dimension finie. Ainsi la convergence présente dans ce résultat est la convergence associée à la norme euclidienne canonique $\|M\| := (\sum_{j,k} m_{jk}^2)^{1/2} = \sqrt{\mathrm{tr}(M^T M)}$. La démonstration de ce théorème est élémentaire à partir des calculs effectués lors de la définition des exponentiels de blocs de Jordan.

On constate que lorsque $\Re(\lambda) = 0$, le comportement diffère selon la taille du bloc de Jordan. Si une valeur propre λ n'a que des blocs de Jordan de taille 1, on dit qu'il s'agit d'une ***valeur propre semi-simple***. Le corollaire immédiat de cette proposition est le théorème suivant :

Théorème 2.6 Théorème de stabilité linéaire **(†)**

On étudie le système linéaire à coefficients constants $\dot{X} = AX$.

(i) L'état stationnaire $0 \in \mathbb{R}^d$ est asymptotiquement stable ssi toutes les valeurs propres de la matrice A sont de partie réelle strictement négative.

(ii) L'état stationnaire $0 \in \mathbb{R}^d$ est Lyapunov stable ssi toutes les valeurs propres de la matrice A sont, soit de partie réelle strictement négative, soit de partie réelle nulle et semi-simples.

Afin de décrire plus précisément les dynamiques linéaires, et le fait qu'elles engendrent *génériquement* des convergences ou des divergences exponentielles, on introduit la une nouvelle notion de stabilité :

Définition 2.11 Stabilité linéaire **(†)**

(i) La solution $t \mapsto \mathscr{S}^t X_0$ est dite **linéairement stable** ssi il existe $\delta, \lambda > 0$ tels que

$$|X_1 - X_0| \leq \delta \qquad \Longrightarrow \qquad \left|\mathscr{S}^t X_1 - \mathscr{S}^t X_0\right| \leq |X_1 - X_0| e^{-\lambda t}$$

(ii) La solution $t \mapsto \mathscr{S}^t X_0$ est dite **linéairement instable** ssi il existe $\lambda > 0$ et il existe une suite (X_n) convergeant vers X_0 tels que

$$\forall\, n \in \mathbb{N}, \qquad \left|\mathscr{S}^t X_n - \mathscr{S}^t X_0\right| \geq |X_n - X_0| e^{\lambda t}$$

Cette notion de stabilité est plus précise que celles introduites par la définition 2.6 car elle nous donne une vitesse de convergence (ou de divergence) du système différentiel. On peut vérifier en particulier les implications suivantes :

$$\begin{array}{lll} \bullet \text{ Stabilité linéaire} & \Longrightarrow & \text{Stabilité asymptotique.} \\ \bullet \text{ Instabilité linéaire} & \Longrightarrow & \text{Instabilité.} \end{array} \tag{2.30}$$

Dans les deux cas la réciproque est fausse. Cependant, dans le cas des équations linéaires à coefficients constants, c'est-à-dire de la forme $\dot{X} = AX$, on peut raffiner les résultats du théorème de stabilité linéaire (Théorème 2.6) :

$$\begin{array}{l} 0 \text{ est asymptotiquement stable} \iff 0 \text{ est linéairement stable} \\ \iff \text{les valeurs propres de } A \text{ sont toutes de partie réelle strictement négative.} \end{array} \tag{2.31}$$

Ces équivalences se démontrent directement à l'aide de l'analyse des exponentielles de blocs de Jordan. Pour l'instabilité, on peut montrer que 0 est linéairement instable si et seulement si la matrice A admet au moins une valeur propre de partie réelle strictement positive.

2.4.2 Allure des solutions d'une équation linéaire

Allure des solutions d'une équation linéaire 2D

L'ensemble des comportements possibles pour les systèmes de 2 équations linéaires est résumé dans le tableau situé à la page suivante. Ces résultats se retrouvent directement à partir du calcul des exponentielles de matrices. Il est important de connaître ces résultats mais il est surtout primordial de savoir les retrouver à l'aide des exponentielles de blocs de Jordan. On a omis dans ce tableau les cas faciles correspondant aux cas où l'une des valeurs propre est nulle en effet, dans ces cas-là on peut se ramener à une dynamique 1D (on parle alors de dynamique *dégénérée*). Lorsqu'on parle de ***décrire l'allure des solutions*** d'une équation linéaire : on fait explicitement référence aux différents cas possibles dans ce tableau (ou bien on dit que la dynamique est 1D le cas échéant).

Remarque : Les courbes tracées représentent l'allure des solutions associées au bloc de Jordan. Pour en déduire le comportement réel de la solution, il faut penser à refaire le changement de base dans l'autre sens. Ainsi une dynamique qui, dans une base adaptée, est périodique sur des cercles, devient une dynamique périodique sur des ellipses inclinées après retour dans la base initiale. Il en de même pour les dynamiques en spirales, en hyperboles, etc...

Allure des solutions d'une équation linéaire en dimension supérieure

L'analyse qui conduit à ce tableau en 2D peut se reproduire sur des blocs de Jordan de toutes les tailles. Il y a peu de différences entre les résultats 2D et leurs équivalents en dimensions supérieures (y compris concernant les dégénérescences dimensionnelles liées à des valeurs propres nulles). Les 2 principales différences importantes sont les suivantes :

1) Les blocs de Jordan de taille plus grande que 2 vont donner des nœuds impropres (stables ou instables) avec des perturbations polynomiales multiplicatives d'ordre de plus en plus grand (l'ordre de la perturbation polynomiale vaut un de moins que la taille du bloc de Jordan).

2) A partir de la dimension 4 on voit apparaître des blocs de Jordan de la forme :

$$\begin{pmatrix} z & 1 \\ 0 & z \end{pmatrix} \qquad \text{et} \qquad \begin{pmatrix} \overline{z} & 1 \\ 0 & \overline{z} \end{pmatrix},$$

avec $z = a + ib$ et $b \neq 0$ (ou des blocs similaires de taille plus grande). Lorsque $a \neq 0$, ces blocs donnent un *point spiral impropre* qui est asymptotiquement stable ssi $a < 0$ et instable sinon et les trajectoires sont des perturbations polynomiales multiplicatives de la spirale logarithmique. Dans le cas où $a = 0$, on obtient également un *point spiral impropre* et celui-ci est instable. La différence étant que la dynamique n'est plus exponentielle mais polynomiale (on parle de spirale polynomiale). Dans le cas où les deux blocs sont de taille 1, on obtient les spirales d'Archimède.

Sens de rotation des spirales

Une question naturelle à partir de ce tableau consiste à savoir s'il est possible de raffiner cette analyse pour les spirales afin de savoir dans quel sens tourne la dynamique associée. Pour cela on commence par séparer la dynamique en 2 blocs formés de la partie exponentiellement divergente (ou convergente) de la partie qui produit la rotation. C'est-à-dire écrire, en notant P la matrice de changement de base qui diagonalise le système (avec $z = a + ib$) :

$$P \begin{pmatrix} z & 0 \\ 0 & \overline{z} \end{pmatrix} P^{-1} = P \begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix} P^{-1} + bP \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix} P^{-1}. \tag{2.32}$$

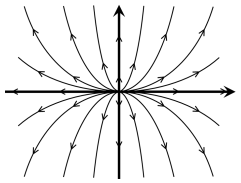
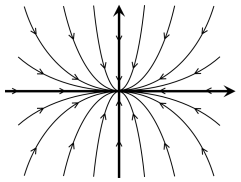
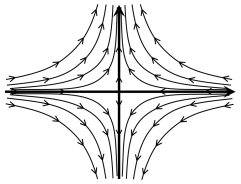
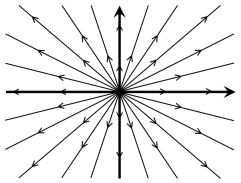
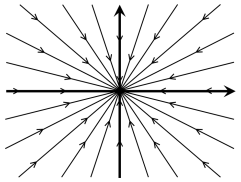
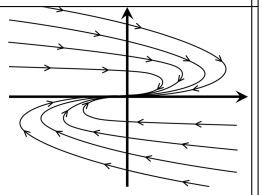
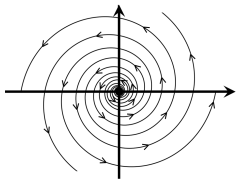
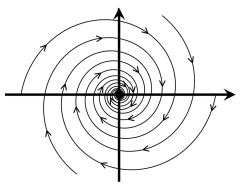
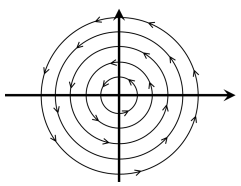
Réduite de Jordan	**Nomenclature**	**Trajectoire**	**Stabilité**	**Allure**
$\begin{pmatrix}\lambda & 0\\ 0 & \mu\end{pmatrix}$ $0<\lambda<\mu$	Nœud instable (ou source)	Divergence exponentielle le long de courbes pseudo-parabolique	Non	
$\begin{pmatrix}\lambda & 0\\ 0 & \mu\end{pmatrix}$ $\mu<\lambda<0$	Nœud stable (ou puits)	Convergence exponentielle le long de courbes pseudo-parabolique	Asymptotique	
$\begin{pmatrix}\lambda & 0\\ 0 & \mu\end{pmatrix}$ $\lambda<0<\mu$	Point col (ou selle)	Divergence exponentielle le long de courbes pseudo-hyperbolique	Non	
$\begin{pmatrix}\lambda & 0\\ 0 & \lambda\end{pmatrix}$ $0<\lambda$	Nœud instable isotrope (étoile instable)	Divergence exponentielle le long de droites linéaires	Non	
$\begin{pmatrix}\lambda & 0\\ 0 & \lambda\end{pmatrix}$ $\lambda<0$	Nœud stable isotrope (étoile stable)	Convergence exponentielle le long de droites linéaires	Asymptotique	
$\begin{pmatrix}\lambda & 1\\ 0 & \lambda\end{pmatrix}$ $\lambda\neq 0$	Nœud impropre	Perturbation polynomiale multiplicative d'une dynamique exponentielle	Asyp. stable ssi $\lambda<0$ sinon : instable	
$\begin{pmatrix}z & 0\\ 0 & \overline{z}\end{pmatrix}$ $z=a+ib$ $0<a,\ b\neq 0$	Noeud spiral instable	Divergence exponentielle le long de spirales logarithmiques	Non	
$\begin{pmatrix}z & 0\\ 0 & \overline{z}\end{pmatrix}$ $z=a+ib$ $a<0,\ b\neq 0$	Noeud spiral stable	Convergence exponentielle le long de spirales logarithmiques	Asymptotique	
$\begin{pmatrix}z & 0\\ 0 & \overline{z}\end{pmatrix}$ $z=a+ib$ $a=0,\ b\neq 0$	Point elliptique	Trajectoire periodique sur des ellipses	Lyapunov mais pas asmptotique	

FIGURE 2.1 – L'ensemble des cas possibles pour l'allure d'une solution d'un système différentiel linéaire 2D non dégénéré et à coefficients constants.

Lorsque $a < 0$, la décomposition donnée par l'équation (2.32) est une décomposition entre une partie dissipative et une partie hamiltonienne. Lorsque $a > 0$, la dissipation est négative (autrement-dit, c'est la dynamique de $t \mapsto \big(x(-t), y(-t)\big)^T$ qui est dissipative). Lorsque $a = 0$, le système est hamiltonien et parcourt une ellipse (qui se calcule à l'aide de la proposition 2.1) avec une vitesse angulaire égale à b. Pour pouvoir étudier cette rotation, on réécrit le bloc hamiltonien à l'aide de la matrice J_1 (matrice de rotation d'angle $\pi/2$). On écrit

$$P \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix} P^{-1} \;=\; Q \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} Q^{-1}.$$

Les nouvelles matrices de changement de base Q et Q^{-1} sont nécessairement à coefficients réels et s'obtiennent très simplement à partir de P et P^{-1} en utilisant la formule suivante (également appelée formule d'Euler - De Moivre, par analogie avec le lien entre trigonométrie et exponentielles complexes) :

$$\begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 1 & i \\ 1 & -i \end{pmatrix} \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ -i & i \end{pmatrix} \tag{2.33}$$

La dynamique associée à la matrice $J = J_1$, c'est-à-dire $\dot{Y} = JY$, est une dynamique périodique sur des cercles centrés dans $\mathbb{R}^2$ à vitesse angulaire 1 et cette rotation constante se fait dans le sens trigonométrique. Ainsi, étudier la dynamique de

$$\dot{X} = Q \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} Q^{-1} X,$$

revient à étudier la dynamique de $\dot{Y} = JY$ modulé par un changement de base $Y = Q^{-1}X$ (ce changement de base n'étant pas forcément orthogonal). On obtient donc une dynamique périodique sur des *ellipses* centrées dans $\mathbb{R}^2$ à vitesse angulaire 1 et cette rotation constante se fait dans le sens suivant :

— Si $\det Q > 0$, alors le sens de la rotation est trigonométrique.

— Si $\det Q < 0$, alors le sens de la rotation est anti-trigonométrique.

En effet, le signe du déterminant signale que le changement de base en 2D préserve l'orientation dans le cas positif (le changement de base est composé uniquement de rotations et de matrices symétriques défini-positives) ou inverse l'orientation dans le cas négatif (dans ce cas, il y a en plus une symétrie axiale). L'équation de l'ellipse stabilisée par le flot est donnée par la préservation de la norme de $Y = Q^{-1}X$. C'est-à-dire :

$$X^T Q^{-T} Q^{-1} X = \text{constante}.$$

2.4.3 Stabilité pour les équations linéaire et théorème du linéarisé

L'un des principaux attrait de l'étude de la stabilité pour les systèmes linéaires consiste dans le fait qu'on peut linéariser une dynamique non-linéaire compliquée au voisinage d'un point stationnaire $X_0 \in \Omega$. En effet, si $\mathcal{F}(X_0) = 0$, on peut écrire pour une perturbation $H \in \mathbb{R}^d$ petite le développement limité suivant :

$$\mathcal{F}(X_0 + H) \;=\; 0 + \nabla\mathcal{F}(X_0)^T H \;+\; \mathcal{O}\big(|H|^2\big). \tag{2.34}$$

Dans cette équation $\nabla\mathcal{F}(X_0)^T$ fait référence à la matrice jacobienne :

$$\nabla\mathcal{F}(X_0)^T = \begin{pmatrix} \dfrac{\partial f_1}{\partial x_1} & \cdots & \dfrac{\partial f_1}{\partial x_d} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial f_m}{\partial x_1} & \cdots & \dfrac{\partial f_d}{\partial x_d} \end{pmatrix}.$$

Ainsi, il est naturel pour étudier la dynamique au voisinage de X_0 de la comparer à celle du linéarisé, c'est-à-dire la dynamique donnée par

$$\frac{\mathrm{d}}{\mathrm{d}t}H \;=\; \nabla\mathcal{F}(X_0)^T H.$$

Cette dynamique, d'inconnue H admet des propriétés de stabilité que l'on obtient avec le théorème 2.6. Il est ensuite naturel de tenter d'en déduire les propriétés de stabilité de X_0 pour la dynamique initiale en analysant l'erreur entre le système non-linéaire et linéarisé.

Le principal intérêt de la notion de stabilité linéaire introduite à la définition 2.11 est qu'il permet de comparer efficacement la dynamique linéaire et la dynamique du linéarisé. En effet, dans le développement limité (2.34) le terme dominant est le terme linéaire ; si celui-ci engendre une convergence ou une divergence exponentielles, alors le terme d'ordre suivant restera négligeable pendant toute la dynamique. Plus précisément nous admettrons le théorème suivant :

Théorème 2.7 Théorème de stabilité du linéarisé (Lyapunov) **(†)**

(i) Si le système linéarisé est linéairement stable en 0 alors le système initial est linéairement stable en X_0.

(ii) Si le système linéarisé est linéairement instable en 0 alors le système initial est linéairement instable en X_0.

Dans de nombreux cas, ce théorème permet de démontrer un résultat de stabilité de manière simple et efficace sans avoir à recourir aux fonctions de Lyapunov. Dans le cas des valeurs propres de partie réelle nulle : on ne peut malheureusement pas conclure. On peut vérifier que ce cas correspond à un dynamique hamiltonienne linéaire, c'est-à-dire un point elliptique (éventuellement perturbé par des blocs de Jordan). L'ajout d'effets dissipatifs ou anti-dissipatifs d'ordre plus élevé que linéaire font basculer le système du côté asymptotiquement stable (dissipation positive) ou instable (dissipation négative). Le seul moyen de conclure consiste alors à construire une fonction de Lyapunov qui fasse apparaître les effets non-linéaires d'ordres plus élevés.

2.5 Bilan du Chapitre et exercices

2.5.1 Ce qu'il faut retenir et savoir-faire

Ce chapitre est premièrement consacré à l'analyse des quantités conservées ou dissipées par une équation différentielle ordinaire et aux propriétés que cela donne sur sa dynamique. La deuxième partie de ce chapitre est consacrée à l'étude de la stabilité des solutions, et plus particulièrement des solutions stationnaires à l'aide de la théorie de Lyapunov qui s'appuie sur la notion de quantité conservée ou dissipée. Les principaux éléments à retenir ou savoir-faire sont les suivants :

(†) Quantités conservées ou dissipées (définition)
- Quantités conservées dans le cas linéaire et critère de Lyapunov

(†) Systèmes hamiltoniens et dissipatifs (définitions)
- Propriété de préservation ou dissipation du hamiltonien

(‡) Deux notions de stabilité et le théorème de stabilité de Lyapunov

(†) Deux notions d'instabilité et le théorème d'instabilité de Lyapunov
- Théorème de stabilité pour les équations de la mécanique

(†) Analyse asymptotique des systèmes linéaires de Jordan

(†) Le tableau de l'allure des solutions d'une équation linéaire 2D.

(†) Le théorème de stabilité du linéarisé.

2.5.2 Exercices

Les exercices ci-dessous proposent d'étudier les quantités conservées et la stabilité des états stationnaires pour quelques équations différentielles, à l'aide des outils présentés dans ce chapitre. Il est recommandé de les traiter dans l'ordre. Les exercices les plus importants sont identifiés avec le symbole (⋆).

Exercice 2.1 (Allures de solutions d'équations linéaires et quantités conservées)**.** (⋆) A l'aide du tableau résumant les différentes allures possibles des solutions des équations linéaire 2D, donner l'allure des solutions pour les équations suivantes :

$$\begin{cases} \dot{x} = x - y \\ \dot{y} = 2x + 3y \end{cases} \qquad \begin{cases} \dot{x} = -x + 4y \\ \dot{y} = -2x - y \end{cases} \qquad \begin{cases} \dot{x} = 3x - y \\ \dot{y} = x + y \end{cases}$$

$$\begin{cases} \dot{x} = -x + 3y \\ \dot{y} = x - y \end{cases} \qquad \begin{cases} \dot{x} = -2x - y \\ \dot{y} = -x - 2y \end{cases} \qquad \begin{cases} \dot{x} = -x + 5y \\ \dot{y} = -2x + y \end{cases}$$

Dans le cas où les valeurs propres sont réelles distinctes ou imaginaires purs : écrire l'équation de la quantité conservée (voir section 2.1.2).

Exercice 2.2 (phénomène de stabilité non-linéaire)**.** (⋆) On considère le système suivant :

$$\begin{cases} \dot{x} = \alpha y + \lambda x^3 \\ \dot{y} = \beta x + \mu y^3 \end{cases}$$

avec $\alpha, \beta, \lambda, \mu \in \mathbb{R}$ des paramètres.

1) Linéariser l'équation au voisinage de $(0, 0)$ et étudier la stabilité linéaire en fonction des valeurs de $\alpha, \beta, \lambda, \mu \in \mathbb{R}$ avec l'aide du lemme 2.5.

2) On travaille pour l'instant dans le cas où $\lambda, \mu = 0$, c'est-à-dire dans le cas linéaire.

2.1) Donner l'allure des solutions en fonction de α et de β.

2.2) Montrer que la dynamique est hamiltonienne dont on précisera le hamiltonien $\mathcal{H}$. Vérifier, en dérivant au cours du temps, que le hamiltonien $\mathcal{H}$ est une quantité conservée.

3) Toujours dans le cas où $\lambda, \mu = 0$, on suppose désormais que $\alpha < 0 < \beta$.

3.1) Montrer que la dynamique reste dans une ellipse dont on donnera l'équation.

3.2) Donner le sens de rotation du système sur cette ellipse (sens horaire ou trigonométrique).

3.3) Montrer que $(0, 0)$ est Lyapunov stable.

4) On suppose toujours que $\alpha < 0 < \beta$, mais désormais $\lambda, \mu < 0$.

4.1) Montrer que le terme cubique est un terme dissipatif.

4.2) En déduire que la solution est globale pour $t \in \mathbb{R}_+$.

4.3) Montrer que $(0, 0)$ est asymptotiquement stable.

4.4) Montrer l'inégalité de Young ($2ab \leq a^2 + b^2$) puis montrer que si $a, b \geq 0$ alors

$$a^2 + b^2 \leq (a + b)^2 \leq 2(a^2 + b^2).$$

En déduire que le hamiltonien vérifie :

$$\frac{\mathrm{d}\mathcal{H}}{\mathrm{d}t} \lesssim -\mathcal{H}^2.$$

4.5) Montrer que, dans l'asymptotique $t \to +\infty$, nous avons : $\mathcal{H} \lesssim \frac{1}{t}$.

4.6) Conclure : $|x| \lesssim \frac{1}{\sqrt{t}}$ et $|y| \lesssim \frac{1}{\sqrt{t}}$.

5) On suppose que α et β sont de signes opposés et que $\lambda, \mu > 0$.

5.1) Montrer que $(0, 0)$ est un état instable.

5.2) Établir l'estimation suivante (expliciter les constantes) : $\frac{\mathrm{d}\mathcal{H}}{\mathrm{d}t} \simeq \mathcal{H}^2$.

5.3) En déduire que le système dynamique n'est pas globalement bien posé.

Exercice 2.3 (Stabilité pour des potentiels 1D)**.** ($\star$) On s'intéresse à la stabilité des états stationnaires pour des équations de la mécanique hamiltonienne en dimension 1 :

$$\frac{\mathrm{d}^2 x}{\mathrm{d}t^2} = -\nabla \mathcal{P}(x).$$

Étudier la stabilité des état stationnaires pour les potentiels suivants :

$$\mathcal{P}(x) = x^2, \qquad \mathcal{P}(x) = -x^4, \qquad \mathcal{P}(x) = x^3,$$
$$\mathcal{P}(x) = x^4 - 2x^2, \qquad \mathcal{P}(x) = \cos(x), \qquad \mathcal{P}(x) = 0.$$

Que se passe-t-il dans si on ajoute des frottements linéaires de la forme $-\lambda \dot{x}$ avec $\lambda > 0$.

Exercice 2.4 (Mouvement de Poinsot et effet Dzhanibekov)**.** ($\star$) On considère un solide ellipsoïdal en apesanteur qui tourne sur lui-même dans les 3 directions d'espace (Mouvement de Poinsot). Ses 3 principaux moments d'inertie sont notés $0 < I_1 < I_2 < I_3$. Les équations d'évolution des 3 vitesses angulaires associées $\omega_1, \omega_2, \omega_3$ sont données par le théorème du moment cinétique :

$$\begin{cases} I_1\,\dot{\omega}_1 = (I_2 - I_3)\,\omega_3\,\omega_2 \\ I_2\,\dot{\omega}_2 = (I_3 - I_1)\,\omega_1\,\omega_3 \\ I_3\,\dot{\omega}_3 = (I_1 - I_2)\,\omega_2\,\omega_1 \end{cases} \tag{2.35}$$

On introduit l'énergie cinétique de rotation $\mathcal{E}(\omega) := \frac{I_1}{2}\omega_1^2 + \frac{I_2}{2}\omega_2^2 + \frac{I_3}{2}\omega_3^2$ et le carré du module du moment cinétique $\mathcal{N}(\omega) := I_1^2\omega_1^2 + I_2^2\omega_2^2 + I_3^2\omega_3^2$.

1) Montrer que $\mathcal{E}$ et $\mathcal{N}$ sont des quantités conservées par la dynamique.

2) Montrer que la dynamique est globale en temps.

3) Quels sont les états stationnaires ?

4) Étudier la stabilité linéaire des états stationnaires (effet Dzhanibekov) [1].

Exercice 2.5 (Retour sur le système de Lotka-Voltera)**.** ($\star$) Cet exercice est dans la continuité de l'exercice sur le système de Lotka-Voltera (Exercice 1.7).

1) Pour chaque état stationnaire, linéariser le système au voisinage de cet état. Donner l'allure de la solution du système linéarisé.

2) Étudier la stabilité linéaire de ces états stationnaires à l'aide du théorème du linéarisé.

3) On pose à présent $p(t) := \ln(y(t))$ et $q(t) := \ln(x(t))$. Écrire les équations différentielles satisfaites par (p, q) et montrer que ce nouveau système est hamiltonien.

4) Étudier la stabilité de Lyapunov du point $(p, q) = (0, 0)$ en utilisant le hamiltonien comme fonction de Lyapunov. En déduire que le point $(x, y) = (1, 1)$ est Lyapunov stable pour la dynamique initiale.

5) Montrer que le hamiltonien $\mathcal{H}$ est une fonction fortement convexe sur $(\mathbb{R}_+^*)^2$ et qu'elle admet un unique minimum global. Tracer qualitativement les trajectoires $t \mapsto (x(t), y(t))$.

Exercice 2.6 (Formulation lagrangienne des EDP de transport)**.** ($\star$)

Soit $\mathcal{F} : [0, +\infty) \times \mathbb{R}^d \to \mathbb{R}^d$ un champ de vitesse globalement lipschitzien pour sa deuxième variable au sens où

$$\lambda_{\mathcal{F}} := \sup_{t \geq 0} \sup_{X \neq Y \in \mathbb{R}^d} \frac{|\mathcal{F}(t, X) - \mathcal{F}(t, Y)|}{|X - Y|} < +\infty.$$

1) Montrer que l'équation différentielle $\dot{X} = \mathcal{F}\big(t, X(t)\big)$ est globalement bien posée pour toute donnée initiale $X \in \mathbb{R}^d$.

1. Il est possible (mais plus difficile) d'utiliser les deux quantités conservées $\mathcal{E}$ et $\mathcal{N}$ pour démontrer que les autres états stationnaires sont stables. Il s'agira dans ce cas-là de stabilité quadratique et non pas linéaire.

2) On note $\mathscr{S}^t$ le flot associé à cette équation. Pour une fonction $g_0 \in \mathcal{C}^1(\mathbb{R}^d;\mathbb{R})$ fixée, on définit [1] $g(t,X) := g_0(\mathscr{S}^{-t}X)$, la fonction transportée par le flot. Représenter graphiquement l'évolution de g au cours du temps pour l'équation $\dot{x} = 1$ (en dimension 1). Même question avec les équations $\dot{x} = x$, puis $\dot{x} = -x$ et enfin $\dot{x} = -x\sin(t)$. Que constate-t-on ?

3) Montrer en toute généralité que g est une solution de l'EDP de transport :

$$\frac{\partial g}{\partial t}(t,X) + \mathcal{F}(t,X)^T \nabla_x g(t,X) = 0.$$

Exercice 2.7 (Retour sur l'oscillateur harmonique)**.** Cet exercice revient sur l'oscilateur harmonique étudié lors de l'exercice 1.9. On rappelle son équation sans source extérieure :

$$m\ddot{x}(t) + \lambda\dot{x} + kx(t) = 0, \tag{2.36}$$

avec m la masse de l'objet, k la raideur du ressort, λ le coefficient de frottement et avec $x : [0,+\infty[\to \mathbb{R}$ la position de l'objet au cours du temps. L'objectif de cet exercice est de manipuler les concepts du cours sur une équation simple (il ne faut donc pas s'aider de la formule qui donne les solutions exactes...).

1) En introduisant la vitesse $v(t) := \dot{x}(t)$ comme variable auxiliaire du système, montrer que l'équation étudiée est équivalente à

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} x \\ v \end{pmatrix}(t) = -JA\begin{pmatrix} x \\ v \end{pmatrix}(t) + \begin{pmatrix} 0 \\ -\mu v \end{pmatrix},$$

où $\mu > 0$ s'exprime à partir des données du problème, J est la matrice de rotation d'angle $\pi/2$ et A est une matrice constante dont on précisera les coefficients.

2) Pour cette question, on travaille sans frottements : $\lambda = 0$.

2.1) Montrer qu'il existe une fonction quadratique $\mathcal{H}$ telle que

$$\nabla_{x,v}\mathcal{H}(x,v) \;=\; A\begin{pmatrix} x \\ v \end{pmatrix}.$$

2.2) En déduire que la dynamique est hamiltonienne (préciser son hamiltonien).

2.3) Montrer que le hamiltonien est une quantité conservée.

2.4) En utilisant le théorème de stabilité de Lyapunov, montrer que $(0,0)$ est un état stable de la dynamique.

3) On travaille désormais *avec* frottements : $\lambda > 0$.

3.1) Montrer que $(0,-\mu v)$ est un terme dissipatif et montrer que le hamiltonien décroît au cours de temps.

3.2) Pour $a,b \in \mathbb{R}$, on considère la fonction suivante :

$$\mathcal{L}_{a,b} \;:=\; ax^2 + 2bxv + v^2.$$

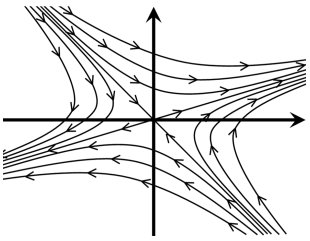
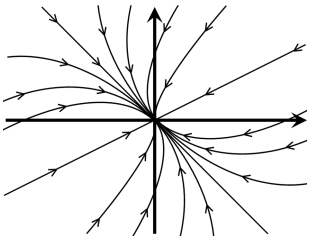
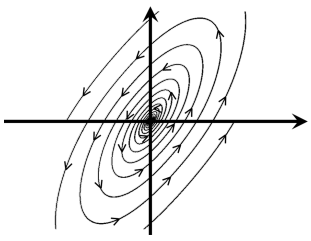
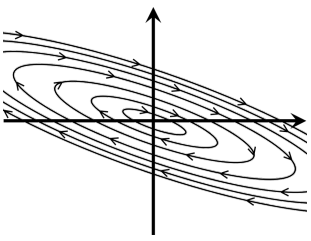
En ajustant les paramètres a et b, trouver une fonction de Lyapunov pour démontrer que l'origine est un point critique asymptotiquement stable.

3.3) Montrer que cette fonction de Lyapunov est une norme. A l'aide du lemme de Grönwall, montrer qu'elle décroît exponentiellement.

1. La fonction g est bien définie car le caractère bien-posé de l'équation implique que pour tout $t > 0$ la fonction flot $\mathscr{S}^t : \mathbb{R}^d \to \mathbb{R}^d$ est une fonction bijective. On peut réécrire cette définition sous la forme lagrangienne : $g(t,\mathscr{S}^t X) = g_0(X)$.

Exercice 2.8 (4 exemples de dynamiques linéaires)**.**

On souhaite étudier en détail les 4 systèmes d'équations linéaires suivants :

$\begin{cases} \dot{x} = x + 3y \\ \dot{y} = x - y \end{cases}$	$\begin{cases} \dot{x} = -3x - 2y \\ \dot{y} = -x - 2y \end{cases}$	$\begin{cases} \dot{x} = 3x - y \\ \dot{y} = 2x + y \end{cases}$	$\begin{cases} \dot{x} = 2x + 6y \\ \dot{y} = -x - 2y \end{cases}$
			
axe stable : $y = -x$ axe instable : $3y = x$	axe rapide : $x = 2y$ axe lent : $x = -y$	grd. axe : $y = \frac{1+\sqrt{5}}{2}x$ rotation sens trigo.	grd. axe : $x = \frac{-5-\sqrt{41}}{4}y$ rotation sens horaire

1) Concernant le premier système d'équations,

1.1) Donner l'allure des solutions.

1.2) Montrer qu'il y a 2 droites laissées invariantes par la dynamique et trouver leurs équations. Montrer qu'il y a une direction le long de laquelle $(0, 0)$ est asymptotiquement stable et que l'autre est instable.

1.3) Montrer que la dynamique suit des branches d'hyperboles. Pour cela, trouver les équations des hyperboles laissées stables par le flot.

1.4) En déduire le dessin tracé pour le premier système.

2) Concernant le deuxième système d'équations,

2.1) Donner l'allure des solutions.

2.2) Montrer qu'il y a 2 droites laissées invariantes par la dynamique et trouver leurs équations. Montrer qu'il y a une direction le long de laquelle la dynamique est plus rapide d'un exposant 4 par rapport à l'autre.

2.3) En déduire que les trajectoires sont des courbes quartiques (c'est-à-dire de la forme $y = x^4$) en trouvant toutes les courbes quartiques laissées stables par le flot.

2.4) En déduire le dessin tracé pour le deuxième système.

3) Concernant le troisième système d'équations,

3.1) Donner l'allure des solutions.

3.2) Cette équation laisse invariantes des spirales logarithmiques à profil elliptique. Donner les équations de ces spirales (en coordonnées polaires : c'est plus simple).

3.3) En déduire le dessin tracé pour le troisième système.

4) Concernant le quatrième système d'équations,

4.1) Donner l'allure des solutions.

4.2) Montrer qu'il s'agit d'une équation de la mécanique hamiltonienne.

4.3) Montrer que cette équation conserve un ensemble d'ellipses du plan dont on donnera les équations ainsi que la direction du grand axe.

4.4) En déduire le dessin tracé pour le quatrième système.

Exercice 2.9 (le pendule simple)**.** L'équation du pendule simple est la suivante :

$$\ddot{\theta} + \lambda\dot{\theta} + \omega_0^2 \sin(\theta) = 0,$$

avec λ le coefficient de dissipation et ω_0 la fréquence caractéristique.

1) Montrer que cette équation est globalement bien posée si on se donne au temps initial un angle $\theta(0) = \theta_0$ et une vitesse angulaire $\dot{\theta}(0) = \theta_1$.

2) Montrer que pour $\lambda = 0$ on a préservation de l'énergie mécanique au cours du temps $E := \dot{\theta}/2 + \mathcal{P}(\theta)$ où le potentiel $\mathcal{P}$ est une fonction à expliciter.

3) Trouver les solutions constantes de cette équation et montrer qu'elles coïncident avec les points-critiques de $\mathcal{P}$.

4) On va maintenant étudier le système dans le cas dissipatif $\lambda > 0$ et montrer que ce système différentiel converge linéairement (c'est-à-dire à vitesse exponentielle).

4.1) Calculer la dissipation instantanée d'énergie mécanique (en fonction de $\dot{\theta}$).

4.2) Pour $a, b > 0$ on considère la quantité suivante :

$$\Gamma_{a,b} \ := \ \frac{\dot{\theta}^2 + 2a\dot{\theta}\sin(\theta) + b\sin^2(\theta)}{2}.$$

En utilisant le fait que $2xy \leq x^2 + y^2$, trouver les valeurs de $a > 0$ pour lesquelles cette quantité est toujours positive quel que soit les valeurs de θ et $\dot{\theta}$.

4.3) Montrer que pour a et b assez petit on a

$$\frac{\mathrm{d}}{\mathrm{d}t}\Gamma_{a,b} \ \leq \ -C\Gamma_{a,b},$$

où C est une constante strictement positive à expliciter (en fonction de a, b, λ et ω_0^2).

4.4) En utilisant le lemme de Grönwall, montrer que $\Gamma_{a,b}$ tend vers 0 lorsque $t \to +\infty$ avec une vitesse exponentielle.

4.4) En déduire que le système converge vers un point-critique du potentiel $\mathcal{P}$.

5) On travaille maintenant dans le cadre $\lambda = 0$. On appelle *séparatrices* les solutions θ telles que $E(\theta) = \max \mathcal{P}$. On rappelle les formules de trigonométrie suivantes :

$$\sin(\arctan(x)) = \frac{x}{\sqrt{1+x^2}}, \qquad \text{et} \qquad \cos(\arctan(x)) = \frac{1}{\sqrt{1+x^2}}.$$

5.1) Montrer que $t \mapsto 2\arctan(\sinh(t))$ est solution de $\ddot{\theta} + \sin(\theta) = 0$. En déduire la formule générale pour toutes les séparatrices.

5.2) Montrer que les solutions sont bornées si et seulement si elles ont une énergie inférieure aux séparatrices. Montrer que si leur énergie est strictement inférieure aux séparatrices alors elles sont périodiques.

5.3) Si elles ont une énergie strictement supérieure aux séparatrices, montrer que $t \mapsto \sin(\theta(t))$ est périodique.

5.4) En déduire le diagramme des phases.

6) Nous allons à présent utiliser la fonction $\Gamma_{a,b}$ pour étudier les propriétés hamiltoniennes du pendule simple.

6.1) Montrer que cette fonction est une fonction de Lyapunov.

6.2) Étudier la stabilité des points stationnaires du pendule simple (distinguer les cas avec ou sans frottements).

6.3) Montrer que le pendule simple est un système mécanique hamiltonien lorsque $\lambda = 0$ et dissipatif lorsque $\lambda > 0$.

7) Effectuer une linéarisation au voisinage des points stationnaire et étudier l'allure de la solution linéarisée.

Exercice 2.10 (Stabilité pour des potentiels 2D)**.** On s'intéresse à la stabilité des états stationnaires pour des équations de la mécanique hamiltonienne en dimension 2 :

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\begin{pmatrix} x \\ y \end{pmatrix} \;=\; -\nabla \mathcal{P}(x,y).$$

Étudier la stabilité des état stationnaires pour les potentiels suivants :

$$\mathcal{P}(x,y) = x^2 + xy + y^2, \qquad \mathcal{P}(x,y) = x^2 + 2xy + y^2, \qquad \mathcal{P}(x,y) = x^3 + y^2,$$
$$\mathcal{P}(x,y) = x^4 - 2x^2 + y^2, \qquad \mathcal{P}(x,y) = x^4 + y^4, \qquad \mathcal{P}(x,y) = x^2 + xy^2,$$
$$\mathcal{P}(x,y) = x^3 - 3xy^2, \qquad \mathcal{P}(x,y) = \cos(x)\cos(y), \qquad \mathcal{P}(x,y) = \cos(x^2 + 2y^2).$$

Que se passe-t-il si on ajoute des frottements linéaires de la forme $-\lambda\dot{x} - \mu\dot{y}$ avec $\lambda, \mu > 0$?

Exercice 2.11 (Équation de Landau-Lifschitz)**.** L'équation de Landau-Lifschitz modélise la dynamique de l'orientation moyenne du spin magnétique d'un atome. Cette équation a été dérivée par Landau et Lifschitz à partir de l'équation de Schrödinger et elle est une bonne approximation de l'évolution de l'aimantation spontanée du nuage électronique d'un atome si on regarde sur ces échelles de temps pas trop courtes. Un terme de dissipation issu de l'équation de Dirac est rajouté pour modéliser les effets relativistes à l'échelle atomique. L'équation de Landau-Lifschitz s'écrit :

$$\frac{\mathrm{d}M}{\mathrm{d}t} = -M \wedge H - \alpha M \wedge (M \wedge H), \tag{2.37}$$

avec $M : [0, T[\to \mathbb{R}^3$ l'aimantation moyenne de l'atome, $H = he_z$ le champ magnétique extérieur supposé uniforme et constant dans la direction verticale, et $\alpha \geq 0$ la constante de Gilbert.

1) Montrer que la norme de M est conservée. En déduire le caractère globalement bien posé.

2) Puisque la norme est conservée, la dynamique de M a lieu sur la sphère (c'est donc une dynamique 2D). Écrire cette équation en coordonnées sphériques.

3) Lorsque $\alpha = 0$, montrer que cette équation est hamiltonienne. Montrer que les trajectoires sont périodiques sur des cercles.

4) Lorsque $\alpha > 0$, montrer que cette équation est dissipative.

5) Étudier la stabilité des états stationnaires (distinguer les cas selon $\alpha = 0$ ou $\alpha > 0$).

6) Donner l'allure des solutions en étudiant le linéarisé au voisinage des états stationnaires.

*
* *

Chapitre 3

Analyse numérique et approximations

Dans de nombreuses situations il n'est pas possible de trouver une formule explicite pour résoudre les équations différentielles ordinaires dans leur formulation générale. Pour pouvoir étudier le comportement des solutions d'une équation différentielle nous avons proposé plusieurs outils d'analyse lors des chapitres précédents. Cependant, ces outils montrent vite leurs limites lorsqu'il s'agit de récupérer des informations précises sur le comportement de la solution pour des applications pratiques et techniques.

L'analyse numérique et la théorie des approximations permet d'approcher la solution exacte à l'aide d'un algorithme que l'on peut programmer sur un ordinateur et ainsi obtenir un résultat exploitable pour des applications dans le cadre d'un travail d'ingénieur. L'idée principale de l'analyse numérique des équations différentielles consiste, étant donné une suite croissante de temps (t_n) finie, à fabriquer récursivement une suite (X_n) qui est une approximation au temps (t_n) de la fonction X solution exacte au problème :

$$X_n \approx X(t_n).$$

Pour que cette approximation soit la meilleure possible, il faut réaliser une bonne approximation de l'équation différentielle elle-même, et en particulier de la dérivée temporelle de X. C'est-à-dire que l'on cherche des approximations (X_n) de la forme :

$$\frac{X_{n+1} - X_n}{t_{n+1} - t_n} \;\approx\; \frac{\mathrm{d}X}{\mathrm{d}t}(t_n). \tag{3.1}$$

La théorie de l'analyse numérique des équations différentielles, introduite dans ce chapitre, se développe dans la continuité de l'analyse classique des équations différentielles présentée aux chapitres 1 et 2.

3.1 Discrétisation des équations différentielles

3.1.1 Premier exemple : Euler explicite et implicite

On considère une équation différentielle mise sous la forme d'un problème de Cauchy :

$$\frac{\mathrm{d}}{\mathrm{d}t}X(t) \;=\; \mathcal{F}\big(t, X(t)\big), \qquad \text{et} \qquad X(0) = X_0, \tag{3.2}$$

avec $\mathcal{F} : [0,T] \times \Omega \to \mathbb{R}^d$ le champs de vitesse, $X_0 \in \Omega$ la donnée initiale et $X : [0,T] \to \Omega$ l'inconnue du problème.

On commence par discrétiser le temps en se donnant un pas de temps $\Delta t > 0$ fixé et on pose $t_0 = 0$ et $t_{n+1} := t_n + \Delta t = n\Delta t$. On note également par N le nombre total de pas de

temps et on doit ajuster N et Δt de sorte que $T = N\Delta t$, où T est le temps final pour l'étude du système dynamique. On construit alors la suite finie $(X_n)_{n=0}^N \in \Omega^{N+1}$ par un algorithme récursif que l'on initialise avec la donnée initiale $X_0 \in \Omega$ et cette suite est construite à l'aide de l'une des différentes méthodes numériques existantes.

Euler explicite

Le premier exemple de discrétisation, le plus naturel et le plus simple, est l'algorithme de Euler-Explicite proposé par le mathématicien suisse Léonhard Euler en 1768. Il s'agit du premier exemple historique d'algorithme de discrétisation d'une équation différentielle. Elle consiste tout simplement à remplacer formellement, dans l'approximation (3.1), la valeur de $\mathrm{d}X/\mathrm{d}t$ au temps $t = t_n$ par sa valeur dans l'équation, à savoir $\mathcal{F}(t_n, X_n)$. On aboutit naturellement à la définition suivante :

Définition 3.1 Méthode de Euler Explicite (‡)

On se donne un problème de Cauchy (3.2), un pas de temps $\Delta t > 0$ et un nombre d'itérations N. On appelle méthode d'***Euler Explicite*** l'algorithme itératif suivant :

$$X_{n+1} \;:=\; X_n + \Delta t\, \mathcal{F}\big(t_n, X_n\big).$$

Pour la méthode d'Euler-Explicite, on obtient bien une forme algorithmique itérative à un pas, c'est-à-dire de la forme : $X_{n+1} = \Phi(n, X_n)$. Dans notre cas, la fonction Φ est donnée par

$$\Phi(n, X) \;=\; X + \Delta t\, \mathcal{F}\big(t_n, X\big).$$

Euler implicite

Une autre approximation naturelle dans la continuité de Euler-Explicite consiste à reprendre le développement limité (3.1) et à le réécrire dans l'autre sens (c'est-à-dire en inversant le rôle joué par le temps t_n et le temps t_{n+1}) :

$$\frac{X_{n+1} - X_n}{\Delta t} \;\approx\; \frac{\mathrm{d}}{\mathrm{d}t} X(t_{n+1}).$$

En remplaçant également à nouveau $\mathrm{d}X/\mathrm{d}t$ par sa valeur donnée par l'équation exacte (mais cette fois-ci au temps t_{n+1}), à savoir $\mathcal{F}(t_{n+1}, X_{n+1})$, on obtient :

Définition 3.2 Méthode de Euler Implicite (‡)

On se donne un problème de Cauchy (3.2), un pas de temps $\Delta t > 0$ et un nombre d'itérations N. On appelle méthode (ou schéma) d'***Euler Implicite*** l'algorithme itératif suivant :

$$X_{n+1} \;:=\; X_n + \Delta t\, \mathcal{F}\big(t_{n+1}, X_{n+1}\big).$$

Contrairement à Euler Explicite, il est beaucoup moins clair ici qu'il soit possible de mettre cet algorithme sous forme itérative. Si on note par $\mathcal{I}_d$ la fonction identité de $\mathbb{R}^d$, on peut réécrire cet algorithme comme suit :

$$\Big(\mathcal{I}_d - \Delta t\, \mathcal{F}\big(t_{n+1}, \mathcal{I}_d\big)\Big)(X_{n+1}) \;=\; X_n.$$

Lorsque Δt est assez petit (par rapport à la constante de Lipschitz de $\mathcal{F}$), la fonction apparaissant à gauche de cette équation devient localement bijective. On peut alors faire une inversion

de fonction et réécrire Euler-Implicite sous forme itérative $X_{n+1} = \Phi(n, X_n)$ avec :

$$\Phi(n, X) \;=\; \Big(\mathcal{I}_d - \Delta t\, \mathcal{F}\Big(t_{n+1},\, \mathcal{I}_d\Big)\Big)^{-1}(X).$$

Méthodes explicites et implicite

Ces deux méthodes sont très simples mais suffisamment riches pour expliquer les principaux concepts et difficultés de l'analyse numérique.

Concernant Euler-Explicite, comme nous le verrons par la suite, cette méthode pose des problèmes de stabilité lorsque le pas de temps est trop important : l'algorithme cesse d'être une bonne approximation de la dynamique car il accumule trop d'erreurs et les amplifie exponentiellement. Il prend alors des valeurs absurdes. Il faut faire des petits pas de temps Δt pour être sûr que l'algorithme renvoie des valeurs exploitables.

Euler-Implicite est une méthode beaucoup plus stable. En revanche, comme son nom l'indique, c'est un schéma *implicite*. C'est-à-dire que pour pouvoir l'écrire sous forme itérative $X_{n+1} = \Phi(n, X_n)$, il est nécessaire de faire le calcul d'une inversion de fonction (ce qui revient à faire une extraction de racine). De manière générale, les méthodes implicites sont souvent beaucoup plus stables que les méthodes explicites mais posent la difficulté supplémentaire de devoir résoudre ce problème d'inversion. Cela qui se révèle parfois très coûteux en terme de temps de calcul (nous y reviendrons dans la section 3.3 consacrée à l'implémentation informatique des méthodes numériques).

3.1.2 Méthodes Numériques et Systèmes Dynamiques Discrets

Afin de pouvoir mener une étude de l'analyse numérique et de l'approximation des équations différentielles, nous allons poser ici quelques définitions. On commence par le concept de base : *les systèmes dynamiques discrets.*

Définition 3.3 Systèmes dynamiques discrets

Soit $\Phi : \mathbb{N} \times \Omega \to \Omega$ (avec Ω ouvert de $\mathbb{R}^d$). On appelle ***système dynamique discret*** sur Ω engendré par Φ l'ensemble des suites $(X_n) \in \Omega^{\mathbb{N}}$ définies par la relation de récurrence :

$$X_{n+1} \;=\; \Phi(n, X_n). \tag{3.3}$$

La fonction Φ est appelée ***fonction de récursion.***

Par analogie, il est fréquent de dire qu'une équation différentielle ordinaire est un *système dynamique continu*, au sens où la solution est paramétrée par un paramètre continu (le temps $t \in \mathbb{R}$) et non un paramètre discret (l'indice $n \in \mathbb{N}$). Si la fonction de récursion Φ est constante par rapport à la variable n, on dit que le système dynamique est *autonome.*

Étant donné une position $X_0 \in \Omega$ et un système dynamique engendrée par une fonction Φ fixée, il existe un unique suite $(X_n) \in \Omega^{\mathbb{N}}$ vérifiant la relation de récurrence (3.3) et ayant X_0 comme position initiale lorsque $n = 0$. En conséquence, on peut définir le flot discret :

Définition 3.4 Flot discret

Étant donné un système dynamique discret engendré par une fonction de récursion Φ, on note $\mathbf{S}^n X$ la valeur à l'itération $n \in \mathbb{N}$ engendrée par la relation de récurrence (3.3) initialisée avec la donnée $X \in \Omega$.

La fonction $\mathbf{S}^n : X \in \Omega \longmapsto \mathbf{S}^n X$ s'appelle le *flot discret* à l'itération n associé au système dynamique engendré par Φ. Si le système dynamique n'est pas autonome, il faut alors préciser l'indice $n_0 \in \mathbb{N}$ à partir duquel on travaille pour définir le flot $\mathbf{S}^n_{n_0}$. De même que dans le cas continu, les flots discrets vérifient la relation de morphisme de semi-groupe : $\mathbf{S}^{n_2}_{n_1} \circ \mathbf{S}^{n_1}_{n_0} = \mathbf{S}^{n_2}_{n_0}$. Dans le cas autonome, on a $\mathbf{S}^{n_1}_{n_0} = \mathbf{S}^{n_1-n_0}$ et on a aussi le morphisme : $\mathbf{S}^{n_2} \circ \mathbf{S}^{n_1} = \mathbf{S}^{n_1+n_2}$ (ce qui est bien un morphisme de groupe si on étend le flot aux indices négatifs $n \in \mathbb{Z}$).

De la même manière que dans le cas continu, on peut définir la notion de *quantité conservée* ou de *quantité dissipée* :

Définition 3.5 Quantités conservées ou dissipées par les systèmes discrets

Étant donné un système dynamique discret donné par un flot $(\mathbf{S}^n)_{n\in\mathbb{N}}$, on dit que la fonction $\mathcal{G} : \Omega \to \mathbb{R}$ est une ***quantité conservée*** par le flot discret ssi

$$\forall\, n \in \mathbb{N}, \quad \forall\, X \in \Omega, \qquad \mathcal{G}(\mathbf{S}^{n+1}X) \;=\; \mathcal{G}(\mathbf{S}^n X).$$

On dit que $\mathcal{G}$ est une ***quantité dissipée*** ssi

$$\forall\, n \in \mathbb{N}, \quad \forall\, X \in \Omega, \qquad \mathcal{G}(\mathbf{S}^{n+1}X) \;\leq\; \mathcal{G}(\mathbf{S}^n X).$$

De la même manière, la théorie de la stabilité des solutions se généralise naturellement au cas des systèmes dynamiques discrets :

Définition 3.6 Stabilité pour les systèmes dynamiques discrets

(i) Étant donné un flot discret $(\mathbf{S}^n)_{n\in\mathbb{N}}$, on dit que la suite $(\mathbf{S}^n X_0)_{n\in\mathbb{N}}$ est ***(Lyapunov) stable*** ssi pour tout $\varepsilon > 0$, il existe un $\delta > 0$ tel que

$$|X - X_0| \leq \delta \qquad \Longrightarrow \qquad \sup_{n\in\mathbb{N}} \left|\mathbf{S}^n X - \mathbf{S}^n X_0\right| \leq \varepsilon.$$

(ii) Elle est ***asymptotiquement stable*** ssi il existe $\delta > 0$ tel que

$$|X - X_0| \leq \delta \qquad \Longrightarrow \qquad \left|\mathbf{S}^n X - \mathbf{S}^n X_0\right| \;\longrightarrow\; 0, \quad \text{lorsque } n \to +\infty.$$

(iii) Elle est ***linéairement stable*** ssi il existe $\delta > 0$ et $\lambda > 0$ tels que

$$|X - X_0| \leq \delta \qquad \Longrightarrow \qquad \left|\mathbf{S}^n X - \mathbf{S}^n X_0\right| \;\leq\; |X - X_0| e^{-\lambda n}.$$

Dans le cadre de notre travail, nous allons étudier uniquement les systèmes dynamiques discrets issus de l'approximation d'un système dynamique continu. En particulier, ce sont des systèmes dynamiques discrets qui dépendent de 2 paramètres : le pas de temps de la discrétisation Δt et le choix du champs de vitesse $\mathcal{F} : [0, T) \times \Omega \to \mathbb{R}^d$ pour l'équation différentielle initiale. Afin de faciliter la lecture, les paramètres seront notés entre crochets (et omis s'il n'y a pas d'ambiguïtés). En particulier, la fonction génératrice du système dynamique discret Φ se note $(n, X) \mapsto \Phi[\Delta t, \mathcal{F}](n, X)$ et le flot discret $X \mapsto \mathbf{S}^n[\Delta t, \mathcal{F}]X$. De même on notera $X \mapsto \mathscr{S}^t[\mathcal{F}]X$ le flot continu associé au champs de vitesse $\mathcal{F}$.

Définition 3.7 Méthode numérique

On appelle ***méthode numérique*** une famille de systèmes dynamiques discrets paramétrée par un réel positif $\Delta t \in \mathbb{R}_+$ et par un champ de vitesse $\mathcal{F} : [0, T) \times \Omega \to \mathbb{R}^d$.

Le flot discret $X \mapsto \mathbf{S}^n[\Delta t, \mathcal{F}]X$ associé à une méthode numérique est donc un flot paramétré par Δt et par $\mathcal{F}$ et s'appelle un ***flot numérique.*** Une fois fixé le champs de vitesse $\mathcal{F}$, on parle alors de *schéma numérique*[1]. La théorie de l'approximation numérique est présentée sur des flots numériques généraux pour présenter les concepts et théorèmes fondamentaux. A partir de la section 3.2, on travaillera avec des méthodes numériques particulières pour lesquelles la fonction génératrice Φ est donnée par une formule analytique explicite en fonction de $\mathcal{F}$ et Δt (méthodes d'Euler par exemple, et leurs généralisations).

3.1.3 Le théorème de convergence numérique

Stabilité d'une méthode numérique

La notion de stabilité pour les systèmes dynamiques discrets est pertinente pour étudier la façon dont vont se propager les erreurs au sein d'un algorithme d'approximation numérique d'une équation différentielle ordinaire. L'origine de ces erreurs est de deux natures différentes. Une première source d'erreur provient des erreurs d'arrondis que fait l'ordinateur au cours de l'exécution de l'algorithme (en effet, il n'est pas possible de faire stocker à une machine la valeur exacte obtenue par l'algorithme puisque celle-ci est nécessairement stockée sur un nombre fini de bits). L'autre source d'erreur est celle qui provient de l'approximation réalisée en passant du système continu au système discret.

Cependant, ces notions de stabilité très générales ne sont pas directement adaptées à la description des instabilités numériques des méthodes de discrétisation pour deux raisons :

1. Lorsqu'on intègre une EDO à l'aide d'un ordinateur, on ne peut pas intégrer jusqu'à $T = +\infty$ puisque la mémoire et le temps de calcul de l'ordinateur est finie. Il nous faut donc une notion de stabilité valable sur les intervalles de temps bornés $[0, T]$.

2. Les notions de stabilité générales ne font pas apparaître la dépendance du flot numérique par rapport au pas de temps Δt ni par rapport au champs de vitesse $\mathcal{F}$, alors que ces paramètres sont cruciaux dans le comportement de la solution numérique.

Nous allons donc nous inspirer de ces définitions pour proposer des outils directement adapté à l'analyse numérique.

Définition 3.8 Stabilité des schémas numériques **(†)**

On considère un schéma numérique défini par un flot numérique discret $(\mathbf{S}^n)_{n\in\mathbb{N}}$ et un champ de vitesses $\mathcal{F} : [0, T) \times \Omega \to \mathbb{R}^d$ fixé.

(i) On dit que ce schéma numérique est ***localement stable*** si pour tout temps final $T > 0$, il existe un *pas de temps critique* $\Delta t^\star = \Delta t^\star[\mathcal{F}, T]$ et une constante de stabilité $C = C[\mathcal{F}, T] > 0$ tels que :

$$\forall\, X, Y \in \Omega, \quad \forall\, \Delta t < \Delta t^\star, \quad \sup_{\substack{n\in\mathbb{N} \\ n\Delta t\leq T}} \left|\mathbf{S}^n[\Delta t, \mathcal{F}]X - \mathbf{S}^n[\Delta t, \mathcal{F}]Y\right| \;\leq\; C\,|X - Y|. \qquad (3.4)$$

(ii) Si de plus la constante C peut être choisie indépendamment du temps final T, on dit que ce schéma numérique est ***stable*** ou bien ***numériquement stable.***

Remarque importante : L'ordre des quantificateurs dans cette définition est crucial car il faut avoir une estimation de l'erreur propagée par le flot numérique qui est **indépendante**

1. Dans la pratique, cette distinction n'est pas toujours respectée et il est fréquent d'utiliser les termes de *méthode numérique* et de *schéma numérique* indistinctement.

du pas de temps de discrétisation Δt (et donc du nombre d'itérations de l'algorithme). Cette uniformité de l'estimation par rapport à Δt nous permet de faire un lien (asymptotiquement) entre la stabilité de Lyapunov pour les systèmes dynamiques discrets et la stabilité numérique : en effet, lorsque $\Delta t \to 0$, le nombre d'itérations n telles que $t_n := n\Delta t$ soit inférieur au temps final $T > 0$ va tendre vers $+\infty$ (on se retrouve bien - *moralement* - à étudier un comportement asymptotique pour une famille de systèmes dynamiques discrets).

Si de plus dans la définition (ii) ci-dessus il est possible de choisir $\Delta t^\star = +\infty$, alors on dit que le schéma est inconditionnellement (numériquement) stable.

On va à présent présenter les notions de stabilité pour les méthodes numériques dont nous allons nous servir dans ce cours. Afin de motiver les définitions qui suivent, il faut souligner qu'il n'est pas possible en général d'espérer une bonne approximation des solutions d'une EDO lorsque celles-ci divergent (même en temps infini). Il est également très difficile d'obtenir une bonne approximation si la solution est très irrégulière. La question de la stabilité numérique pour nous se posera donc dans un cadre restreint qui est celui des *flots numériquement admissibles* :

Définition 3.9 Flots numériquement admissibles

On dit qu'un champ de vitesse $\mathcal{F}$ engendre un *flot numériquement admissible* si $\mathcal{F}$ est globalement lipschitz est les solutions de $\dot{X} = \mathcal{F}(t, X(t))$ sont toutes bornées. Autrement dit

$$\forall\, X \in \Omega, \qquad \sup_{t\geq 0} \left| \mathscr{S}^t[\mathcal{F}]X \right| < +\infty. \tag{3.5}$$

Par métonymie, on dira alors que $\mathcal{F}$ est numériquement admissible. La stabilité d'une méthode numérique est définie par :

Définition 3.10 Stabilité des méthodes numériques **(†)**

Soit une méthode numérique définie par un flot numérique discret $(\mathbf{S}^n)_{n\in\mathbb{N}}$ paramétré par $\mathcal{F}$ et Δt.

(i) On dit que cette méthode numérique est ***stable sous condition*** si, pour tout champ de vitesse $\mathcal{F}$ numériquement admissible, le schéma numérique $\Delta t \longmapsto \mathbf{S}^n[\Delta t, \mathcal{F}]$ est numériquement stable.

(ii) Si de plus le pas de temps critique est toujours infini : $\Delta t^\star = +\infty$, alors la méthode est ***inconditionnellement stable.***

Autrement dit, dès qu'on applique la méthode numérique avec un champ de vitesse $\mathcal{F}$ qui engendre un flot borné on obtient un schéma stable au sens de la définition 3.8-(ii). On remarquera que la notion de *stabilité locale* telle qu'introduite par la définition 3.8-(i) n'est pas utilisée dans cette définition. De manière générale, elle est assez peu utilisée car lorsque la constante dépend du temps final, cette dépendance est souvent exponentielle ! Cependant, dans certains cas il n'est pas possible d'obtenir mieux.

Erreur de consistance et erreur totale

Pour pouvoir apprécier la qualité d'un schéma numérique, nous devons étudier sa précision, c'est-à-dire la différence entre les valeurs obtenues grâce à ce schéma et la solution exacte. Comme il n'est pas forcément possible de contrôler les erreurs sur des intervalles de temps infinis, on travaille sur un intervalle de temps borné $[0, T]$. On travaille avec un paramètre de

discrétisation temporelle $\Delta t > 0$ (ie : le *pas de temps*) choisi tel que $T = N\Delta t$ où $N \in \mathbb{N}$ est le nombre d'itérations de l'algorithme. Autrement dit, on ajuste Δt pour avoir $T/\Delta t \in \mathbb{N}$.

L'objectif est d'étudier *l'erreur totale* définie de la manière suivante, en notant $\mathscr{S}$ le flot exact et $\mathbf{S}$ le flot numérique :

Définition 3.11 Erreur totale (‡)

On appelle ***erreur totale*** la différence en norme ℓ^∞ entre la solution exacte et l'approximation numérique :

$$\mathrm{Err}(\Delta t) \ := \ \sup_{n=1\ldots N} |\mathscr{S}^{t_n} X - \mathbf{S}^n X|. \tag{3.6}$$

Il est également possible d'étudier la différence entre la solution exacte et son approximation numérique en utilisant d'autres normes pour les suites (les normes ℓ^1 et ℓ^2 sont notamment très utilisées). Dans le cadre de ce cours nous étudierons uniquement l'erreur en norme ℓ^∞ pour la suite de pas de temps. La valeur de l'erreur de convergence va dépendre du choix de la donnée initiale $X \in \Omega$ et du champs de vitesse $\mathcal{F}$, mais elle va également dépendre du choix du pas de temps Δt et du choix du temps final T. C'est la dépendance par rapport à Δt qui est la plus importante à étudier.

Il n'est pas aisé en général d'estimer l'erreur totale. La première erreur que nous pouvons estimer plus facilement à l'aide de développements limités est l'erreur de troncature locale. C'est-à-dire l'erreur que commet l'algorithme à chaque nouvelle itération. Les définitions se font également à l'aide de la notation flot :

Définition 3.12 Erreur de troncature locale et erreur de consistance (‡)

On définit l'***erreur de troncature locale*** à l'étape $n+1$, notée η^{n+1}, comme étant l'erreur commise par l'algorithme entre les instants t_n et t_{n+1} (pour le calcul de X_{n+1}) par rapport à la résolution exacte de l'équation différentielle sur $[t_n, t_{n+1}]$ avec $X(t_n) = \mathscr{S}^{t_n} X$ comme donnée initiale. Autrement dit,

$$\eta_{n+1}(\Delta t) \ := \ \left|\mathbf{S}_n^{n+1} X(t_n) - \mathscr{S}_{t_n}^{t_{n+1}} X(t_n)\right|.$$

L'***erreur de consistance*** du schéma est la somme des erreurs de troncature locale :

$$\eta(\Delta t) \ := \ \sum_{n=1}^{N} \eta_n(\Delta t).$$

Remarque : L'erreur de consistance *n'est pas* l'erreur totale qui est la différence entre la solution exacte et la solution numérique. En effet, lorsqu'on calcule l'erreur de troncature locale, on regarde le flot à partir du point $X(t_n)$ (qui lui-même n'est connu que par une approximation X_n) alors que dans la formule (3.6), on considère le flot discret et continu depuis l'origine. Il y a donc des erreurs qui s'accumulent au fur-et-à-mesure de l'exécution de l'algorithme de résolution numérique *qui ne seront pas prise en compte* lors du calcul de l'erreur de consistance !

Énoncé du théorème de convergence numérique

Pour énoncer le théorème de convergence numérique, il nous faut introduire deux notions de convergence (dont nous allons montrer qu'elles sont équivalentes) :

Définition 3.13 Méthode numérique convergente (†)

On dit qu'une méthode numérique est ***convergente*** si pour tout champ de vitesse $\mathcal{F}$ globalement lipschitzien pour sa variable d'espace, pour toute donnée initiale $X \in \Omega$ et pour tout temps final $T > 0$ on a

$$\mathrm{Err}(\Delta t) \longrightarrow 0, \qquad \text{lorsque} \quad \Delta t \longrightarrow 0. \tag{3.7}$$

Définition 3.14 Méthode numérique consistante (†)

On dit qu'une méthode numérique est ***consistante*** si pour tout champ de vitesse $\mathcal{F}$ globalement lipschitzien pour sa variable d'espace, pour toute donnée initiale $X \in \Omega$ et pour tout temps final $T > 0$ on a

$$\eta(\Delta t) \longrightarrow 0, \qquad \text{lorsque} \quad \Delta t \longrightarrow 0. \tag{3.8}$$

Remarque : Pour simplifier la présentation, on va ignorer les erreurs d'arrondies faites par la machine lors de l'exécution de l'algorithme. Si on souhaite les incorporer, l'analyse qui va suivre reste valide avec des modifications mineures.

L'idée générale du théorème de convergence numérique consiste à montrer que si l'erreur de consistance tend vers zéro lorsque le pas de temps Δt tend vers zéro alors l'erreur totale aussi. Pour qu'une telle propriété soit possible, il est impératif d'avoir un schéma stable pour contrôler la propagation des erreurs par l'algorithme. Dans le cadre de ce cours, nous allons étudier la précision des différents schémas numériques à l'aide du théorème fondamental suivant :

Théorème 3.1 Le théorème de convergence numérique (‡)

Si une méthode numérique est consistante et localement stable, alors elle est convergente lorsque $\Delta t \to 0$ et

$$\mathrm{Err}(\Delta t) \leq C\,\eta(\Delta t), \tag{3.9}$$

où $C = C[T, \mathcal{F}]$ est la constante de stabilité du schéma.

La plus ancienne démonstration connue de résultats de convergence numérique est celle de Louis-Augustin Cauchy, retrouvée dans ses notes de cours de l'École Polytechnique datées de l'année 1824. Cette démonstration portait uniquement sur la convergence pour Euler Explicite. Cette version plus précise, plus simple et plus générale a été le fruit du travail de différents mathématiciens au cours du XIX^e siècle.

Démonstration. Soit $\mathcal{F} : [0, +\infty) \times \mathbb{R}^d \to \mathbb{R}^d$ un champ de vitesse, soit $X \in \mathbb{R}^d$ une donnée initiale et $n \in \mathbb{N} \mapsto \mathbf{S}^n[\Delta t, \mathcal{F}]$ un flot numérique localement stable et consistant. On étudie la différence entre solution exacte et numérique à l'aide de la méthode dite des "*sommes télescopiques*" :

$$\begin{aligned} \mathscr{S}^{t_n} X - \mathbf{S}^n X &= \sum_{k=1}^{n} \left(\mathbf{S}_k^n \mathscr{S}^{t_k} X - \mathbf{S}_{k-1}^n \mathscr{S}^{t_{k-1}} X \right) \\ &= \sum_{k=1}^{n} \left(\mathbf{S}_k^n \mathscr{S}^{t_k} X - \mathbf{S}_k^n \mathbf{S}_{k-1}^k \mathscr{S}^{t_{k-1}} X \right) \end{aligned} \tag{3.10}$$

où pour la deuxième égalité nous avons utilisé les relations de semi-groupe pour la composition

des flots discrets. Par l'inégalité triangulaire, on a :

$$\left|\mathscr{S}^{t_n}X - \mathbf{S}^n X\right| \;\leq\; \sum_{k=1}^{n} \left|\mathbf{S}_k^n \mathscr{S}^{t_k}X - \mathbf{S}_k^n \mathbf{S}_{k-1}^k \mathscr{S}^{t_{k-1}}X\right|. \tag{3.11}$$

La propriété de stabilité numérique (voir Définition 3.8) appliquée à la différence entre $\mathscr{S}^{t_k}X$ et $\mathbf{S}_{k-1}^k \mathscr{S}^{t_{k-1}}X$ nous donne

$$\left|\mathbf{S}_k^n \mathscr{S}^{t_k}X - \mathbf{S}_k^n \mathbf{S}_{k-1}^k \mathscr{S}^{t_{k-1}}X\right| \;\leq\; C\left|\mathscr{S}^{t_k}X - \mathbf{S}_{k-1}^k \mathscr{S}^{t_{k-1}}X\right|, \tag{3.12}$$

où C est la constante de stabilité de la méthode numérique (voir Définition 3.8). En utilisant les relations de semi-groupe pour la composition des flots continus, on remarque que le terme de droite de l'inégalité ci-dessus est égal à l'erreur de troncature locale :

$$\left|\mathscr{S}^{t_k}X - \mathbf{S}_{k-1}^k \mathscr{S}^{t_{k-1}}X\right| \;=\; \left|\mathscr{S}_{t_{k-1}}^{t_k} \mathscr{S}^{t_{k-1}}X - \mathbf{S}_{k-1}^k \mathscr{S}^{t_{k-1}}X\right| \;=\; \eta_k. \tag{3.13}$$

Par conséquent, l'équation (3.11) implique que

$$\left|\mathscr{S}^{t_n}X - \mathbf{S}^n X\right| \;\leq\; C\sum_{k=1}^{n}\eta_k \;\leq\; C\eta. \tag{3.14}$$

En prenant le supremum pour la variable $n \leq T/\Delta t$ à gauche de l'inégalité ci-dessus, on obtient (3.9). □

Remarque : A la lecture de la preuve on voit que la vitesse de convergence du schéma semble dépendre de la constante de stabilité du schéma et on constate empiriquement que c'est effectivement le cas ; c'est donc un enjeu d'analyse numérique de construire des schémas avec les meilleures propriétés de stabilité possibles.

Remarque 2 : On pourrait naturellement se demander si l'inégalité (3.9) est optimale, c'est-à-dire est-ce que nous avons

$$\mathrm{Err}(\Delta t) \;\simeq\; \eta(\Delta t), \qquad \text{lorsque} \quad \Delta t \;\longrightarrow\; 0. \tag{3.15}$$

Il n'est pas possible de conclure dans le cas général car il existe de rares cas particuliers pour lesquels les erreurs de troncatures locales se compensent lors de leur propagation par l'algorithme. Cependant, dans le cas (très) général on peut affirmer que l'on a la propriété (3.15) par une étude approfondie des termes d'erreur (études complétées et confirmées par les observations empiriques lors de l'exécution des algorithmes).

3.1.4 Étudier la convergence d'un schéma

Le théorème de convergence numérique est l'outil central pour nous permettre d'étudier la convergence des schémas, c'est-à-dire pour estimer si notre approximation numérique est suffisamment proche de la solution exacte. Ce théorème nous permet de nous ramener à étudier séparément la troncature locale sur un pas de temps et à l'étude de propriétés de stabilité numérique.

Troncature locale et développements limités

L'Étude de l'erreur de troncature locale se fait à l'aide des formules de développements limités de Taylor. En effet, si on écrit $X(t_{n+1})$ en fonction de $X(t_n)$ à l'aide d'un développement limité par rapport à Δt (vu comme un *petit paramètre*) on a :

$$X(t_{n+1}) \;=\; X(t_n) \;+\; \dot{X}(t_n)\,\Delta t \;+\; \mathcal{O}\big(\Delta t^2\big) \tag{3.16}$$

On utilise alors le fait que X est solution de l'équation $\dot{X}(t) = \mathcal{F}(t, X(t))$ pour écrire

$$X(t_{n+1}) \;=\; X(t_n) \;+\; \mathcal{F}\big(t_n, X(t_n)\big)\,\Delta t \;+\; \mathcal{O}\big(\Delta t^2\big) \tag{3.17}$$

Ce développement limité nous permet de comparer $X(t_{n+1})$ par rapport à son approximation numérique sur un pas de temps (à savoir $\mathbf{S}_n^{n+1}X(t_n)$) en calculant $X(t_{n+1}) - \mathbf{S}_n^{n+1}X(t_n)$ et en regardant quels sont les termes du développement limité qui vont être annulés par l'expression de $\mathbf{S}_n^{n+1}X(t_n)$. Plus l'approximation est précise et plus on a de termes qui s'annulent dans le développement limité de $X(t_{n+1}) - \mathbf{S}_n^{n+1}X(t_n)$. Si on souhaite estimer la constante qui apparaît dans le $\mathcal{O}$, on utilise l'inégalité de Taylor-Lagrange.

Stabilité numérique et propagation d'erreurs

Pour établir des résultats de stabilité numérique au sens de la définition 3.8, il nous faut étudier la façon dont se propagent les erreurs lors de l'exécution de l'algorithme. Cela se fait à l'aide de la théorie de la stabilité (Chapitre 2) et de la théorie de la comparaison des solutions et du lemme de Grönwall (Chapitre 1). Plus précisément, à l'aide d'une version discrète de ces théories, on peut démontrer les résultats de stabilité numérique des principaux schémas numériques utilisés par les ingénieurs.

Dans le cadre de ce cours, nous ferons usage du résultat suivant :

THÉORÈME 3.2 Théorème de propagation des erreurs numériques (†)

On considère un schéma numérique donné par un flot discret $(\mathbf{S}^n)_{n\in\mathbb{N}}$ et un champ de vitesse $\mathcal{F}$ fixé. On suppose qu'il existe une constante $\mu \in \mathbb{R}$ et un pas de temps critique $\Delta t^\star$ tels que pour tout $\Delta t \leq \Delta t^\star$ on ait la propriété de régularité lipschitzienne suivante :

$$\forall\, X, Y \in \Omega, \qquad \big\|\mathbf{S}^1 X - \mathbf{S}^1 Y\big\| \;\leq\; \big(1 + \mu\,\Delta t\big)\,\|X - Y\|, \tag{3.18}$$

pour une certaine norme $\|\_\|$ définie sur $\mathbb{R}^d$. Alors pour tout $T > 0$, le schéma numérique admet la propriété de stabilité suivante :

$$\sup_{\substack{n\in\mathbb{N}\\ n\Delta t\leq T}} \big|\mathbf{S}^n X - \mathbf{S}^n Y\big| \;\leq\; K^2\,|X - Y|\,\max\big\{1, e^{\mu T}\big\}, \tag{3.19}$$

où K est la constante de comparaison des normes : $\forall\, X \in \mathbb{R}^d, \quad \dfrac{1}{K}|X| \;\leq\; \|X\| \;\leq\; K|X|.$

En d'autres termes, le schéma est localement stable avec une constante de stabilité égale au maximum entre 1 et $e^{\mu T}$ multiplié par K^2. On constate que si $\mu \leq 0$ alors cette constante de stabilité est indépendante du temps final (stabilité numérique).

Démonstration. Par récurrence immédiate, nous avons :

$$\forall\, n \in \mathbb{N}, \qquad \big\|\mathbf{S}^n X - \mathbf{S}^n Y\big\| \;\leq\; \big(1 + \mu\,\Delta t\big)^n\,\|X - Y\|. \tag{3.20}$$

Dans le cas où $\mu \leq 0$ on obtient que $(1+\mu\,\Delta t)^n \leq 1$ et donc

$$\forall\, n \in \mathbb{N}, \qquad \big\|\mathbf{S}^n X - \mathbf{S}^n Y\big\| \;\leq\; \|X - Y\|, \tag{3.21}$$

On utilise à présent la comparaison de la norme $\|\_\|$ avec la norme euclidienne canonique :

$$\frac{1}{K}\big|\mathbf{S}^n X - \mathbf{S}^n Y\big| \;\leq\; \big\|\mathbf{S}^n X - \mathbf{S}^n Y\big\|, \qquad \text{et} \qquad \|X - Y\| \;\leq\; K|X - X|. \tag{3.22}$$

En combinant (3.21) et (3.22) on obtient

$$\forall\, n \in \mathbb{N}, \qquad \frac{1}{K}\big|\mathbf{S}^n X - \mathbf{S}^n Y\big| \;\leq\; K|X - Y|, \tag{3.23}$$

ce qui nous donne (3.19) puisque si $\mu \leq 0$ on a $\max\{1, e^{\mu T}\} = 1$.

Dans le cas où $\mu > 0$, on utilise le fait que l'on considère seulement des $n \in \mathbb{N}$ tels que $n\Delta t \leq T$ pour écrire à partir de (3.20) :

$$\sup_{\substack{n\in\mathbb{N}\\ n\Delta t\leq T}} \big\|\mathbf{S}^n X - \mathbf{S}^n Y\big\| \;\leq\; \big(1+\mu\,\Delta t\big)^{\frac{T}{\Delta t}}\|X - Y\|. \tag{3.24}$$

Vu que le logarithme est une fonction concave, on a $\ln(1+x) \leq x$ et donc

$$\big(1+\mu\,\Delta t\big)^{\frac{T}{\Delta t}} \;=\; \exp\left(\frac{T}{\Delta t}\ln\big(1+\mu\,\Delta t\big)\right) \;\leq\; \exp\big(\mu T\big). \tag{3.25}$$

En réinjectant cette inégalité dans (3.24) et en utilisant la comparaison de la norme $\|\_\|$ avec la norme euclidienne canonique :

$$\frac{1}{K}\sup_{\substack{n\in\mathbb{N}\\ n\Delta t\leq T}} \big|\mathbf{S}^n X - \mathbf{S}^n Y\big| \;\leq\; K\exp\big(\mu T\big)|X - Y|. \tag{3.26}$$

On obtient bien (3.19) puisque si $\mu > 0$ alors on a $\max\{1, e^{\mu T}\} = e^{\mu T}$. □

Remarque importante : Dans beaucoup de situations on utilise ce théorème directement avec la norme euclidienne canonique (et dans ce cas $K = 1$) mais il est parfois pertinent de choisir une norme plus judicieuse. On rappelle par exemple que si P est une matrice inversible alors $X \longmapsto |P^{-1}X|$ est une norme. Il peut s'avérer utile d'utiliser une telle norme avec P une matrice de changement de base ; par exemple un changement de base qui diagonaliserait la matrice A si on étudie l'équation linéaire $\dot{X} = AX$.

Notons que l'existence de cette constante K est toujours vérifiée en dimension finie (théorème d'équivalence des normes) mais fausse en dimension infinie.

Notion de A-stabilité

Bien que le théorème de propagation des erreurs numérique soit très performant pour démontrer qu'un schéma est numériquement stable, il est parfois difficile d'utilisation dans de nombreux cas pratiques. Dans beaucoup de situations non-linéaires, il se révèle inutilisable et une étude à l'aide de fonctions de Lyapunov est alors nécessaire (ce qui rend l'analyse très technique).

L'idée derrière la notion de A-stabilité consiste à dire que si on a de bonnes propriétés de stabilité dans le cas des dynamiques linéaires à coefficients constants, alors on peut raisonnablement espérer qu'il en est de même dans le cas général (par des arguments de linéarisation

similaires à ceux présentés au chapitre 2). On observe empiriquement que c'est très souvent le cas, et ceci est largement confirmé par des études théoriques, bien que des contre-exemples existent (ces situations rares s'étudient au cas-par-cas).

Définition 3.15 Méthode numérique A-stable

On considère une méthode numérique donnée par un flot discret $(\mathbf{S}^n)_{n\in\mathbb{N}}$ paramétré par un champ de vitesse $\mathcal{F}$ et un pas-de-temps Δt. On dit que cette méthode est ***A-stable*** ssi pour toute matrice A et pour tout pas de temps Δt, on a l'implication suivante :

$$\sup_{t\geq 0}\left|e^{tA}\right| < +\infty, \qquad \Longrightarrow \qquad \forall\, X\in\Omega,\quad \sup_{n\in\mathbb{N}}\left|\mathbf{S}^n[A,\Delta t]X\right| < +\infty, \tag{3.27}$$

où dans cette expression on a identifié la matrice A avec l'endomorphisme $X \mapsto AX$.

Le concept de A-stabilité coïncide avec celui de stabilité inconditionnelle sauf que l'on se restreint aux dynamiques linéaires à coefficients constants. Si l'implication ci-dessus n'est vraie que pour des pas de temps Δt assez petits, on parle de *A-stabilité sous condition*. Pour trouver le pas-de-temps en deçà duquel le schéma devient stable on étudie le domaine de A-stabilité :

Définition 3.16 Domaine de A-stabilité

Le ***domaine de A-stabilité*** d'une méthode numérique est défini par

$$\left\{A\in\mathcal{M}_d(\mathbb{R})\ :\ \forall\, X\in\Omega,\quad \sup_{n\in\mathbb{N}}\left|\mathbf{S}^n[A,1]X\right| < +\infty\right\}. \tag{3.28}$$

En pratique : on étudie la A-stabilité pour l'équation 1D linéaire suivante : $\dot{x} = \lambda x$ avec $\lambda \in \mathbb{C}$ pour un pas de temps $\Delta t = 1$. On en déduit le cas général en faisant une diagonalisation de matrice (quitte a gérer d'éventuels blocs de Jordan). Pour ce qui est du pas de temps, le comportement s'obtiennent généralement par une simple homothétie à partir du pas $\Delta t = 1$.

Les tracés des domaines de A-stabilité donnent de précieuses informations sur la stabilité de la méthode numérique. Plus ce domaine est grand et plus la méthode est stable. Si le domaine de A-stabilité d'une méthode englobe une large part de l'axe réel négatif alors cette méthode est adaptée à l'étude des équations fortement dissipatives. Si au contraire le domaine de A-stabilité englobe une large part de l'axe imaginaire, alors il s'agit d'une méthode adaptée aux équations conservatives (comme les équations hamiltoniennes par exemple).

A retenir : l'étude du domaine de A-stabilité se révèle largement suffisant pour s'assurer de la stabilité (conditionnelle ou inconditionnelle) d'une méthode numérique dans le cas général.

Autres notions de stabilité

Dans la continuité de cette notions de A-stabilité é, toute une théorie a été développée pour décrire la façon dont se propagent les erreurs au cours du temps dans les algorithmes numériques dans le cas linéaire et non-linéaire. De nombreuses autres notions de stabilité ont été introduites par les mathématiciens au fil des décennies (L-stabilité, B-stabilité, zéro-stabilité, etc...) pour tenter de décrire tout ce qui peut se passer lors de l'exécution de l'algorithme et construire les méthodes numériques les plus stables possibles (en fonction de chaque problème).

3.2 Discrétisations d'ordre plus élevé

3.2.1 Ordre de convergence d'un schéma

Une fois analysé la stabilité d'un schéma, on souhaite montrer qu'il est convergent en montrant que son erreur de consistance tend vers 0. Celle-ci s'étudie à l'aide de développements limités sur l'intervalle $[t_n, t_{n+1}]$ par rapport au petit paramètre $\Delta t > 0$. Cependant, tous les schémas ne convergent pas à la même vitesse lorsque $\Delta t \to 0^+$. Pour les classifier, on définit :

Définition 3.17 Ordre de convergence d'un schéma (†)

L'***ordre de convergence*** d'un schéma numérique est le plus grand entier $k \in \mathbb{N}^*$ tel que l'erreur totale est comparable avec $\Delta t^{\,k}$:

$$\mathrm{Err}(\Delta t) \lesssim \Delta t^{\,k}, \qquad \text{lorsque } \Delta t \to 0^+.$$

Les constantes multiplicatives qui apparaissent dans la définition du symbole $\simeq$ sont pas trop grandes en pratique (plus précisément, ça dépend de la régularité du champs de vitesse $\mathcal{F}$). Dans le cas où le théorème de convergence numérique s'applique, l'ordre de convergence est égale à l'ordre de consistance. Pour simplifier on peut interpréter la notion d'ordre comme suit :

- **Ordre 1 :** si je divise Δt par 10, je divise l'erreur par 10^1 (je gagne **1** décimale).
- **Ordre 2 :** si je divise Δt par 10, je divise l'erreur par 10^2 (je gagne **2** décimales).
- **Ordre 3 :** si je divise Δt par 10, je divise l'erreur par 10^3 (je gagne **3** décimales).

etc...

On peut à présent énoncer le théorème relatif à la convergence de la méthode de Euler explicite ou implicite :

Proposition 3.1 Ordre de convergence pour les schémas de Euler (‡)

On suppose que le champ de vitesse $\mathcal{F}$ est globalement lipschitz pour la variable d'espace. Dans ce cas :

(i) Le schéma de Euler Explicite est stable sous condition et son ordre vaut 1.

(ii) Le schéma de Euler Implicite est inconditionnellement stable et son ordre vaut 1.

La démonstration de ce théorème est proposée comme exercice à la fin de ce chapitre (dans un cadre simplifié).

3.2.2 Crank-Nicolson et Point-Milieu implicite

L'objectif de cette section est de reprendre les méthodes de Euler pour les améliorer et ainsi obtenir deux méthodes d'ordre 2 (la méthode de Crank-Nicolson et celle du point-milieu) qui vont avoir de bien meilleures propriétés de convergence.

Méthode de Crank-Nicolson

L'un des grands défauts de la méthode de Euler-Explicite est qu'elle sous-estime systématiquement la courbure de la solution exacte et donc tend significativement à s'en éloigner (d'où une convergence lente à l'ordre 1). La méthode de Euler Implicite, au contraire, va sur-estimer la courbure et s'éloigner de la courbe exacte mais dans la direction contraire (voir Figure 3.1 pour

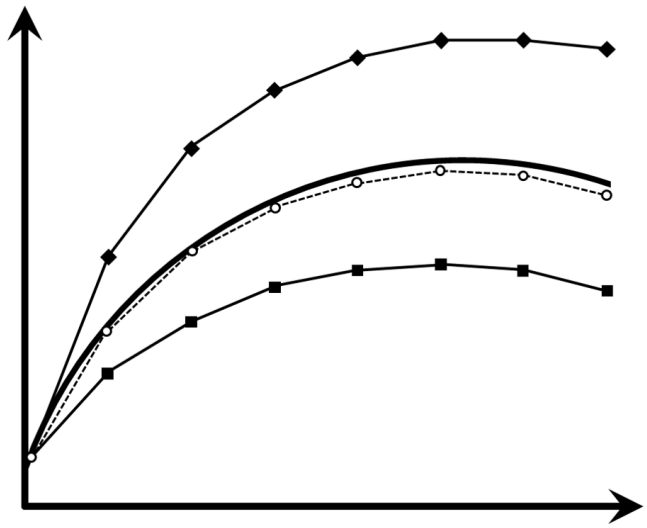


FIGURE 3.1 – Comparaison du schéma de Euler-Explicite (sur les losanges), Euler Implicite (sur les carrés) et Crank-Nicolson (en pointillés) avec la solution exacte (en gras).

une illustration). Si on fait le développement limité à l'ordre 2 pour estimer l'erreur de troncature locale, on obtient effectivement une erreur liée à un terme de courbure de la trajectoire $t \mapsto X(t)$:

- ***Euler explicite :***

$$X_1 - \mathscr{S}^{t_1} X_0 = X_0 + \Delta t \mathcal{F}(t_0, X_0) - \left(X_0 + \int_0^{\Delta t} \mathcal{F}(s, X(s))\, \mathrm{d}s \right) = -\frac{\Delta t^2}{2} \ddot{X}(t_0) + \mathcal{O}(\Delta t^3)$$

- ***Euler implicite :***

$$X_1 - \mathscr{S}^{t_1} X_0 = X_0 + \Delta t \mathcal{F}(t_1, X_1) - \left(X_0 + \int_0^{\Delta t} \mathcal{F}(s, X(s))\, \mathrm{d}s \right) = \frac{\Delta t^2}{2} \ddot{X}(t_0) + \mathcal{O}(\Delta t^3)$$

On constate que le terme d'erreur d'ordre 2 est le même au signe près. Il est naturel de vouloir faire disparaître ce terme d'ordre 2 en faisant la moyenne de deux schémas :

DÉFINITION 3.18 Schéma de Crank-Nicolson (‡)

On considère un problème de Cauchy (3.2), un pas de temps $\Delta t > 0$ et un nombre d'itérations N. On appelle méthode (ou schéma) de ***Crank-Nicolson*** l'algorithme itératif suivant :

$$X_{n+1} := X_n + \frac{\Delta t}{2} \Big(\mathcal{F}\big(t_n, X_n\big) + \mathcal{F}\big(t_{n+1}, X_{n+1}\big) \Big).$$

PROPOSITION 3.2 Ordre de convergence pour le schéma de Crank-Nicolson (‡)

On suppose que le champ de vitesse $\mathcal{F}$ et son gradient $\nabla \mathcal{F}$ sont globalement lipschitz pour la variable d'espace. Dans ce cas :

Le schéma de Crank-Nicolson est inconditionnellement stable et son ordre vaut 2.

L'hypothèse de régularité supplémentaire sur $\mathcal{F}$ s'explique par le fait que, dans la démonstration, on effectue un développement limité à un ordre plus grand par rapport à Euler.

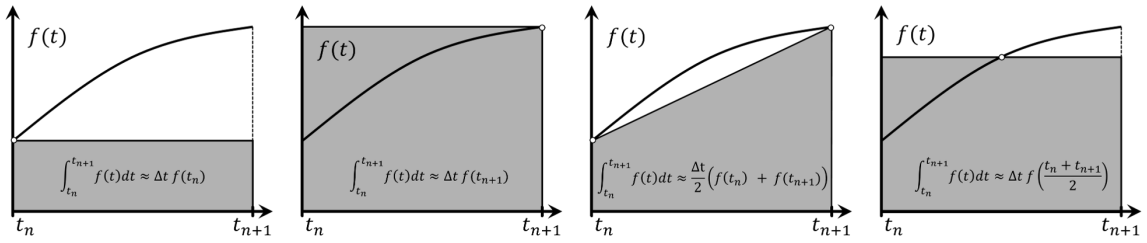


FIGURE 3.2 – Les 4 approximations d'intégrales les plus importantes (de gauche à droite) : Approximation de Riemann à gauche, Approximation de Riemann à droite, Méthode des trapèzes (ou interpolation linéaire), Méthode du point-milieu.

Méthode du point-milieu implicite

Pour appréhender la seconde méthode présentée dans ce paragraphe, il nous faut faire le lien entre la discrétisation des EDO et l'approximation numérique des intégrales (voir Figure 3.2). En effet, on réécrit l'EDO à l'aide de la formule de Duhamel sous la forme d'un calcul intégral :

$$X(t_{n+1}) = X(t_n) + \int_{t_n}^{t_{n+1}} \mathcal{F}\big(t, X(t)\big)\,\mathrm{d}t.$$

On peut donc obtenir des méthodes pour la résolution numérique d'une l'EDO simplement en utilisant une méthode d'approximation numérique de l'intégrale de Duhamel :

- L'intégration de *Riemann à gauche* donne le schéma de *Euler-Explicite.*
- L'intégration de *Riemann à droite* donne le schéma de *Euler-Implicite.*
- L'intégration par *Méthode des trapèzes* donne le schéma de *Crank-Nicolson.*
- L'intégration par *Méthode du point-milieu* va donner le schéma du *point-milieu implicite* :

$$X_{n+1} \;:=\; X_n + \Delta t\, \mathcal{F}\left(t_n + \frac{\Delta t}{2}, X_{n+\frac{1}{2}}\right). \tag{3.29}$$

Malheureusement en l'état actuel cette équation n'est pas utilisable car le terme $X_{n+\frac{1}{2}}$ n'est pas bien défini. Comme la valeur de X au temps $t = t_n + \Delta t/2$ n'est par connue, il faut remplacer $X_{n+\frac{1}{2}}$ par une approximation qui fasse intervenir X_n et X_{n+1}. Le plus naturel est de prendre la moyenne entre les deux :

DÉFINITION 3.19 Schéma du point-milieu implicite

On considère un problème de Cauchy (3.2), un pas de temps $\Delta t > 0$ et un nombre d'itérations N. On appelle méthode (ou schéma) du ***point-milieu implicite*** l'algorithme itératif suivant :

$$X_{n+1} \;:=\; X_n + \Delta t\, \mathcal{F}\left(t_n + \frac{\Delta t}{2}, \frac{X_n + X_{n+1}}{2}\right).$$

PROPOSITION 3.3 Ordre de convergence pour le schéma du point milieu implicite

On suppose que le champ de vitesse $\mathcal{F}$ et son gradient $\nabla\mathcal{F}$ sont globalement lipschitz pour la variable d'espace. Dans ce cas :

Le schéma du point-milieu implicite est inconditionnellement stable et son ordre vaut 2.

Dans la pratique : Ce schéma est peu utilisé car il ne rentre pas dans la théorie des schémas de Runge-Kutta (présentée plus loin) et on constante empiriquement qu'il est souvent légèrement

moins précis que le schéma de Crank-Nicolson pour un temps de calcul souvent équivalent... Ce qui est intéressant en revanche réside dans la formule (3.29) : si on a une meilleure approximation de $X_{n+\frac{1}{2}}$ alors on obtient un schéma avec de bonnes propriétés.

3.2.3 Point-Milieu explicite et Méthode de Heun

Point-Milieu explicite

Une solution pour estimer $X_{n+\frac{1}{2}}$ consiste à faire une prédiction $\widetilde{X}_{n+\frac{1}{2}}$ de sa valeur possible à l'aide d'un schéma numérique simple et de préférence explicite. C'est le principe des schémas par ***prédiction-correction*** (on parle aussi d'une *explicitation approchée*). Dans notre cas, on va expliciter $X_{n+\frac{1}{2}}$ en remplaçant ce terme par l'approximation de *Euler explicite* :

Définition 3.20 Schéma du point-milieu explicite

On considère un problème de Cauchy (3.2), un pas de temps $\Delta t > 0$ et un nombre d'itérations N. On appelle méthode (ou schéma) du ***point-milieu explicite*** l'algorithme itératif suivant :

$$\begin{aligned} \widetilde{X}_{n+\frac{1}{2}} &:= X_n + \frac{\Delta t}{2}\mathcal{F}\big(t_n, X_n\big), \\ X_{n+1} &:= X_n + \Delta t\, \mathcal{F}\left(t_n + \frac{\Delta t}{2}, \widetilde{X}_{n+\frac{1}{2}}\right). \end{aligned} \tag{3.30}$$

Comme il s'agit d'un schéma explicite, on retrouve sans surprise des effets d'instabilité lorsque Δt n'est pas assez petit :

Proposition 3.4 Ordre de convergence pour le schéma du point milieu explicite

On suppose que le champs de vitesse $\mathcal{F}$ et son gradient $\nabla\mathcal{F}$ sont globalement lipschitz pour la variable d'espace. Dans ce cas :

Le schéma du point-milieu explicite est stable sous condition et son ordre vaut 2.

Ce schéma est plus rapide que Crank-Nicolson tout en ayant une précision comparable, mais il est moins stable.

Méthode de Heun

La méthode de Heun réside dans la même démarche de *prédiction-correction* mais appliquée à la méthode Crank-Nicolson pour faire une *explicitation approchée.* On va approcher le X_{n+1} apparaissant dans Crank-Nicolson en faisant un pas de Euler Explicite :

$$\begin{aligned} \widetilde{X}_{n+1} &:= X_n + \Delta t\, \mathcal{F}\big(t_n, X_n\big), \\ X_{n+1} &:= X_n + \Delta t\, \frac{1}{2}\Big(\mathcal{F}\big(t_n, X_n\big) + \mathcal{F}\big(t_{n+1}, \widetilde{X}_{n+1}\big)\Big). \end{aligned} \tag{3.31}$$

Afin de diminuer le nombre d'appel à la fonction $\mathcal{F}$ (ce qui peut être coûteux en temps de calcul) on peut reformuler cet algorithme de la manière suivante :

Définition 3.21 Méthode de Heun (‡)

On considère un problème de Cauchy (3.2), un pas de temps $\Delta t > 0$ et un nombre d'itérations N. On appelle méthode de ***Heun*** l'algorithme itératif suivant :

$$\begin{aligned} \widetilde{X}_{n+1} &:= X_n + \Delta t\, \mathcal{F}\big(t_n, X_n\big), \\ \widehat{X}_{n+1} &:= X_n + \Delta t\, \mathcal{F}\big(t_{n+1}, \widetilde{X}_{n+1}\big), \\ X_{n+1} &:= \frac{1}{2}\big(\widetilde{X}_{n+1} + \widehat{X}_{n+1}\big). \end{aligned} \tag{3.32}$$

La première étape s'appelle l'***étape de prédiction*** et la seconde l'***étape de correction.*** La deuxième étape vient corriger la première qui a fait une erreur à cause du terme de courbure. Comme la plupart des schémas explicites, la méthode de Heun a un domaine d'instabilité lorsque Δt est trop grand. En pratique le domaine de stabilité s'avère être assez grand, le temps de calcul plus faible que Crank-Nicolson (surtout si la fonction $\mathcal{F}$ est couteuse à calculer) et la précision est assez comparable.

Proposition 3.5 Ordre de convergence pour la méthode de Heun (‡)

On suppose que le champs de vitesse $\mathcal{F}$ et son gradient $\nabla\mathcal{F}$ sont globalement lipschitz pour la variable d'espace. Dans ce cas :

La méthode de Heun est stable sous condition et son ordre vaut 2.

3.2.4 Méthodes de Runge-Kutta

Principe général

Le principe des méthodes de Runge-Kutta consiste à formuler dans un cadre général le principe de *prédiction-correction* présenté précédemment. Nous allons donc introduire des points intermédiaires de calculs en temps $t_{n,i}$ ainsi que des valeurs intermédiaires de prédictions $X_{n,i}$ avec $i = 1, \dots, q$. Ces temps $t_{n,i}$ sont caractérisés par un nombre $c_i \in [0, 1]$ qui est une donnée de la méthode et on écrit

$$t_{n,i} := t_n + c_i \Delta t.$$

Pour chaque point intermédiaire $(t_{n,i}, X_{n,i})$ on évalue le champ de vitesse en utilisant l'équation

$$p_{n,i} = \mathcal{F}\big(t_{n,i}, X_{n,i}\big).$$

Si on considère une solution exacte $t \mapsto X(t)$, on a

$$X(t_{n,i}) = X(t_n) + \Delta t \int_0^{c_i} \mathcal{F}\Big(t_n + u\Delta t,\, X(t_n + u\Delta t)\Big)\, du,$$

On fait ici une approximation de cette intégrale à l'aide de méthodes d'approximation : la valeur de la fonction intégrée est connue seulement aux points c_i. On fait cette quadrature pour $g(u) := \mathcal{F}(t_n + u\Delta t, X(t_n + u\Delta t))$:

$$\int_0^{c_i} g(u)\, \mathrm{d}u \;\approx\; \sum_{k=1}^{i-1} a_{ik}\, g(c_k), \;=\; \sum_{k=1}^{i-1} a_{ik}\, p_{n,k}$$

La valeurs des coefficients a_{ik} va dépendre du choix de la quadrature retenue. Pour avoir des méthodes de discrétisation des EDO d'ordre élevé, il faut choisir des quadratures d'intégrales

d'ordre élevé. On effectue également une quadrature pour le calcul de toute l'intégrale :

$$\int_0^1 g(u)\,\mathrm{d}u \;\approx\; \sum_{k=1}^{q} b_k g(c_k) \;=\; \sum_{k=1}^{q} b_k\, p_{n,k}.$$

Définition 3.22 Méthodes de Runge-Kutta (†)

On considère un problème de Cauchy (3.2), un pas de temps $\Delta t > 0$ et un nombre d'itérations N. On appelle méthodes de ***Runge-Kutta*** à q étapes la famille d'algorithmes itératifs suivante :

Pour i allant de 1 à q :

$$t_{n,i} \;:=\; t_n + c_i \Delta t, \qquad X_{n,i} \;:=\; X_n + \Delta t \sum_{k=1}^{q} a_{ik}\, p_{n,k}, \qquad p_{n,i} \;:=\; \mathcal{F}\big(t_{n,i}, X_{n,i}\big).$$

Ensuite on calcule : $\quad X_{n+1} = X_n + \Delta t \sum_{k=1}^{q} b_k\, p_{n,k}.$

(3.33)

Afin d'y voir plus clair, les différents coefficients c_i, a_{ik} et b_k sont rangés dans une matrice appelé ***tableau de Butcher :***

$$\begin{array}{c|c} C & A \\ \hline & B \end{array} \;=\; \begin{array}{c|ccccc} c_1 & a_{1,1} & a_{1,2} & & & \\ c_2 & a_{2,1} & a_{2,2} & \ddots & & \\ c_3 & a_{3,1} & a_{3,2} & \ddots & & \\ \vdots & \vdots & & \ddots & \ddots & \\ c_q & a_{q,1} & a_{q,2} & \dots & a_{q,q-1} & a_{q,q} \\ \hline & b_1 & b_2 & \dots & b_{q-1} & b_q \end{array} \tag{3.34}$$

Proposition 3.6 Méthode de Runge-Kutta explicite

Une méthode de Runge-Kutta est explicite ssi la matrice A du tableau de Butcher associé est triangulaire inférieure stricte.

Par convention les cases vides d'un tableau de Butcher correspondent à la valeur 0. Dans le cadre de ce cours nous allons travailler avec des méthodes de Runge-Kutta expicite puisque le but de ces méthodes de prédicteur-correcteur est de modifier des méthodes implicites pour les rendre explicites avec une précision comparable. Cependant, il existe de nombreuses méthodes de Runge-Kutta implicites avec de très bonnes propriétés de stabilité.

Théorème 3.3 Convergence pour Runge-Kutta (†)

Les méthodes de Runge-Kutta explicites sont stables sous condition et convergentes ssi

$$\sum_{k=1}^{q} b_k \;=\; 1.$$

Remarque 1 : Les méthodes de Euler Explicite et Implicite rentrent dans la (petite) famille des méthodes de Runge-Kutta à 1 étape. Leur tableau de Butcher respectif est

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array} \qquad \text{et} \qquad \begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}$$

Remarque 2 : Le plus grand ordre de convergence possible pour une méthode de Runge-Kutta explicite à q étapes est q lorsque les coefficients sont bien ajustés (à l'aide d'une bonne méthode de quadrature d'intégrale). Dans le cas des méthodes de Runge-Kutta implicites à q étapes, il existe des méthode d'ordre plus élevé que q.

Runge-Kutta 2

Les méthodes de Runge-Kutta 2 explicites d'ordre 2 sont entièrement paramétrées par un unique paramètre $\alpha \in [0, 1]$ et le tableau de Butcher correspondant s'écrit

$$\begin{array}{c|cc} 0 & & \\ \alpha & \alpha & 0 \\ \hline & \frac{2\alpha-1}{2\alpha} & \frac{1}{2\alpha} \end{array}$$

PROPOSITION 3.7 Méthode de Runge-Kutta 2 explicite **(†)**

Les méthodes de Runge-Kutta 2 associées à ce tableau de Butcher sont les seules méthodes de Runge-Kutta explicites à 2 étapes dont l'ordre de consistance vaut 2 (si $\mathcal{F}$ est suffisament régulière).

La démonstration de cette proposition est proposée au guise d'exercice (voir la section consacrée aux exercices) et elle se fait à l'aide d'un développement limité sur le terme d'erreur de troncature locale.

- Lorsque $\alpha = 1/2$, on retrouve la *méthode du point-milieu explicite*.
- Lorsque $\alpha = 1$ on retrouve la *méthode de Heun* (pour cette raison, la méthode de Heun est souvent simplement appelée *Runge-Kutta 2* sans plus de précision).
- Lorsque $\alpha = 2/3$ la méthode obtenue s'appelle la *méthode de Ralston*. C'est une méthode intermédiaire entre les deux précédentes et se révèle souvent légèrement plus précise car elle minimise (pour le paramètre α) l'erreur de troncature locale moyenne.

Si on regarde les méthodes de Runge-Kutta 2 implicites, on retrouve la méthode de *Crank-Nicolson* avec le tableau de Butcher suivant :

$$\begin{array}{c|cc} 0 & & \\ 1 & 1/2 & 1/2 \\ \hline & 1/2 & 1/2 \end{array}$$

Dans le cas implicite, on peut construire des méthodes de Runge-Kutta à 2 étapes mais d'ordre plus élevé que 2. C'est le cas par exemple de la méthode de Crouziex 2 (qui est d'ordre 3) :

$$\begin{array}{c|cc} \frac{1}{2}+\frac{\sqrt{3}}{6} & \frac{1}{2}+\frac{\sqrt{3}}{6} & 0 \\ \frac{1}{2}-\frac{\sqrt{3}}{6} & -\frac{\sqrt{3}}{3} & \frac{1}{2}+\frac{\sqrt{3}}{6} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

Runge-Kutta 3

Dans le cas des méthodes de Runge-Kutta explicites à 3 étapes, celles-ci s'écrivent à l'aide d'un tableau de Butcher qui prend la forme générale suivante ($\alpha, \beta, \gamma \in \mathbb{R}$) :

$$\begin{array}{c|ccc} 0 & & & \\ \alpha & \alpha & & \\ \beta+\gamma & \beta & \gamma & \\ \hline & b_1 & b_2 & b_3 \end{array}$$

Comme précédemment, on estime l'erreur de troncature locale avec des développements limités et on ajuste les paramètres α, β, γ pour obtenir des schémas d'ordre 3. On présente ici quelques tableaux de Butcher classiques pour l'explicite d'ordre 3 :

Heun d'ordre 3 :

0			
1/3	1/3		
2/3	0	2/3	
	1/4	0	3/4

Ralston d'ordre 3 :

0			
1/2	1/2		
3/4	0	3/4	
	2/9	1/3	4/9

RK3 fortement stable :

0			
1	1		
1/2	1/4	1/4	
	1/6	1/6	1/3

Kutta d'ordre 3 :

0			
1/2	1/2		
1	−1	2	
	1/6	2/3	1/6

Van der Houwen-Wray d'ordre 3 :

0				
8/15	8/15			
2/3	1/4	5/12		
	1/4	0	3/4	

Runge-Kutta 4

Il existe de nombreuses méthodes pour Runge-Kutta à 4 étapes. Cependant, lorsqu'on parle de ***Runge-Kutta 4***, on fait le plus souvent référence à une méthode très classique qui est celle associée au tableau de Butcher suivant :

0				
1/2	1/2			
1/2	0	1/2		
1	0	0	1	
	1/6	1/3	1/3	1/6

Une autre méthode de Runge-Kutta à 4 étapes appelée la *régle des 3 huitièmes* . Elle est légèrement plus stable et plus précise mais un peu plus lente. Elle est donnée par le tableau de Burcher suivant :

0				
1/3	1/3			
2/3	−1/3	1		
1	1	−1	1	
	1/8	3/8	3/8	1/8

3.2.5 Méthodes de Adams-Bashforth

Les méthodes de Runge-Kutta présentées précédemment appartiennent à la catégories des méthodes à 1 pas (et ce malgré la présence de pas de prédiction-correction) car ces méthodes n'exploitent que la position X_n (avec l'équation) pour déduire la position X_{n+1}. L'idée des ***méthodes à pas multiples*** consistent à exploiter l'information donnée par X_n, X_{n-1}, X_{n-2}, etc... En règle générale, les méthodes à pas multiples sont moins précises que les méthodes de Runge-Kutta (si on compare au même ordre) mais elles sont en revanche plus rapides en temps de calcul car on évalue moins souvent la fonction $\mathcal{F}$. Dans cette section, on va présenter le principe des pas multiples sur les méthodes les plus classiques à savoir Adams-Bashforth. Ces méthodes ne sont pas toujours très stables et nécessitent parfois des pas de temps Δt assez petits.

Pour construire les méthodes à pas multiples d'ordre k, on commence par écrire la formule de Duhamel :

$$X(t_{n+1}) \;=\; X(t_n) + \int_{t_n}^{t_{n+1}} \mathcal{F}\big(t, X(t)\big)\,\mathrm{d}t.$$

On va alors approcher la fonction sous l'intégrale à l'aide des valeurs de cette fonction aux instants t_n, t_{n-1},...,t_{n-k+1}. Ces valeurs sont connues de manière approchée respectivement par les valeurs $\mathcal{F}(t_n, X_n)$, $\mathcal{F}(t_{n-1}, X_{n-1})$,..., $\mathcal{F}(t_{n-k}, X_{n-k+1})$. La ***méthode d'Adams-Bashforth*** d'ordre k consiste à approcher l'intégrale étudiée en faisant une interpolation de Lagrange sur ces temps t_n, t_{n-1},...,t_{n-k+1}. On obtient formellement

$$\mathcal{F}(t, X(t)) \;\approx\; \sum_{j=0}^{k-1} F_{n-j}\, L_{n,j,k}(t),$$

où $F_j = \mathcal{F}(t_j, X_j)$ et où le polynôme de Lagrange unitaire est donné par la célèbre formule

$$L_{n,j,k}(t) \;:=\; \prod_{\substack{\ell=0 \\ j\neq j}}^{k-1} \frac{t - t_{n-\ell}}{t_{n-j} - t_{n-\ell}}.$$

Si on injecte cette approximation dans la formule de Duhamel on obtient :

Définition 3.23 Méthodes d'Adams-Bashforth (†)

On considère un problème de Cauchy (3.2), un pas de temps $\Delta t > 0$ et un nombre d'itérations N. On appelle méthodes d'***Adams-Bashforth*** à k étapes l'algorithme itératif explicite suivant :

$$\begin{aligned} X_{n+1} &= X_n + \Delta t \sum_{j=0}^{k-1} F_{n-j}\, b_{j,k}, \\ F_{n+1} &= \mathcal{F}\big(t_{n+1}, X_{n+1}\big), \end{aligned} \tag{3.35}$$

avec : $b_{j,k} := \dfrac{1}{\Delta t} \displaystyle\int_{t_n}^{t_{n+1}} L_{n,j,k}(t)\,\mathrm{d}t$ les coefficients d'Adams-Bashforth.

On peut en effet démontrer aisément que les coefficients d'Adams-Bashforth ainsi définis ne dépendent pas de n ni de Δt. Chacune de leur valeur est un nombre universel fixé. Les valeurs de ces nombres sont résumés dans le *tableau d'Adams-Bashforth* (ne pas confondre avec Butcher). Les coefficients s'écrivent tous sous la forme $b_{j,k} = \beta_{j,k}/a_k$ où les a_k et les $\beta_{j,k}$ sont les entiers :

k	$\beta_{0,k}$	$\beta_{1,k}$	$\beta_{2,k}$	$\beta_{3,k}$	$\beta_{4,k}$	a_k
1	1					1
2	3	-1				2
3	23	-16	5			12
4	55	-59	37	-9		24
5	1901	-2774	2616	-1274	251	720

On peut vérifier aisément que l'Algorithme d'Admas-Bashforth n'évalue la fonction $\mathcal{F}$ qu'une seule fois par pas de temps. C'est donc une bonne méthode, malgré sa faible stabilité, si le calcul de $\mathcal{F}$ est lent (par exemple parce qu'il provient d'un maillage de discrétisation spatiale très fine pour un problème d'EDP).

Théorème 3.4 Méthodes d'Adams-Bashforth

Pour $k \leq 5$, le schéma d'Adams-Bashforth à k pas est stable sous conditions et il est consistant d'ordre k (à condition que $\mathcal{F}$ soit suffisamment régulière).

3.2.6 Liste d'autres méthodes numériques

Voici une liste non-exhaustive d'autres méthodes numériques. Chacune d'entre-elle a différents avantages ou inconvénients, certaines étant spécifiques à des équations particulières (notamment adaptée au cas de la mécanique hamiltonienne) :

- Méthodes de Euler semi-implicite
- Euler exponentiel et intégrateurs exponentiels
- Méthode de Newmark-β
- Algorithme de Beeman
- Méthodes saute-moutons
- Méthodes à pas de temps adaptatif
- Méthodes de Runge-Kutta adaptatives
- Méthodes de Verlet et Störmer-Verlet
- Méthode de Adams-Moulton
- Méthode de Gauss-Legendre
- Backward differentiation formula
- Algorithmes de Yoshida
- etc...

3.3 Implémentation informatique

Dans cette section on présente les bases de l'implémentation informatique des méthodes d'analyse numérique afin de pouvoir s'en servir dans le cadre des séances de travaux pratiques consacrées à l'analyse numérique des EDO mais aussi pour la réalisation des différents projets.

3.3.1 Implémentation des méthodes explicites

Concernant l'implémentation informatique d'une méthode explicite, il n'y a pas de grandes difficultés : il suffit de faire une grande boucle `for`, d'exécuter l'algorithme présenté précédemment et de stocker le résultat dans un tableau. Concernant le nombre d'itérations, il est en général préférable de faire varier le pas de temps Δt et le temps final T et d'en déduire le nombre d'itérations N. On va illustrer ces différents codes avec l'exemple particulier de l'équation linéaire $\dot{x} = -3x + 2\sin(9t)$ intégrée sur $[0, 1]$.

Méthode de Euler explicite

Sur un code en langage `Python`, l'implémentation de la méthode de Euler-Explicite présentée à la définition 3.1 peut s'écrire de la façon suivante :

```
from pylab import *              # Importation de la bibliotheque maths

dt=0.001                         # Le pas de temps
T=1.                             # Le temps final
N=int(ceil(T/dt))                # Le nombre d'iteration (entier)
dt=T/N                           # Petite correction pour avoir T=N*dt
x_0=1.                           # Donnee initiale

# Le choix du champ de vitesse F :
def F(t,x):
    return -3*x+2*sin(9*t)
```

```
# Initialisation de l'algorithme :
t=[0.]                          # La liste des temps
x=[x_0]                         # La liste des positions au cours du temps

# La methode de Euler-Explicite proprement dite :
for k in range(N):
    x_new=x[k]+dt*F(t[k],x[k])  # Nouvelle position par Euler explicite
    x.append(x_new)             # stockage des resultats
    t.append(t[k]+dt)

# trace le graphe de la solution obtenue :
plot(t,x,'.k')
```

La façon de programmer ci-dessus est dite ***impérative.*** C'est la manière de programmer la plus naturelle mais ce n'est pas la plus pratique. Voici une simple réécriture de l'implémentation de la méthode de Euler-Explicite avec une programmation ***fonctionnelle :***

```
from pylab import *

# Le choix du champ de vitesse F :
def F(t,x):
    return -3*x+2*sin(9*t)

# La methode de Euler-Explicite dans une fonction :
def EulerExplicite(F,dt=0.01,T=1.,x_0=0.):
    N=int(ceil(T/dt))
    dt=T/N
    t=[0.]
    x=[x_0]
    for k in range(N):
        x.append(x[k]+dt*F(t[k],x[k]))
        t.append(t[k]+dt)
    return t,x

# trace le graphe de la solution obtenue :
t,x=EulerExplicite(F,0.001,1.,1.)
plot(t,x,'.k')
```

La bibliothèque `Pylab` est très facile d'utilisation pour tracer des courbes simples. Si on souhaite faire des tracés plus complexes, la bibliothèque `MatplotLib` dispose de plus d'options.

Remarque : Il est également possible de coder Euler-Explicite en utilisant une fonction auxiliaire récursive (une fonction qui s'appelle elle-même) au lieu d'une boucle `for`. Le temps de calcul est similaire mais la programmation de l'algorithme est plus hardue et source d'erreurs.

Méthode de Heun

Concernant la méthode Heun (définition 3.21), il n'y a pas de grands changements. Il faut simplement rajouter le calcul du prédicteur `p` et du correcteur `c` à chaque itération :

```
from pylab import *

# Le choix du champ de vitesse F :
def F(t,x):
    return -3*x+2*sin(9*t)

# La methode de Heun dans une fonction :
def Heun(F,dt=0.01,T=1.,x_0=0.):
    N=int(ceil(T/dt))
    dt=T/N
    t=[0.]
    x=[x_0]
    for k in range(N):
        p=x[k]+dt*F(t[k],x[k])   #Calcul du predicteur
        c=x[k]+dt*F(t[k]+dt,p)   #Calcul du correcteur
        x.append(0.5*(p+c))      #Le pas avec correction
        t.append(t[k]+dt)
    return t,x

# trace le graphe de la solution obtenue :
t,x=Heun(F,0.001,1.,1.)
plot(t,x,'.k')
```

Méthodes de Runge-Kutta

Les méthodes de Runge-Kutta d'ordre plus élevé s'implémentent de la même façon que la méthode de Heun. La seule différence est le nombre de correcteurs qui augmente avec l'ordre de la méthode. Par exemple, la méthode de Runge-Kutta 4 classique peut s'écrire de la manière suivante :

```
from pylab import *

# Le choix du champ de vitesse F :
def F(t,x):
    return -3*x+2*sin(9*t)

# La methode de Runge-Kutta 4:
def RK4(F,dt=0.01,T=1.,x_0=0.):
    N=int(ceil(T/dt))
    dt=T/N
    t=[0.]
    x=[x_0]
    for k in range(N):
        k1 = F(t[k], x[k])
        k2 = F(t[k]+dt/2, x[k]+dt*k1/2)
        k3 = F(t[k]+dt/2, x[k]+dt*k2/2)
        k4 = F(t[k]+dt, x[k]+dt*k3)
        x.append(x[k] + dt*(k1+2*k2+2*k3+k4)/6)
        t.append(t[k]+dt)
    return t,x

# trace le graphe de la solution obtenue :
t,x=RK4(F,0.001,1.,1.)
plot(t,x,'.k')
```

Méthode de Adams-Bashforth 2

Pour les méthodes à pas multiples comme Adams-Bashforth , il y a une subtilité concernant l'initialisation. En effet, comme on utilise les valeurs de X_n, X_{n-1}, X_{n-2}, etc... pour calculer X_{n+1}, celles-ci ne sont pas bien définies pour les petites valeurs de n. Il faut donc faire les premiers pas avec une méthode explicite à un pas qui a le même ***ordre de troncature locale*** ou meilleur. Pour Adams-Bashforth 2, nous avons fait le premier pas avec la méthode de Heun :

```
from pylab import *

# Le choix du champ de vitesse F :
def F(t,x):
    return -3*x+2*sin(9*t)

# La methode de Heun dans une fonction :
def Heun(F,dt=0.01,T=1.,x_0=0.):
    N=int(ceil(T/dt))
    dt=T/N
    t=[0.]
    x=[x_0]
    for k in range(N):
        p=x[k]+dt*F(t[k],x[k])   #Calcul du predicteur
        c=x[k]+dt*F(t[k]+dt,p)   #Calcul du correcteur
        x.append(0.5*(p+c))      #Le pas avec correction
        t.append(t[k]+dt)
    return t,x

# La methode de Adams-Bashforth 2 dans une autre fonction :
def AdamsBashforth2(F,dt=0.01,T=1., x_0=0.):
    N=int(ceil(T/dt))
    dt=T/N
    t,x=Heun(F,dt,dt,x_0) # le premier pas de l'algorithme se fait avec Heun
    for k in range(1,N):
        x.append(x[k]+dt*(3.*F(t[k],x[k])/2.-F(t[k-1],x[k-1])/2.))
        t.append(t[k]+dt)
    return t,x

# trace le graphe de la solution obtenue avec Adams-Bashforth
t,x=AdamsBashforth2(F,0.001,1.,x_0)
plot(t,x,'.k')
```

3.3.2 Implémentation des méthodes implicites

Cas explicitement inversible

Pour les méthodes implicites il y a une difficulté liée au fait que pour obtenir X_{n+1} à partir de X_n, il y a une fonction à inverser. Dans le cas où cette inverse peut s'exprimer explicitement et de manière exacte à l'aide de manipulations algébriques, il faut privilégier l'expression exacte. Cette situation se présente notoirement dans le cas linéaire $\dot{X} = A(t)X$ avec A une matrice qui peut dépendre du temps. Euler implicite dans ce cas va s'écrire :

$$\big(I_d - \Delta t A(t_{n+1})\big) X_{n+1} = X_n.$$

Si on note $\mu \geq 0$ la plus grande valeur des modules pour chaque valeur propre de $A(t_{n+1})$, on sait que cette matrice est inversible dès que $\Delta t < \mu$ et dans ce cas un simple pivot de Gauss (ou d'autres algorithmes plus rapides) permet d'avoir une expression explicite de l'inverse. Si la matrice est constante, il suffit de faire le pivot de Gauss une seule fois avant la simulation et ainsi gagner beaucoup de temps de calcul.

Si on fait les mêmes manipulations sur la méthode de Crank-Nicolson, on arrive à l'équation (inversible dès que $\Delta t < \mu/2$) suivante :

$$\left(I_d - \frac{\Delta t}{2} A(t_{n+1})\right) X_{n+1} = \left(I_d + \frac{\Delta t}{2} A(t_n)\right) X_n.$$

Méthode de Euler Implicite

Dans le cas où on ne peut pas inverser la partie implicite de manière algébrique simple et peu coûteuse, on résout le problème de façon approchée à l'aide d'un algorithme de recherche de racine, par exemple la méthode de Newton (ou une de ses généralisations comme la méthode de la sécante). Dans le cas des fonctions de plusieurs variables, la méthode de Newton pour trouver une racine d'une fonction $\mathcal{G}$ s'écrit formellement :

$$X_{k+1} := X_k - \nabla \mathcal{G}(X_k)^{-T} \mathcal{G}(X_k). \tag{3.36}$$

Pour initialiser l'algorithme de recherche de racine, on fait un pas de Euler Explicite. Dans le cas de la méthode de Euler Implicite on obtient le code suivant (en utilisant les fonctions anonymes `lambda`) :

```
from pylab import *

# Le choix du champ de vitesse F :
def F(t,x):
    return -3*x+2*sin(9*t)

# Le calcul de sa derivee (pour Newton) :
def dF(t,x):
    return -3

# l'algorithme de Newton prend en argument une fonction f, sa derivee df,
# une donnee initiale a et une erreur maximale e
def Newton(f,df,a,e):
    delta = 1
    while delta > e:
        x = -f(a)/df(a) + a
        delta = abs(x - a)
        a = x
    return x

# Euler Implicite avec la methode de Newton
def EulerImplicite(F,dF,dt=0.01,T=1.,x_0=0.):
    N=int(ceil(T/dt))
    dt=T/N
    t=[0.]
    x=[x_0]
    for k in range(N):
        f = lambda y : y-dt*F(t[k]+dt,y)-x[k]    # la fonction a inverser
        df = lambda z : 1-dt*dF(t[k]+dt,z)       # sa derivee
        x_init = x[k]+dt*F(t[k],x[k])            # initialise Newton
        x.append(Newton(f,df,x_init,1.e-12))     # un pas de Euler implicite
        t.append(t[k]+dt)
    return t,x

# trace le graphe de la solution obtenue :
t,x=EulerImplicite(F,dF,0.001,1.,1.)
plot(t,x,'.k')
```

Méthode de Crank-Nicolson

L'implémentation de la méthode de Crank-Nicolson est similaire à celle de Euler-Implicite. Seule la fonction à inverser est différente

```
from pylab import *

# Le choix du champ de vitesse F :
def F(t,x):
    return -3*x+2*sin(9*t)

# Le calcul de sa derivee (pour Newton) :
def dF(t,x):
    return -3

# l'algorithme de Newton prend en argument une fonction f, sa derivee df,
# une donnee initiale a et une erreur maximale e
def Newton(f,df,a,e):
    delta = 1
    while delta > e:
        x = -f(a)/df(a) + a
        delta = abs(x - a)
        a = x
    return x

# Crank-Nicolson avec la methode de Newton
def CrankNicolson(F,dF,dt=0.01,T=1.,x_0=0.):
    N=int(ceil(T/dt))
    dt=T/N
    t=[0.]
    x=[x_0]
    for k in range(N):
        f = lambda y : y-dt*F(t[k]+dt,y)/2-x[k]-dt*F(t[k],x[k])/2
        df = lambda z : 1-dt*dF(t[k]+dt,z)/2
        x_init = x[k]+dt*F(t[k],x[k])
        x.append(Newton(f,df,x_init,1.e-12))
        t.append(t[k]+dt)
    return t,x

# trace le graphe de la solution obtenue :
t,x=CrankNicolson(F,dF,0.001,1.,1.)
plot(t,x,'.k')
```

3.3.3 Calculer l'ordre empirique d'un schéma

Une fois que nous avons intégré numériquement une équation différentielle sur un intervalle de temps $[0, T]$ on doit étudier l'erreur empirique réalisée par le schéma. On pourra alors calculer l'ordre de convergence empirique (pour le comparer à l'ordre théorique). A toutes fins utiles, on rappelle que la définition de l'erreur est la différence entre la solution exacte et la solution approchée :

$$\mathrm{Err}(\Delta t) \;:=\; \max_{\substack{n\in\mathbb{N}\\ n\Delta t\leq T}} \Big|\mathbf{S}^n[\Delta t]X - \mathscr{S}^{n\Delta t}X\Big|. \tag{3.37}$$

Si on divise le pas de temps par 10, on s'attend à ce que l'erreur (sur le même intervalle de temps) soit divisée par 10^k ssi le schéma est d'ordre k. Autrement dit, formellement :

$$\mathrm{Err}(\Delta t) \;\approx\; 10^k\,\mathrm{Err}\left(\frac{\Delta t}{10}\right). \tag{3.38}$$

Il est donc naturel de définir l'ordre empirique comme suit :

DÉFINITION 3.24 Ordre empirique du schéma

Etant donné une méthode numérique $(\mathbf{S}^n)_{n\in\mathbb{N}}$, on définit son ***ordre empirique*** par

$$\mathrm{Ord}(\Delta t) \;:=\; \log_{10}\left(\frac{\mathrm{Err}(\Delta t)}{\mathrm{Err}(\Delta t/10)}\right). \tag{3.39}$$

Remarque : il est également possible de diviser le pas de temps par 2 et non par 10 (ce qui se révèle parfois plus adapté puisque les temps de calculs augmentent linéairement avec le nombre de pas de temps). Dans ce cas, il faut prendre le logarithme à base 2 dans la définition précédente.

Pour la pratique : on calcule le logarithme en base 10 de l'ordre empirique pour Δt, $\Delta t/10$, $\Delta t/10^2$, $\Delta t/10^3$, *etc...* et on trace le graphe ainsi obtenu (avec les deux axes de coordonnées munies d'échelles logarithmiques). On peut alors faire une régression linéaire et la pente de la droite obtenue est alors l'ordre empirique moyen du schéma. En python avec la bibliothèque `pylab`, on utilise la commande `loglog` pour avoir des graphiques avec des échelles logarithmiques sur les axes.

Si la solution exacte n'est pas connue, ce qui est le cas dans la majorité des situations, il n'est pas possible de calculer l'erreur exacte (3.37). Dans ce cas, Une solution possible et qui fonctionne très bien consiste à choisir $\mathbf{S}^n[\Delta t/10]X_0$ comme solution *presque exacte* lors du calcul de l'erreur. Plus précisément, on définit :

DÉFINITION 3.25 Erreur approchée et Ordre empirique approché

Étant donné une méthode numérique $(\mathbf{S}^n)_{n\in\mathbb{N}}$, on définit l'erreur approchée par :

$$\mathscr{E}(\Delta t) \;:=\; \max_{\substack{n\in\mathbb{N}\\ n\Delta t\leq T}} \left|\mathbf{S}^n[\Delta t]X - \mathbf{S}^{10n}[\Delta t/10]X\right|. \tag{3.40}$$

On définit ensuite l'ordre empirique approché par analogie avec (3.39) :

$$\mathscr{O}(\Delta t) \;:=\; \log_{10}\left(\frac{\mathscr{E}(\Delta t)}{\mathscr{E}(\Delta t/10)}\right). \tag{3.41}$$

Ainsi, on résout numériquement l'équation différentielle pour les pas de temps Δt, $\Delta t/10$, $\Delta t/10^2$, $\Delta t/10^3$, *etc...* et on stocke en mémoire les résultats obtenus. On peut alors calculer les erreur approchées et les tracer en double échelle logarithmique. Ces points vont s'aligner sur une droite dont la pente sera l'ordre empirique approché moyen du schéma.

3.3.4 Choisir le bon schéma et le bon pas de temps

Choisir le bon schéma

Il n'existe pas de méthodologie générale pour bien choisir son schéma et il fait partie du métier d'ingénieur de bien connaître et s'approprier un bon nombre de schémas numérique afin d'utiliser judicieusement celui qui est le plus adapté au problème étudié. Afin d'éclairer les choix de schémas à faire, voici un tableau récapitulant les idées principales qui guident ce choix :

Problème :	**Solutions classiques :**
Je souhaite une solution très précise	Runge-Kutta 3 ou 4 et petits Δt
Je souhaite une solution rapidement	Méthode de Heun avec Δt moyens
J'ai des problèmes de stabilité	Impliciter le schéma
Je souhaite connaître le comportement en temps long	Schéma implicite et grand Δt
Les trajectoires ont un fort rayon de courbure	Méthode de Ralston d'ordre 2 ou 3
$\mathcal{F}$ est très longue à calculer	Adams-Bashforth d'ordre 2 ou 3
Je souhaite préserver le Hamiltonien	Störmer-Verlet ou Beeman
Mon champ de vitesse est "presque linéaire"	Intégrateur exponentiel
Variations brusques du champs de vitesse	Méthodes à pas de temps adaptatif
Variations brusques avec instabilités	Backward Differentiation Formula

etc...

Important : il s'agit d'un guide et non d'une vérité absolue ; il peut y avoir de nombreuses surprises quant au comportement d'un schéma particulier sur un problème particulier (et il n'y a pas de miracles non plus...). Le mieux reste quand même d'essayer de nombreuses méthodes différentes afin de les comparer. Voire dans certains cas, fabriquer sa propre méthode numérique adaptée au problème étudié : il faut savoir rester ouvert et inventif ;-)

Choisir le bon pas de temps

Choisir le bon pas de temps en revanche est beaucoup plus simple. L'objectif est de minimiser le temps de calcul de la machine étant donnée une précision voulue. La précision se calcule et s'estime à l'aide des outils de la section 3.3.3. Le temps écoulé se mesure en `Python` à l'aide de la bibliothèque `datetime`. Ajuster le pas de temps se fait de la manière suivante :

Méthode (choisir le bon pas de temps) :

- On commence par faire une intégration très grossière (par exemple avec $\Delta t = 1$).
- On fait tourner l'algorithme avec le pas de temps $\Delta t/10$
  - → Si le temps d'exécution du programme dépasse le temps maximal autorisé : on arrête l'algorithme en on renvoie Δt comme temps optimal.
  - → Sinon, on calcule l'erreur approchée (3.40) en prenant la solution à $\Delta t/10$ comme solution "quasi-exacte".
- Si l'erreur est inférieur au seuil fixé, on a trouvé que Δt était le pas de temps optimal.
- Sinon on remplace Δt par $\Delta t/10$ en on recommence à l'étape 2.

Il est bien-sûr possible d'être plus précis quand à la façon dont on ajuste le pas de temps.

Remarque TRES importante : Lorsque l'on évalue les mérites comparés des différents schémas numériques pour un problème donné, on les classe en général en fonction du temps que met chacun pour résoudre le problème. Pour classer les différents schémas en fonction de leur temps de calcul ***il ne FAUT PAS prendre le même pas de temps Δt !...***

Méthode (choisir le bon schéma) :

- On commence par ***fixer une précision*** $\varepsilon > 0$ (qui va dépendre des données physiques et techniques du problème).
- Pour chaque schéma numérique considéré, on met en place la méthode précente à partir de laquelle on déduit le pas de temps Δt optimal. ***Chaque schéma va avoir son propre pas de temps optimal...***
- Pour chaque schéma *muni de son propre pas de temps optimal* et on mesure le temps que met une machine informatique pour exécuter l'algorithme. On classe alors les schémas en fonction les temps obtenus.

Choisir la bonne précision

La question centrale réside donc davantage dans le choix de la bonne précision. C'est un problème d'ingénierie délicat pour lequel il n'y a pas vraiment de réponse générale. Il faut choisir au cas-par-cas en fonction de deux contraintes antagoniques :

- La fiabilité du résultat obtenu.
- Le temps de calcul nécessaire à la résolution.

Les ingénieurs débutant ont parfois tendance à faire des calculs trop longs pour une précision inutile et qui n'a pas toujours de sens physique. Nous fournissons ci-dessous un tableau de quelques grandeurs physiques et qui permet de savoir si notre précision reste pertinente dans le cadre de la mécanique classique (en fonction du problème étudié). Ce type de tableau est très utile pour détecter des algorithmes qui sont trop précis étant donné l'objet physique et les modèles considérés. Par exemple, si on étudie la déformation du pneu d'une voiture franchissant un trottoir, il est inutile d'avoir un algorithme précis à $10^{-15}\ m$...

Grandeur physique considérée	Valeur en unité SI
Diamètre moyen de l'atome d'hydrogène dans l'état fondamental	$1.2 \times 10^{-10}\ m$
Distance de la terre au soleil au périhélie	$1.5 \times 10^{11}\ m$
Vitesse de la lumière dans le vide intersidéral	$3.0 \times 10^{8}\ m.s^{-1}$
Fréquence de transition hyperfine de l'atome de Césium	$9.2 \times 10^{9}\ s^{-1}$
Masse de l'atome d'hydrogène dans l'état fondamental	$1.7 \times 10^{-27}\ kg$
Masse du soleil	$2.0 \times 10^{30}\ kg$
Température de fusion nucléaire au centre du soleil	$1.5 \times 10^{10}\ K$
Formation des paires de Cooper dans les cristaux de mercure	$2.4\ K$
Énergie d'une liaison covalente entre 2 atomes d'hydrogène	$7.3 \times 10^{-19}\ kg.m^2.s^{-2}$
Nombre d'Avogadro (lien entre l'échelle moléculaire et macro.)	6.0×10^{23}
Moment magnétique d'un atome d'hydrogène	$9.3 \times 10^{-24}\ A.m^2$

etc...

3.4 Bilan du Chapitre et exercices

3.4.1 Ce qu'il faut retenir et savoir-faire

Ce dernier chapitre est consacré à présenter les principales idées fondamentales qui président à l'analyse numérique des EDO et à l'approximation des solutions à l'aide de programmes informatiques. Les principaux concepts à retenir et savoir-faire de ce chapitre sont les suivants :

(‡) Méthodes de Euler, Crank-Nicolson et Heun (formules, ordres, stabilité).
(‡) Erreur de consistance, erreur totale et théorème de convergence numérique.
(†) Stabilité numérique et théorème de propagation des erreurs numériques.
(†) Ordre de consistance d'un schéma (définition, développements limités).
(†) Méthodes A-stables et domaines de A-stabilité.

- Schéma du point-milieu explicite (formule, ordre, stabilité).
- Principes généraux des méthodes de Runge-Kutta (tableau de Butcher, ordres, stabilité).
- Les méthodes de Adams-Bashforth (formules, ordres, stabilité).
- Savoir implémenter ces méthodes numériques sur ordinateur.
- Savoir tracer les courbes d'ordre empirique d'une méthode.
- Savoir choisir le bon schéma et le bon pas de temps.

3.4.2 Exercices

Les exercices ci-dessous proposent d'étudier la consistance et la stabilité des méthodes numériques. Il est recommandé de traiter les exercices dans l'ordre. Les exercices les plus importants sont identifiés avec le symbole (⋆).

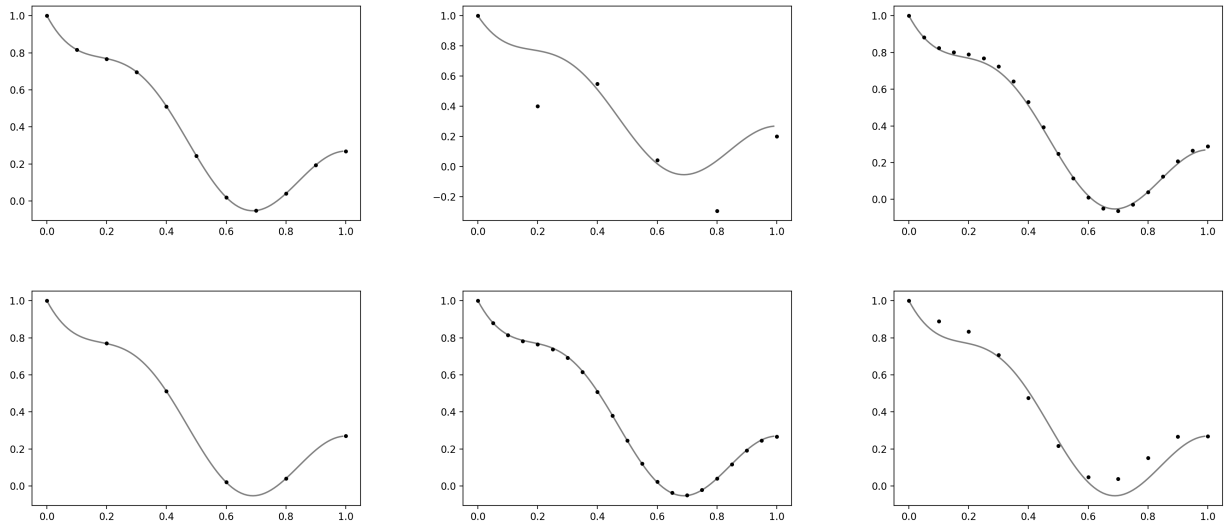


FIGURE 3.3 – Résolution numérique du problème de Cauchy :

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -3x + 2\sin(9t), \qquad \text{et} \qquad x(0) = 1, \tag{3.42}$$

sur l'intervalle de temps $[0, 1]$ avec différentes méthodes numériques et différents pas de temps pour l'exercice 3.1 (la solution exacte a été tracée en trait continu).

Exercice 3.1 (Précision des méthodes numériques)**.** ($\star$)

Nous avons résolu numériquement le problème de Cauchy (3.42) sur l'intervalle de temps $[0, 1]$ avec les méthodes numériques et pas de temps suivants :

Euler explicite ($\Delta t = 0.2$)	Euler implicite ($\Delta t = 0.1$)	Crank-Nicholson ($\Delta t = 0.05$)
Runge-Kutta 4 ($\Delta t = 0.2$)	Runge-Kutta 3 ($\Delta t = 0.1$)	Adams-Bashforth ($\Delta t = 0.05$)

Les résultats obtenus ont été tracés sur la figure 3.3 à la page précédente. Pour chaque tracé obtenu, dire à quelle méthode et quel pas de temps il correspond parmi le tableau ci-dessus. Justifier la réponse.

Exercice 3.2 (Stabilité des méthodes numériques)**.** ($\star$)

Même question pour les tracés ci-dessous sur l'intervalle de temps $[0, 20]$ avec :

Euler explicite ($\Delta t = 1$)	Euler explicite ($\Delta t = 0.05$)	Euler implicite ($\Delta t = 1$)

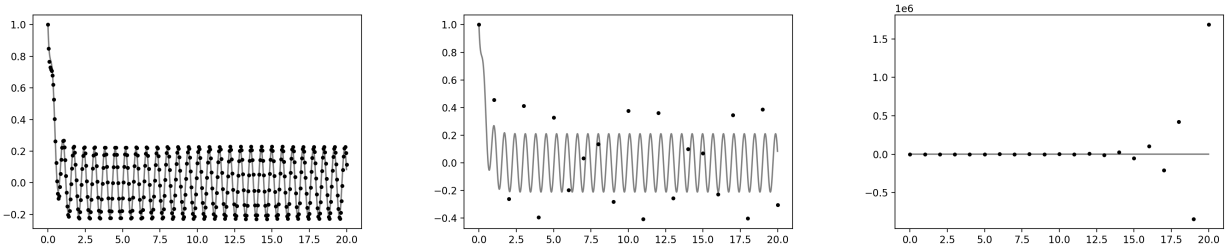


FIGURE 3.4 – Tracé de la solution exacte et de la résolution numérique (Exercice 3.2).

Exercice 3.3 (Temps de calculs relatifs)**.** ($\star$)

On a résolu un problème de Cauchy $\dot{X} = \mathcal{F}(t, X)$, $X(0) = X_0$ à l'aide d'un schéma numérique. La fonction $\mathcal{F}$ choisie étant lente à calculer, nous avons obtenu des temps très différentes en fonction des schémas et des pas de temps. Les performances obtenues sont les suivantes :

$1.51s$	$1.68s$	$6.3s$	$15.8s$	$64.7s$	$1760s$

Les schémas numériques qui ont été testés sont les suivants :

- Euler explicite avec $\Delta t = 0.01$,
- Euler explicite avec $\Delta t = 0.001$,
- Crank-Nicolson avec $\Delta t = 0.01$,
- Runge-Kutta 4 avec $\Delta t = 0.01$,
- Runge-Kutta 4 avec $\Delta t = 0.001$,
- Adams-Bashforth 4 avec $\Delta t = 0.01$.

Pour chaque schéma numérique et pas de temps dans la liste ci-dessus associer le temps de calcul mesuré parmi ceux du tableau. Justifier votre réponse.

Exercice 3.4 (Identifier l'ordre empirique d'un schéma)**.** ($\star$) Pour 4 schémas numériques, nous avons résolu le problème de Cauchy (3.42) avec différents pas de temps. Les schémas choisis sont : Euler explicite, Heun, Adams-Bashforth 2 et Runge-Kutta 3. Comme on sait que la solution exacte est égale à

$$X(t) = \Big(x_0 + \frac{1}{5}\Big)e^{-3t} + \frac{\sin(9t) - 3\cos(9t)}{15}, \tag{3.43}$$

on peut alors calculer l'erreur commise par chaque méthode numérique et pour chaque pas de temps Δt. On a ainsi tracé en échelle `log-log` l'erreur numérique en fonction du pas de temps (voir figure 3.5 ci dessous).

1) Pour chacune de ces 4 courbes, identifier à quel schéma numérique celle-ci correspond.

2) A l'aide de la figure 3.5, identifier pour chaque schéma l'ordre de grandeur du pas de temps qui permet d'atteindre une précision d'environ 10^{-6}.

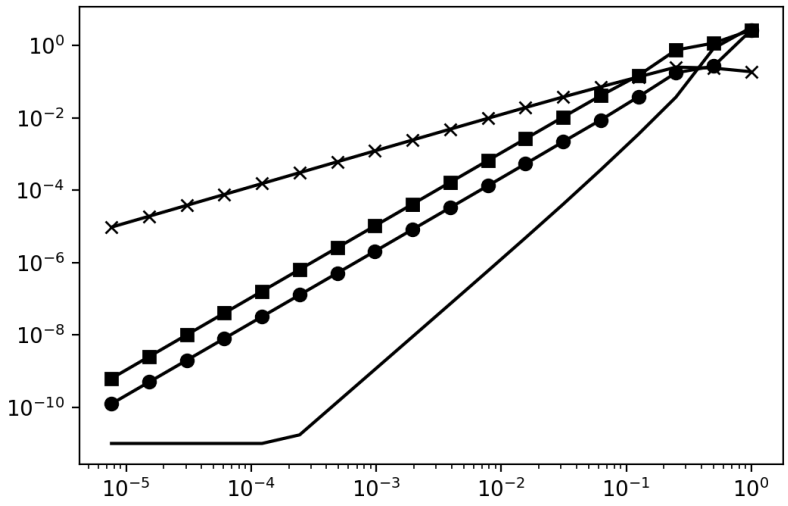


FIGURE 3.5 – Tracé de l'erreur en fonction du pas de temps pour 4 schémas : Euler explicite, Heun, Adams-Bashforth 2, Runge-Kutta 3.

Exercice 3.5 (Choisir le meilleur schéma)**.** ($\star$) On travaille de nouveau avec le problème de Cauchy (3.42) et à l'aide de la solution exacte (3.43) on peut calculer explicitement l'erreur obtenue (ainsi que les temps de calculs grâce à la bibliothèque `time` en python). Nous obtenons les erreurs et les temps de calcul suivants :

	Δt =0.1	Δt =0.01	Δt =0.001	Δt =1.e-4	Δt =1e-5	Δt = 1.e-6
EE	0.9 (0.0ms)	0.14 (0.1ms)	0.013 (1.0ms)	0.0012 (7.9ms)	1.3e-5 (74ms)	1.2e-6 (760ms)
EI	0.8 (0.0ms)	0.12 (0.2ms)	0.0012 (2.9ms)	1.3e-4 (34ms)	1.2e-5 (330ms)	1.3e-6 (3270ms)
Heun	1.9e-2 (0.0ms)	2.0e-4 (0.1ms)	2.2e-6 (1.4ms)	2.1e-8 (15ms)	2.1e-10 (149ms)	3.1e-12 (1410ms)
CN	2.0e-2 (0.1ms)	2.3e-4 (0.7ms)	2.2e-6 (4.1ms)	2.1e-8 (46ms)	2.2e-10 (446ms)	4.3e-12 (4457ms)
AB2	0.1 (0.0ms)	0.0011 (0.1ms)	1.2e-5 (1.2ms)	1.0e-7 (10ms)	1.1e-9 (110ms)	1.e-11 (1060ms)
RK3	1.2e-3 (0.0ms)	1.3e-6 (0.4ms)	1.1e-9 (2.5ms)	1.0e-12 (27ms)	1.1e-13 (240ms)	1.e-13 (2510ms)

1) Pour chacun de ces 6 schémas numériques, donner le pas de temps Δt pour lequel la précision de la solution obtenue est de l'ordre de grandeur 10^{-6}.

2) Pour chaque schéma, donner le temps de calcul pour lequel le schéma résout l'équation différentielle avec une précision de l'ordre de 10^{-6}. Quel est le meilleur schéma pour ce problème ?

3) Combien de temps de calcul est nécessaire pour résoudre cette équation avec Euler explicite pour obtenir une précision 10^{-12} (convertir en heures) ? Comparer avec Runge-Kutta 3.

Exercice 3.6 (Quelques exemples non-linéaires)**.** ($\star$) L'objectif de cet exercice est de s'entraîner à écrire les algorithmes standards pour des équations non-linéaires.

1) Écrire le schéma de Euler explicite pour l'équation de la sortie du tuyau (Exercice 1.3).

2) Écrire le schéma de Euler implicite pour l'équation logistique (Exercice 1.5).

3) Écrire le schéma de Euler explicite pour l'équation de Lotka-Volterra (Exercice 1.7).

4) Écrire le schéma de Crank-Nicolson pour l'équation du pendule simple (Exercice 2.9). Comparer avec le point-milieu implicite.

Exercice 3.7 (Stabilité de la méthode de Euler explicite dans le cas linéaire)**.** ($\star$)

1) On cherche à résoudre, avec la méthode de Euler-explicite, l'équation linéaire $\dot{x} = -3x$ avec la donnée initiale $x(0) = 1$. Écrire l'algorithme de Euler explicite pour cette équation et montrer que la suite (x_n) obtenue est une suite géométrique dont on précisera la raison. Pour quelles valeurs de $\Delta t > 0$ cette suite est-elle positive ? bornée ? convergente ?

2) En déduire la condition nécessaire et suffisante sur Δt pour que ce schéma associé à l'équation $\dot{x} = -3x$ soit numériquement stable en reconnaissant une série géométrique.

3) Retrouver cette condition de stabilité à l'aide du théorème de propagation des erreurs numériques (Théorème 3.2).

4) On souhaite à présent étudier le problème de Cauchy (3.42). Écrire l'algorithme de Euler explicite pour cette équation. Pour quelles valeurs de $\Delta t > 0$ cette suite est-elle bornée ?

5) Que devient la condition de stabilité obtenue précédemment ?

6) A l'aide d'une diagonalisation de matrice, trouver la condition nécessaire et suffisante pour que la méthode de Euler-explicite appliquée au système d'équations suivant soit stable :

$$\frac{\mathrm{d}x}{\mathrm{d}t} = y - x \qquad \text{et} \qquad \frac{\mathrm{d}y}{\mathrm{d}t} = -x - y. \tag{3.44}$$

7) Retrouver cette condition de stabilité à l'aide du théorème de propagation des erreurs numériques (Théorème 3.2).

8) Généraliser ce résultat à toutes les équations linéaires à coefficients constants $\dot{X} = AX$ en montrant que le schéma d'Euler explicite associé est stable sous condition sur Δt si et seulement si les valeurs propres de A sont toutes de partie réelle strictement négative.

9) Représenter l'ensemble des $\lambda \in \mathbb{C}$ tels que la méthode d'Euler-explicite appliquée à l'équation scalaire $\dot{x} = \lambda x$ donne une suite bornée lorsque $\Delta t = 1$ (domaine de A-stabilité).

Exercice 3.8 (Convergence de la méthode de Euler explicite)**.** ($\star$) Soit $\mathcal{F} : [0, T[\times\Omega \to \mathbb{R}^d$ un champ de vitesse un $\Delta t > 0$ un pas de temps. On note $X : [0, T[\to \mathbb{R}^d$ la solution exacte et (X_n) la solution approchée par le schéma d'Euler explicite.

1) A l'aide d'un développement limité, montrer que

$$X(t_1) - X_1 = \mathcal{O}(\Delta t^2).$$

En déduire que l'erreur de troncature locale est de l'ordre $\mathcal{O}(\Delta t^2)$ et que l'erreur de consistance est de l'ordre $\mathcal{O}(\Delta t)$.

2) On suppose à présent que le schéma est stable si Δt est inférieur à un certain seuil $\Delta t^\star$. Montrer que la méthode de Euler-explicite est convergente d'ordre 1 à l'aide du théorème de convergence numérique (Théorème 3.1).

3) On souhaite maintenant utiliser la méthode d'Euler explicite pour résoudre numériquement l'équation de l'oscillateur harmonique amorti ($\lambda > 0$) donnée par

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\theta + \lambda\frac{\mathrm{d}}{\mathrm{d}t}\theta + \theta = 0. \tag{3.45}$$

Montrer que le schéma de Euler explicite appliqué à cette équation est un schéma convergent d'ordre 1 lorsque $\Delta t \to 0$.

Exercice 3.9 (Stabilité et convergence de la méthode de Euler implicite)**.** ($\star$) On cherche à résoudre les équations linéaires dissipative avec la méthode de Euler implicite.

1) Écrire l'algorithme de Euler implicite pour cette équation et montrer que la suite (x_n) obtenue est une suite géométrique dont on précisera la raison. Pour quelles valeurs de $\Delta t > 0$ cette suite est-elle positive ? bornée ? convergente ?

2) Montrer que le schéma d'Euler implicite pour l'équation différentielle $\dot{x} = \lambda x$ avec $\lambda < 0$ est numériquement stable.

3) Représenter graphiquement l'ensemble des $\lambda \in \mathbb{C}$ tels que la méthode de Euler implicite intègre l'équation $\dot{x} = \lambda x$ de manière bornée lorsque $\Delta t = 1$ (domaine de A-stabilité).

4) A l'aide d'un développement limité, montrer que l'erreur de troncature locale est de l'ordre $\mathcal{O}(\Delta t^2)$. En déduire que le schéma est convergent d'ordre 1.

5) Étudier la stabilité de ce schéma pour l'équation de l'oscillateur harmonique amorti (3.45).

6) Généraliser ce résultat à toutes les équations linéaires à coefficients constants $\dot{X} = AX$ en montrant que le schéma d'Euler implicite est inconditionnellement stable dès que les valeurs propres de A sont toutes de partie réelle négative ou nulle.

Exercice 3.10 (Stabilité et convergence de la méthode de Crank-Nicolson)**.** ($\star$) L'objectif de cet exercice est d'étudier les propriétés de la méthode de Crank-Nicolson.

1) Représenter graphiquement l'ensemble des $\lambda \in \mathbb{C}$ tels que la méthode de Crank-Nicolson intègre l'équation $\dot{x} = \lambda x$ de manière bornée lorsque $\Delta t = 1$ (domaine de A-stabilité).

2) A l'aide de développements limités, montrer que l'erreur de troncature locale est de l'ordre de $\mathcal{O}(\Delta t^3)$. En déduire que cette méthode est convergente d'ordre 2 pour résoudre numériquement le système suivant :

$$\dot{x} = 2y - x, \qquad \text{et} \qquad \dot{y} = -4y - x. \tag{3.46}$$

3) En utilisant le théorème de propagation des erreurs numériques avec une norme bien choisie, montrer que la méthode de Crank-Nicolson donne un schéma convergent d'ordre 2 pour ce système :

$$\dot{x} = 2x + 5y, \qquad \text{et} \qquad \dot{y} = -x - 2y. \tag{3.47}$$

4) Montrer que la méthode de Crank-Nicolson est A-stable.

Exercice 3.11 (Stabilité et convergence pour les méthodes de Runge-Kutta 2)**.** ($\star$) Les méthodes de Runge-Kutta-2 explicites sont données par le tableau de Butcher suivant ($a, b_1, b_2, c \in [0,1]$) :

$$\begin{array}{c|cc} 0 & & \\ c & a & \\ \hline & b_1 & b_2 \end{array}$$

1) Pour quelles valeurs de a, b_1, b_2, c est-ce que l'on obtient la méthode du point-milieu explicite ? Celle de Heun ?

2) Étude de la méthode du point-milieu explicite :

2.1) Calculer $\ddot{x}(t)$ pour $x : [0,T] \to \mathbb{R}$ solution d'une équation autonome 1D : $\dot{x} = f(x(t))$.

2.2) En déduire le développement de Taylor en Δt de $x(t+\Delta t)$ à l'ordre 2.

2.3) On pose $k_1 = f(x(t))$ et $k_2 = f(x(t) + \Delta t k_1/2)$. Calculer le développement de Taylor en Δt à l'ordre 2 de $x(t) + \Delta t k_2$.

2.4) Conclure que le point-milieu explicite a une erreur de troncature de l'ordre $\mathcal{O}(\Delta t^3)$.

3) Étude de la méthode de Heun (parfois appelée méthode des trapèzes explicite) :

3.1) On pose $k_1 = f(x(t))$ et $k_2 = f(x(t) + \Delta t k_1)$. Calculer le développement limité en Δt de k_2 à l'ordre 1.

3.2) En déduire le développement limité en Δt à l'ordre 2 de $x(t) + \Delta t(k_1 + k_2)/2$.

3.3) Conclure que la méthode de Heun a une erreur de troncature de l'ordre $\mathcal{O}(\Delta t^3)$.

4) Généralisation : pour quelles valeurs de a, b_1, b_2, c est-ce que l'erreur de troncature locale est de l'ordre $\mathcal{O}(\Delta t^3)$? En déduire que, sous une hypothèse de stabilité numérique, ce sont des méthodes convergentes d'ordre 2.

5) On se concentre maintenant uniquement sur les méthodes d'ordre 2. Trouver une condition sur Δt pour pouvoir intégrer l'équation $\dot{x} = \lambda x$ de manière stable avec $\lambda < 0$?

6) Pour quels $z \in \mathbb{C}$ est-ce qu'il existe un $t_z > 0$ tel que pour tout $t \in [0, t_z]$ on ait tz qui appartienne au domaine de A-stabilité ?

7) Calculer t_z dans le cas où $z = (-1, 1)$. En déduire une condition sur Δt pour que ces schémas de Runge-Kutta 2 explicites intègrent de manière stable le système (3.44) de l'exercice 3.7.

Exercice 3.12 (Point-Milieu implicite)**.**

1) A l'aide de développements limités, montrer que l'erreur de troncature locale pour la méthode du point-milieu implicite est $\mathcal{O}(\Delta t^3)$.

2) On souhaite appliquer la méthode du point-milieu implicite au système suivant :

$$\dot{x} = 2x + 5y - 3z, \qquad \dot{y} = -x - 2y + z, \quad \text{et} \quad \dot{z} = 4x - 2y - 5z. \tag{3.48}$$

Montrer que le schéma obtenu est numériquement stable et convergent d'ordre 2.

3) La méthode du point-milieu implicite donne-t-elle un schéma convergent pour le système (3.47) de l'exercice 3.10 ?

Exercice 3.13 (Quantités conservées par le flot numérique)**.** Soit une EDO autonome

$$x'(t) = f(x(t)),$$

où $f\colon \mathbb{R}^d \to \mathbb{R}^d$. On suppose qu'il existe une fonction $\mathcal{E}\colon \mathbb{R}^d \to \mathbb{R}$ qui est une quantité conservée par le flot. C'est à dire telle que, pour toute solution x de l'EDO, l'application $t \mapsto \mathcal{E}(x(t))$ est constante.

1) Soit x_n la n^{e} itérée obtenue en appliquant la méthode d'Euler Explicite à l'EDO ci-avant. Montrer que

$$\mathcal{E}(x_{n+1}) = \mathcal{E}(x_n) + \mathcal{O}(\Delta t^2).$$

2) Plus généralement, montrer que si x_n est la n^{e} itérée obtenue en appliquant une méthode d'ordre p à l'EDO ci-avant, on a

$$\mathcal{E}(x_{n+1}) = \mathcal{E}(x_n) + \mathcal{O}(\Delta t^{p+1}).$$

Exercice 3.14 (Quantités conservées par le flot numérique (2))**.**

1) Calculer la solution exacte du système suivant :

$$\begin{aligned} x'(t) &= -y(t), \\ y'(t) &= x(t), \end{aligned}$$

avec $x(0) = 1$ et $y(0) = 0$. Quelle est alors la trajectoire du point $(x(t), y(t))$? En déduire le comportement de $x^2(t) + y^2(t)$.

2) Appliquer le schéma d'Euler explicite à ce système et calculer $x_n^2 + y_n^2$. Que conclure ?

3) Faire de même avec Euler Implicite et la méthode de trapèzes implicites (Crank-Nicolson).

4) Écrire une formule de Runge-Kutta explicite à 2 étages pour ce problème. Quels sont les coefficients qui permettent d'obtenir le comportement souhaité de $x_n^2 + y_n^2$?

Exercice 3.15 (Application à la discrétisation d'une EDP)**.** On souhaite utiliser les techniques de résolution numérique des EDO pour résoudre des EDP discrétisées en espace. On va travailler avec l'équation de la chaleur 1D sans second membre définie pour $x \in [0, 1]$:

$$\frac{\partial \theta}{\partial t} - \lambda \frac{\partial^2 \theta}{\partial x^2} = 0, \qquad \text{et} \qquad \theta(t, 0) = \theta(t, 1) = 0.$$

où $\theta(t, x)$ représente la température au temps t et à la position x et $\lambda > 0$ la diffusivité thermique du matériau. Nous allons à présent discrétiser la dérivation spatiale avec $d+2$ points de discrétisation (en comptant les extrémités). Nous allons pour cela approcher la fonction $x \mapsto \theta(t, x)$ pour tout t par un vecteur $\Theta(t) \in \mathbb{R}^d$. Plus précisément, on cherche à construire Θ pour avoir $(k = 1, \dots, d)$

$$\theta\left(t, x = \frac{k}{d}\right) \approx \Theta_k(t). \tag{3.49}$$

Pour ce faire on approche le laplacien à l'aide d'une différence finie :

$$\frac{\partial^2}{\partial x^2}\theta\left(t, x = \frac{k}{d}\right) \approx \frac{1}{\Delta x^2}\Big(\Theta_{k-1}(t) - 2\Theta_k(t) + \Theta_{k+1}(t)\Big).$$

Dans l'expression ci-dessus on mets la convention de remplacer Θ_0 et Θ_{d+1} par la donnée au bord nulle.

1) Montrer que l'évolution en temps du vecteur $\Theta \in \mathbb{R}^d$ est donnée par une EDO linéaire à coefficients constants $\dot{\Theta} = A\Theta$ où A est une matrice à préciser.

2) Montrer que le polynôme caractéristique de A est scindé à racines simples et calculer ses racines.

3) Trouver la condition de stabilité sur Δt pour le schéma de Euler-Explicite appliqué à cette EDO. Même question avec Euler Implicite.

4) On souhaite améliorer la stabilité pour Euler explicite en procédant à une implicitation sur la diagonale de la discrétisation spatiale. Autrement-dit :

$$\frac{\partial^2}{\partial x^2}\theta\left(t_n, x = \frac{k}{d}\right) \approx \frac{1}{\Delta x^2}\Big(\Theta_{k-1}(t_n) - 2\Theta_k(t_{n+1}) + \Theta_{k+1}(t_n)\Big).$$

Trouver la condition de stabilité pour ce nouveau schéma. Quel est l'intérêt de ce schéma ?

Exercice 3.16 (Méthodes d'Adams-Bashforth et conditions initiales)**.**

1) Calculer les expressions des méthodes d'Adams-Bashforth à 1, 2 et 3 pas. Donner l'inéquation satisfaite par les éléments du domaine de A stabilité pour Adams-Bashforth 2.

Indication : on peut toujours réécrire les suite récurrentes linéaires d'ordre 2 *:*

$$x_{n+1} = a\, x_n + b\, x_{n-1},$$

sous la forme d'une suite récurrente d'ordre 1 *vectorielle :*

$$\begin{pmatrix} x_{n+1} \\ x_n \end{pmatrix} = \begin{pmatrix} a & b \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x_n \\ x_{n-1} \end{pmatrix}.$$

2) En déduire que Adams-Bashforth est stable sous conditions pour intégrer l'équation $\dot{x} = -\lambda x$ avec $x : [0, T] \to \mathbb{R}$ et $\lambda > 0$. Donner la condition de stabilité sur le pas de temps Δt (en fonction de λ).

3) On souhaite à présent appliquer la méthode d'Adams-Bashforth à 3 pas. Afin d'obtenir les premières valeurs de (x_n), on choisit d'utiliser la méthode d'Adams Bashforth à 1 pas pour calculer x_1 et celle à 2 pas pour calculer x_2. On considère $x'(t) = -\lambda x(t)$, $\lambda > 0$.

4) Donner une estimation de l'erreur commise pour obtenir x_1 en fonction du pas Δt.

5) Donner une estimation de l'erreur commise pour faire $N - 2$ itérations de la méthode d'Adams à 3 pas en fonction de Δt où $N = T/\Delta t$. Que peut-on en conclure ?

Exercice 3.17 (Une méthode de Runge-Kutta 3)**.** On commence par rappeler que si g est suffisamment régulière, la formule de quadrature

$$\int_0^1 g(s)\mathrm{d}s \approx \frac{1}{4}g(0) + \frac{3}{4}g(2/3)$$

a pour degré d'exactitude 3. On veut s'inspirer de cette méthode de quadrature pour obtenir une méthode de Runge-Kutta d'ordre 3. Soit x une solution de $x'(t) = f(t, x(t))$.

1) Ecrire la formule de Duhamel pour x entre t et $t + \Delta t$.

2) Utiliser la formule du point milieu pour approcher $x(t_n + 2\Delta t/3)$ à l'ordre 2 .

3) En déduire avec la formule de quadrature une méthode de Runge-Kutta à $s = 3$ étages.

4) En utilisant les développements de Taylor, vérifier que la méthode est d'ordre 3.

5) Calculer le domaine de A-stabilité de cette méthode (donner l'équation).

Exercice 3.18 (Méthodes d'Euler Symplectiques et de Störmer-Verlet)**.** Les méthodes d'Euler Symplectiques et de Störmer-Verlet sont des méthodes multi-pas qui permettent de préserver le hamiltonien numériquement avec un ordre assez élevé. On considère $\phi\colon \mathbb{R} \to \mathbb{R}$ de classe $\mathcal{C}^1$ et le système d'EDO mécanique :

$$\begin{aligned} v' &= -\phi'(x), \\ x' &= v. \end{aligned}$$

1) La méthode d'Euler symplectique-a est donnée par la formule de récurrence

$$\begin{aligned} v_{n+1} &= v_n - \Delta t \phi'(x_n), \\ x_{n+1} &= x_n + \Delta t v_{n+1}. \end{aligned}$$

Estimer l'erreur de troncature, l'ordre de consistance et l'ordre de convergence pour cette méthode (Pour l'ordre de convergence, on se restreindra au cas linéaire).

2) Faire pareil pour la méthode d'Euler symplectique-b donnée par :

$$\begin{aligned} x_{n+1} &= x_n + \Delta t v_n, \\ v_{n+1} &= v_n - \Delta t \phi'(x_{n+1}). \end{aligned}$$

3) Faire pareil pour la méthode de Störmer–Verlet donnée par :

$$x_{n+1} \;=\; 2x_n - x_{n-1} - \Delta t^2 \phi'(x_n)$$

4) Même question pour Velocity-Verlet (Verlet en vitesse) donnée par :

$$\begin{aligned} x_{n+1} &= x_n + \Delta t v_n - \frac{\Delta t^2}{2} \phi'(x_n), \\ v_{n+1} &= v_n - \frac{\Delta t}{2} \Big(\phi'(x_{n+1}) + \phi'(x_n) \Big). \end{aligned}$$

Remarque : La différence entre l'ordre de consistance et l'ordre de convergence s'explique parce que la méthode est multi-pas. Le théorème de convergence numérique énoncé dans le cours ne s'applique pas ici.

5) On définit $\mathcal{H}$ le Hamiltonien de ce système par

$$\mathcal{H} \;:=\; \frac{v^2}{2} + \phi(x).$$

Pour chacune de ces 4 méthodes, calculer l'ordre de consistance et de convergence pour le Hamiltonien $\mathcal{H}$ et les comparer avec les résultats de l'exercice 3.13.

Exercice 3.19 (Une propriété des méthodes de Runge-Kutta)**.** Dans cet exercice, nous souhaitons démontrer qu'il est judicieux d'avoir $c_i = \sum_{j<i} a_{i,j}$. Soit une méthode de Runge-Kutta donné par le tableau de Butcher générique de l'équation (3.34).

1) Montrer que dans le cas d'une EDO autonome, x_{n+1} est indépendant du choix des c_i.

2) On considère le problème de Cauchy non autonome

$$\dot{x}(t) = \mathcal{F}\big(t, x(t)\big), \qquad \text{et} \qquad x(0) = x_0,$$

où $\mathcal{F} : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$. Montrer que le problème de Cauchy précédent est équivalent au problème de Cauchy autonome suivant (dont les inconnues sont y et σ) :

$$\begin{cases} \dot{y}(t) = \mathcal{F}\big(\sigma(t), y(t)\big), \\ \dot{\sigma}(t) = 1. \end{cases} \qquad \text{et} \qquad \begin{cases} x(0) = x_0, \\ \sigma(0) = 0. \end{cases}$$

3) Écrire la méthode de Runge-Kutta générale pour ce système autonome.

4) Montrer que, lorsque $c_i = \sum_{j<i} a_{i,j}$, appliquer une méthode de Runge-Kutta au premier ou au second système est équivalent.

5) En déduire qu'il suffit de chercher l'ordre d'une méthode de Runge- Kutta seulement pour les systèmes autonomes.

Exercice 3.20 (Formule de Milne)**.** La formule de Milne est la suivante :

$$x_{n+1} = x_{n-1} + \frac{\Delta t}{3}\Big(\mathcal{F}(t_{n+1}, X_{n+1}) + 4\mathcal{F}(t_n, X_n) + \mathcal{F}(t_{n-1}, X_{n-1})\Big).$$

1) Montrer que c'est une méthode multipas d'ordre 4.

2) Calculer le domaine de A-stabilité de cette méthode (donner l'équation). Cette méthode est-elle A-stable ?

Exercice 3.21 (Backward Differentiation Formula)**.** La *Backward Differentiation Formula* est une méthode multi-pas très similaire à la méthode de Adams-Bashforth. La principale différence étant ici que la méthode est implicite. Il s'agit d'une méthode plutôt efficace pour gérer des problèmes de stabilité liées à une variation brusque du champ de vitesse. Cette méthode à l'ordre k, notée BDF-k, s'écrit

$$\sum_{j=0}^{k} \beta_{j,k}\, X_{n+j} \;=\; \Delta t\, a_k\, \mathcal{F}\big(t_{n+k}, X_{n+k}\big)$$

Les coefficients $\beta_{j,k}$ et a_k sont calculés à l'aide des polynômes de Lagrange comme pour Adams-Bashforth. Cependant, le calcul des coefficients d'Adams-Bashforth va donner un résultat différents (puisque l'approximation de la dérivée est différente). On les range dans le tableau suivant :

k	$\beta_{0,k}$	$\beta_{1,k}$	$\beta_{2,k}$	$\beta_{3,k}$	$\beta_{4,k}$	$\beta_{5,k}$	a_k
1	-1	1					1
2	1	-4	3				2
3	-2	9	-18	11			6
4	3	-16	36	-48	25		12
5	-12	75	-200	300	-300	147	60

1) Montrer que BDF-1 est équivalente à Euler-Implicite.

2) Calculer l'erreur de troncature locale pour les méthodes BDF-k.

3) On dit qu'un schéma numérique est *zéro-stable* s'il intègre de manière stable l'équation $\dot{X} = 0$. Montrer que ces schémas BDF sont zéro-stables. (*Remarque : Le schéma BDF-*6 *est également zéro-stable mais pas les schémas BDF d'ordre plus élevé !*)

4) On admet que ces schémas sont stables. Montrer alors que l'ordre de consistance est égal à l'ordre de convergence (faire le cas linéaire).

Exercice 3.22 (Runge-Kutta 4)**.** Montrer que la méthode de Runge-Kutta 4 classique est bien une méthode d'ordre 4 stable sous condition. Même question pour la règle des $3/8^{\text{ièmes}}$.

*
* *