Equations différentielles ordinaires
Equations aux dérivées partielles

Analyse théorique et numérique

Franck Boyer

Master Mathématiques et Applications
Première année

Université de Provence - Université Paul Cézanne

5 janvier 2014

Table des matières

Chapitre I

Théorie des équations différentielles

Dans toute la première partie de ce chapitre, on va se concentrer sur les équations du premier ordre. Les équations d'ordre supérieur seront évoquées par la suite.

En introduction nous pouvons motiver cette étude par les quelques exemples suivants, issus de la modélisation en dynamique des populations.

– Modèle de Malthus : Il s'agit de dire que la population totale $t \mapsto N(t)$ évolue seulement au gré des naissances (taux $b > 0$ par unité de temps) et des décès (taux $d > 0$ par unité de temps). On obtient l'équation différentielle

$$N'(t) = (b-d)N(t).$$

Celle-ci peut s'intégrer à vue, ce qui donne

$$N(t) = N_0 e^{(b-d)t}, \quad \forall t > 0.$$

Son comportement dépend donc du signe de $b - d$, ce qui est intuitivement clair :

– Si $b > d$, les naissances sont prépondérantes sur les décès et la population croît exponentiellement au cours du temps.

– Si $b < d$, c'est le phénomène inverse et la population décroît exponentiellement.

– Si $b = d$, on a un équilibre parfait.

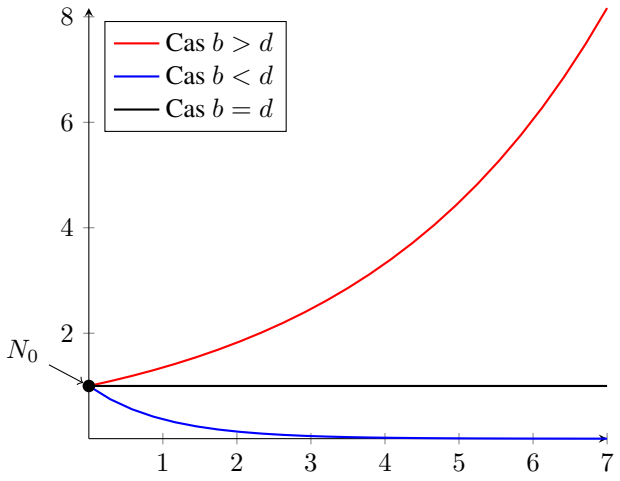


FIGURE I.1 – Le modèle de Malthus

– Modèle logistique : Le modèle de Malthus a le défaut de ne pas prévoir de limitation dans la capacité d'extension de la population. Or, on comprend bien que ce phénomène peut avoir une importance cruciale, par exemple si on imagine que les ressources en nourriture disponibles sont limitées par exemple. C'est pourquoi le modèle malthusien a été modifié de la façon suivante

$$N'(t) = (b-d)N(t)\left(1 - \frac{N(t)}{K}\right).$$

Le paramètre $K > 0$ modélise la taille critique de la population considérée. On peut comprendre ce nouveau modèle à partir du modèle de Malthus de plusieurs façons.

– Ou bien on voit le facteur $(1 - N/K)$ comme un correctif au taux global d'évolution $b - d$ qui a tendance a diminuer plus N augmente

– Ou bien, en développant on voit le nouveau terme dans l'équation

$$-\frac{b-d}{K}N^2,$$

comme un terme de compétition qui modélise le fait que quand N est grand les individus de l'espèce considérée on tendance à lutter pour les ressources et ce de façon d'autant plus importante que leur probabilité de rencontre est élevée (celle-ci étant proportionnelle au carré de N).

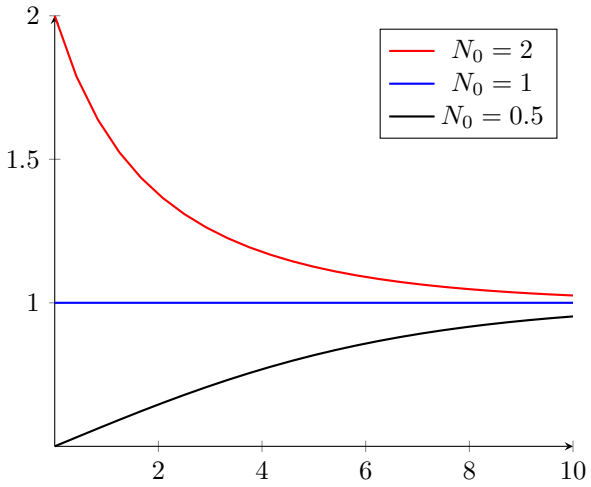


FIGURE I.2 – Le modèle logistique

On observe dans ce cas que toutes les solutions semblent converger en temps long vers la valeur K (qu'on a pris égal à 1 dans cette figure).

– Le modèle de Lotka-Volterra. Ce modèle met en jeu deux espèces : les proies (dont la population à l'instant t est notée $x(t)$) et les prédateurs (notés $y(t)$). On peut penser aux sardines et aux requins par exemple.

En l'absence de prédateurs, les proies se développent selon une loi de type Malthus (on suppose qu'elles disposent de nourriture en quantité illimitée). En l'absence de proies, les prédateurs ont tendance à disparaître selon une loi de type Malthus à taux négatif. Il faut maintenant modéliser le fait que les "rencontres" proies-prédateurs donnent lieu à la disparation de proies mangées par les prédateurs, ce qui participe à la survie des prédateurs. Mathématiquement, le nombre moyen de telles rencontres par unité de temps est proportionnel au produit des deux populations concernées et chacune de ces rencontres contribue positivement à l'évolution de y et négativement à l'évolution de x. On obtient le système suivant

$$\begin{cases} x'(t) = ax(t) - bx(y)y(t), \\ y'(t) = -cy(t) + dx(t)y(t). \end{cases}$$

On peut montrer que les solutions de ce système (on trace ici la courbe $t \mapsto (x(t), y(t))$ dans le plan x/y appelé **plan de phases**) ont l'allure donnée dans la figure suivante :

On observe un phénomène attendu de périodicté des solutions.

I Définitions de base

Définition I.1

Soit $I \subset \mathbb{R}$ un intervalle ouvert de $\mathbb{R}$ et $F : I \times \mathbb{R}^d \mapsto \mathbb{R}^d$ une application. On appelle **solution** *de l'équation différentielle*

$$y' = F(t, y), \tag{I.1}$$

tout couple (J, y) où $J \subset I$ est un sous-intervalle de I et y une fonction dérivable définie sur J telle que

$$\forall t \in J, \quad y'(t) = F(t, y(t)).$$

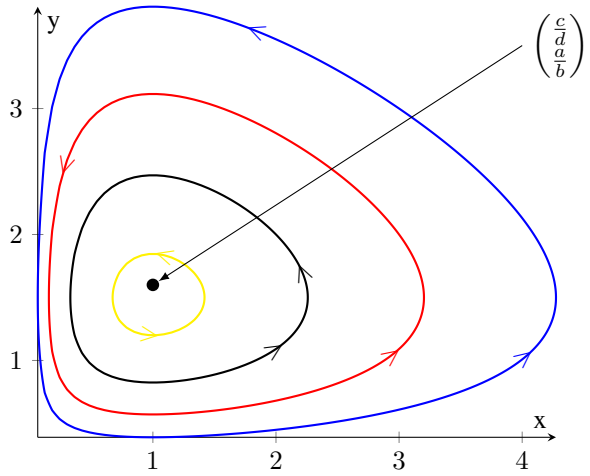


FIGURE I.3 – Portrait de phase pour le modèle de Lotka-Volterra

Remarque I.2

- *On peut tout à fait étendre ces définitions au cas où F est définie sur un ouvert quelconque de $\mathbb{R} \times \mathbb{R}^d$.*
- *L'équation* (I.1) *est appelée* du premier ordre *car elle ne fait intervenir que les dérivées premières de la fonction inconnue.*
- *La forme la plus générale d'une équation différentielle est de la forme (dite* non résolue*)*

$$F(t, y, y') = 0,$$

mais nous ne traiterons pas ce cas ici. Notons que le théorème des fonctions implicites, permet de se ramener au cas résolu dans un certain nombre de cas.

Dans le cas général, il est tout à fait possible que l'intervalle J ne soit pas égal à I et qu'on ne puisse pas faire mieux. La notion de solution maximale permet de donner un sens précis à *"on ne peut pas faire mieux"*.

Définition I.3 (Solution maximale)

On dit que (J, y) est une solution maximale de (I.1) *s'il n'existe pas de solution $(\tilde{J}, \tilde{y})$ vérifiant $J \subsetneq \tilde{J}$ et $\tilde{y}_{|J} = y$.*

La figure I.4 illustre la notion de prolongement de solution : la solution (J_2, y_2) prolonge la solution (J_1, y_1).

En utilisant le lemme de Zorn, on peut montrer le résultat suivant, sans aucune hypothèse sur l'équation considérée.

Proposition I.4

Pour toute solution (J, y) de (I.1)*, il existe* **au moins une** *solution maximale $(\tilde{J}, \tilde{y})$ qui prolonge (J, y), c'est-à-dire telle que $J \subset \tilde{J}$ et $\tilde{y}_{|J} = y$.*

Définition I.5

Toute solution (J, y) de (I.1) *définie sur l'intervalle $J = I$ tout entier est dite* **globale**.

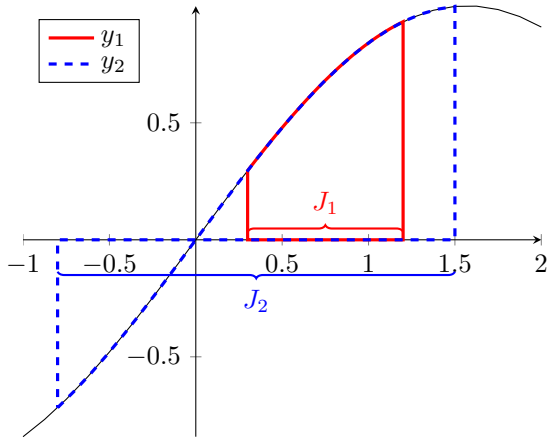


FIGURE I.4 – Notion de prolongement de solution et de solution maximale

II Théorie de Cauchy-Lipschitz

II.1 Le problème de Cauchy : définition et énoncé du théorème principal

Il arrive qu'on ne recherche pas toutes les solutions d'une EDO mais seulement celles qui vérifient certaines conditions, dites *conditions initiales de Cauchy* ou tout simplement *conditions de Cauchy*.

Définition I.6 (Problème de Cauchy)

Soit $F : I \times \mathbb{R}^d \mapsto \mathbb{R}^d$, $t_0 \in I$ et $y_0 \in \mathbb{R}^d$. On appelle solution (resp. solution maximale) du problème de Cauchy associé à la donnée (t_0, y_0) toute solution (J, y) (resp. solution maximale) de $y' = F(t, y)$ vérifiant de plus

$$t_0 \in \mathring{J}, \quad et \quad y(t_0) = y_0.$$

Le théorème **fondamental** de ce chapitre est le suivant

Théorème I.7 (Cauchy-Lipschitz, forme faible)

Si la fonction $F : I \times \mathbb{R}^d \mapsto \mathbb{R}^d$ est de classe $\mathcal{C}^1$ alors pour toute donnée de Cauchy $(t_0, y_0) \in I \times \mathbb{R}^d$, il existe, au voisinage de t_0, une unique solution du problème de Cauchy associé.
En particulier, pour toute telle donnée, il existe une unique solution maximale associée et toute autre solution vérifiant la condition de Cauchy est une restriction de cette solution maximale.

En réalité, ce théorème est encore vrai sous des hypothèses plus faibles.

Définition I.8

Une fonction continue $F : I \times \mathbb{R}^d \mapsto \mathbb{R}^d$ est dite **localement Lipschitzienne** *par rapport à la variable d'état (ou* à la seconde variable*). Si pour tout $(t_0, y_0) \in I \times \mathbb{R}^d$, il existe $C_{t_0,y_0} > 0$ et un voisinage U de (t_0, y_0) dans $I \times \mathbb{R}^d$ tel que*

$$\forall t \in I, \forall y_1, y_2 \in \mathbb{R}^d, \text{tels que } (t, y_1) \in U \text{ et } (t, y_2) \in U, \quad \text{on a } \|F(t, y_1) - F(t, y_2)\| \leq C_{t_0,y_0} \|y_1 - y_2\|. \quad \text{(I.2)}$$

Grâce au théorème des accroissements finis, on vérifie que toute fonction de classe $\mathcal{C}^1$ est localement lipschitzienne par rapport à sa deuxième variable. C'est pourquoi le théorème suivant est bien plus fort que le précédent.

Théorème I.9 (Cauchy-Lipschitz, forme forte)

Le théorème de Cauchy-Lipschitz est encore vrai si F est continue et localement lipschitzienne par rapport à la variable d'état.

Ce théorème admet plusieurs démonstrations qu'il est peut être bon de connaître, ce sera l'objet du paragraphe suivant.

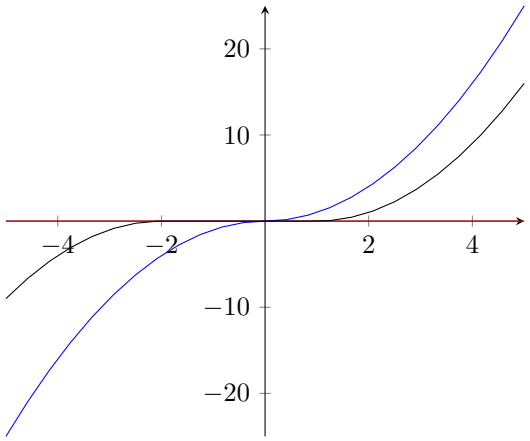
Remarque I.10

– *La propriété d'existence de solutions maximales persiste sous la seule hypothèse de continuité de F (Théorème de Cauchy-Arzela).*

– *L'exemple canonique d'équation pour laquelle le problème de Cauchy n'a pas de solution unique est l'équation suivante*

$$y' = 2\sqrt{|y|},$$

qui possède une infinité de solutions vérifiant $y(0) = 0$ dont la fonction identiquement nulle et la fonction $y(t) = |t|t$ (qui est bien de classe $\mathcal{C}^1$!). Nous dessinons trois telles solutions dans la figure ci-dessous :



II.2 Démonstrations du théorème de Cauchy-Lipschitz

La démonstration de ce théorème peut se faire de plusieurs manières.

Elles partent toutes de la constatation que (J, y) est solution du problème de Cauchy si et seulement si $t_0 \in \circ J$ et si y est une fonction continue sur J qui vérifie

$$y(t) = y_0 + \int_{t_0}^{t} F(s, y(s))\, ds, \quad \forall t \in J. \tag{I.3}$$

Il s'agit donc de résoudre l'équation intégrale (I.3) dans l'espace fonctionnel $\mathcal{C}^0(J, \mathbb{R}^d)$, c'est pourquoi il est naturel que les preuves du théorème utilisent de façon fondamentale les grands théorèmes de l'analyse fonctionnelle : ou bien le théorème d'Ascoli (méthode de compacité) ou bien le théorème du point fixe de Banach.

II.2.1 Le lemme de Gronwall

Un outil central dans tous les problèmes d'équations différentielles est le lemme de Gronwall qui permet de déduire des bornes sur les solutions à partir d'inégalité intégrales qu'elles vérifient.

Lemme I.11 (de Gronwall)

Soit $[a, b[\subset \mathbb{R}$, $C \in \mathbb{R}$ et z, φ deux fonctions continues sur $[a, b[$ à valeurs réelles. On suppose que

– *φ est positive.*

– *L'inégalité suivante est vérifiée*

$$z(t) \leq C + \int_a^t \varphi(s) z(s)\, ds, \quad \forall a \leq t < b.$$

Alors, on a l'estimation

$$z(t) \leq C \exp\left(\int_a^t \varphi(s)\, dt \right), \quad \forall a \leq t < b.$$

Preuve :

On pose

$$h(t) = C + \int_a^t \varphi(s)z(s)\,ds.$$

Comme φ et z sont continues, h est de classe $\mathcal{C}^1$ et on a

$$h'(t) = \varphi(t)z(t) \leq \varphi(t)h(t),$$

car φ est positive. On en déduit que la fonction $e^{-\int_a^t \varphi}h(t)$ est décroissante, ce qui fournit l'inégalité attendue. ■

II.2.2 Unicité

Dans le cas où f est localement Lipschitzienne par rapport à sa seconde variable, on peut démontrer l'unicité d'une éventuelle solution en utilisant le Lemme de Gronwall.

En effet, soient (J_1, y_1), (J_2, y_2) deux solutions du même problème de Cauchy en t_0. On veut montrer que y_1 et y_2 sont égales sur $J_0 = J_1 \cap J_2$. Pour cela on introduit l'ensemble

$$S = \{t \in J_0, \text{ tel que } y_1(s) = y_2(s),\ \forall s \in [t_0, t]\},$$

où $[t_0, t]$ est remplacé par $[t, t_0]$ si $t < t_0$.

Cet ensemble est non vide car il contient t_0 (y_1 et y_2 vérifient la même donnée de Cauchy à l'instant t_0). On va montrer que $S \cap [t_0, +\infty[= J_0 \cap [t_0, +\infty[$ (la même idée montrerait l'égalité de $S \cap]-\infty, t_0]$ et de $J_0 \cap]-\infty, t_0]$).

Supposons que $S \cap [t_0, +\infty[\neq J_0 \cap [t_0, +\infty[$. On pose alors $t^* = \sup(S)$. On a $t^* \geq t_0$ et $t^* \in \overset{\circ}{J}_0$. En effet, si ce n'était pas le cas on aurait $t^* \in \partial J_0$ et alors $y_1 = y_2$ sur $[t_0, \sup J_0[$ et donc $y_1 = y_2$ sur $J_0 \cap [t_0, +\infty[$ par continuité de y_1 et y_2. Ceci contredit l'hypothèse. Par ailleurs, par continuité de y_1 et y_2, on sait que $y_1(t^*) = y_2(t^*) = \tilde{y}$.

Soit L une constante de Lipschitz de f sur le compact $K = [t^*, t^* + 1] \times \bar{B}(\tilde{y}, 1)$. Par continuité, il existe $\delta > 0$ tel que $t^* + \delta \in J_0$ et tel que

$$y_i(t) \in \bar{B}(\tilde{y}, 1),\ \forall t \in [t^*, t^* + \delta], \forall i = 1, 2.$$

Par ailleurs, comme y_1 et y_2 vérifient l'équation on a

$$y_i(t) = y_i(t^*) + \int_{t^*}^t F(s, y_i(s))\,ds.$$

Par soustraction, on trouve

$$|y_1(t) - y_2(t)| \leq \int_{t^*}^t |F(s, y_1(s)) - F(s, y_2(s))|\,ds,\ \ \forall t \in [t^*, t^* + \delta].$$

Comme y_1 et y_2 prennent leur valeurs dans K, on en déduit

$$|y_1(t) - y_2(t)| \leq L \int_{t^*}^t |y_1(s) - y_2(s)|\,ds,\ \ \forall t \in [t^*, t^* + \delta].$$

Le lemme de Gronwall donne alors $y_1(t) = y_2(t)$ pour tout $t \in [t^*, t^* + \delta]$. Ceci montre que $t^* + \delta$ est dans S et contredit donc la définition de t^*.

II.2.3 Existence : la preuve par le théorème du point fixe

Soit $J = [t_0 - \alpha, t_0 + \alpha]$ un intervalle contenant t_0 dans son intérieur. On pose $y^0(t) = y_0$ pour tout $t \in J$ et on construit, par récurrence, la suite de fonctions

$$y^{n+1}(t) = y_0 + \int_{t_0}^t F(s, y^n(s))\,ds,\ \ \forall t \in J.$$

Ceci revient à définir $y^{n+1} = \Phi(y^n)$ où $\Phi : \mathcal{C}^0(J, \mathbb{R}^d) \mapsto \mathcal{C}^0(J, \mathbb{R}^d)$ est l'application qui à y associe

$$\Phi(y)(t) = y_0 + \int_{t_0}^t F(s, y(s))\,ds,\ \ \forall t \in J.$$

Résoudre l'équation (I.3) revient à trouver un point fixe de l'application Φ. Comme J est compact on peut munir $E = \mathcal{C}^0(J, \mathbb{R}^d)$ de la norme infinie, ce qui en fait un espace complet. On peut donc espérer appliquer le théorème du point fixe de Banach à cette fonction. Pour cela, il faudrait montrer que Φ est contractante. Comme on ne possède aucune information globale sur F, il se peut que $\|F(s, y)\|$ soit très grand quand $\|y\|$ est grand et il y a donc aucune chance que nous arrivions à montrer que Φ est contractante sur E.

On va donc essayer d'appliquer le théorème sur le sous-espace fermé $F = \mathcal{C}^0(J, \bar{B}(y_0, R))$ de E (qui est donc bien complet). Pour cela, on peut jouer sur les paramètres α et R pour faire en sorte que $\Phi(F) \subset F$ et que Φ soit contractante.

– Fixons une valeur $\alpha_0 > 0$ et un nombre $R_0 > 0$ tels que le compact $K_0 = [t_0 - \alpha_0, t_0 + \alpha_0] \times \bar{B}(y_0, R_0)$ soit inclus dans l'ouvert U sur lequel (I.2) est vraie.

– On note maintenant $M = \sup_{[t_0-\alpha_0,t_0+\alpha_0]\times\bar{B}(y_0,R_0)} \|F\|$. Ainsi, pour toute fonction $y \in \mathcal{C}^0([t_0 - \alpha_0, t_0 + \alpha_0], \bar{B}(y_0, R_0))$ on a

$$\|\Phi(y)(t) - y_0\| \leq \left\| \int_{t_0}^{t} F(s, y(s))\, ds \right\| \leq |t - t_0| M.$$

Si on veut s'assurer que $\Phi(y)(t)$ reste dans la boule $\bar{B}(y_0, R_0)$, il faut se restreindre à un intervalle $[t_0 - \alpha, t_0 + \alpha]$ avec $0 < \alpha \leq \alpha_0$ choisi pour que

$$\alpha M \leq R_0. \tag{I.4}$$

Ainsi, l'espace $F_\alpha = \mathcal{C}^0([t_0 - \alpha, t_0 + \alpha], \bar{B}(y_0, R_0))$ est laissé fixe par Φ dès que (I.4) est vérifiée.

– Essayons maintenant d'étudier le caractère contractant de Φ sur un tel espace. Soient $y, z \in F_\alpha$, on a

$$\|\Phi(y)(t) - \Phi(z)(t)\| \leq \left| \int_{t_0}^{t} \|F(s, y(s)) - F(s, z(s))\|\, ds \right| \leq C_{t_0,y_0} |t - t_0| \|y - z\|_\infty,$$

et donc

$$\|\Phi(y) - \Phi(z)\|_\infty \leq C_{t_0,y_0} \alpha \|y - z\|_\infty.$$

En conclusion, Φ sera contractante dès que

$$\alpha C_{t_0,y_0} < 1. \tag{I.5}$$

– En conclusion, on va choisir $0 < \alpha \leq \alpha_0$ qui satisfait (I.4) et (I.5), ce qui est bien entendu possible. La fonction Φ laisse alors invariant le sous-espace fermé $F_\alpha \subset E$ et elle est contractante dans cet espace.

D'après le théorème du point fixe de Banach, il existe donc une unique solution $y \in F_\alpha$ à l'équation (I.3) et ainsi $([t_0 - \alpha, t_0 + \alpha], y)$ est une solution du problème de Cauchy considéré. C'est également l'unique solution sur cet intervalle qui prend ses valeurs dans la boule $\bar{B}(y_0, R_0)$.

– Il reste à montrer que toute autre solution éventuelle z du problème de Cauchy définie sur un intervalle de la forme $[t_0 - \beta, t_0 + \beta]$ avec $\beta \leq \alpha$ coincide avec y.

  – Si z prend ses valeurs dans la boule $\bar{B}(y_0, R_0)$, alors la propriété d'unicité dans le théorème du point fixe donne le résultat.

  – Si z ne prend pas ses valeurs dans cette boule, on note $\tilde{\beta}$ le plus grand nombre dans $[0, \beta]$ tel que $z([t_0 - \beta, t_0 + \beta])$ est contenu dans cette boule. On a $\tilde{\beta} < \beta$ par hypothèse et $\tilde{\beta} > 0$ car $z(t_0) = y_0$ est dans l'intérieur de la boule et que z est continue.

  On a alors

$$\|z(t) - y_0\| \leq |t - t_0| M \leq \tilde{\beta} M < \beta M \leq \alpha M \leq R_0, \ \ \forall t \in [t_0 - \tilde{\beta}, t_0 + \tilde{\beta}],$$

  ce qui contredit la maximalité de $\tilde{\beta}$.

Remarque I.12

La méthode de point fixe ne permet pas réellement, en général, le calcul effectif des solutions (ou d'une approximation) des équations différentielles.
Calculer, à titre d'exemple, les approximations successives par la méthode de Picard appliquée à la résolution du problème de Cauchy $y' = y$ et $y(0) = 1$.

II.2.4 Existence : la preuve *via* la méthode d'Euler

Pour montrer l'existence d'une solution sous la seule hypothèse que F est continue, on va prouver que l'approximation obtenue par la méthode d'Euler converge. On pourra se référer, par exemple, à [**?**, page 133] bien que la preuve ci-dessous soit rédigée un peu différemment.

Soit M une borne de f sur le compact $[t_0, t_0 + 1] \times \bar{B}(y_0, 1)$. On pose maintenant $T = \min(1, 1/M)$.

On fixe un nombre $N > 0$, on pose $\Delta t = T/N$, $t^n = t_0 + n\Delta t$, et on construit l'approximation d'Euler comme suit

$$\begin{cases} y^0 = y_0, \\ y^{n+1} = y^n + \Delta t F(t^n, y^n), \ \ \forall n \in \{0, ..., N-1\}. \end{cases}$$

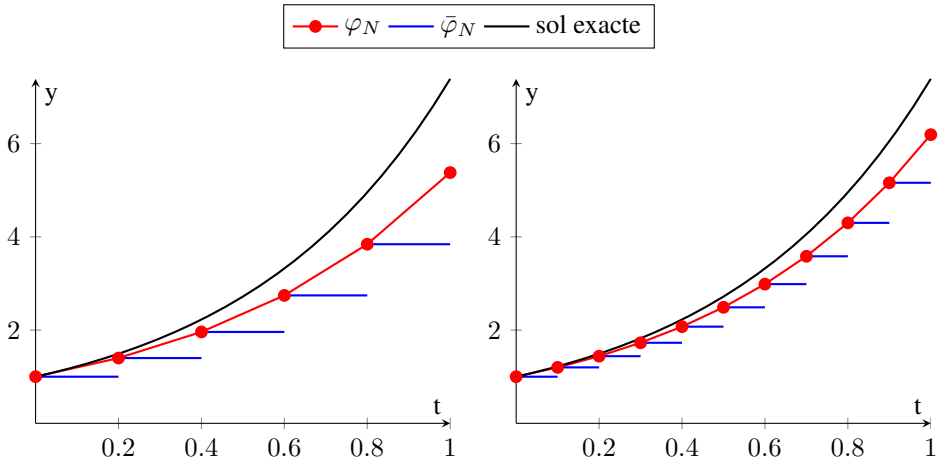
– On vérifie aisément par récurrence que $y^n \in \bar{B}(y_0, 1)$ pour tout $n \in \{0, ..., N\}$.

– A l'aide de cette suite, on construit l'unique fonction continue affine par morceaux φ_N vérifiant (voir Figure I.5)

$$\varphi_N(t^n) = y^n, \ \ \forall n \in \{0, ..., N\},$$

et l'unique fonction constante par morceaux $\bar{\varphi}_N$ définie par

$$\bar{\varphi}_N(t) = y^n, \ \ \forall t \in [t^n, t^{n+1}[.$$



(a) $N = 5$ (b) $N = 10$

FIGURE I.5 – Illustration de la méthode d'Euler explicite pour l'équation $y' = 2y$

On voit que φ_N est Lipschitzienne sur $[t_0, t_0 + T]$ et que $\mathrm{Lip}(\varphi_N) \leq M$. De plus, les fonctions φ_N sont uniformément bornées sur $[t_0, t_0 + T]$.
Par ailleurs, par construction, nous avons (en regardant ce qui se passe sur chaque intervalle de longueur Δt)

$$\|\varphi_N - \bar{\varphi}_N\|_\infty \leq M\Delta t = \frac{MT}{N} \xrightarrow[N\to+\infty]{} 0. \tag{I.6}$$

– La suite de fonctions $(\varphi_N)_N$ est donc bornée dans $\mathcal{C}^0([0,T], \mathbb{R}^d)$ et également équiuniformément continue. On peut donc appliquer le théorème d'Ascoli et obtenir l'existence d'une sous-suite $(\varphi_{N_k})_k$ qui converge uniformément vers une fonction continue $\varphi : [0,T] \times \mathbb{R}^d$.
D'après (I.6), on a également la convergence uniforme de la suite $(\bar{\varphi}_{N_k})_k$ vers **la même limite** φ.
Constatons maintenant que, par contruction, φ_{N_k} et $\bar{\varphi}_{N_k}$ vérifient

$$\varphi_{N_k}(t) = y_0 + \int_{t_0}^{t} F(\bar{\varphi}_{N_k}(s))\, ds, \ \ \forall t \in [t_0, t_0 + T]. \tag{I.7}$$

On va chercher à passer à la limite dans (I.7) et ainsi prouver que la limite $\varphi \in \mathcal{C}^0([0,T], \mathbb{R}^d)$ est bien solution de l'équation recherchée.
On note ω le module d'uniforme continuité de F sur le compact $[t_0, t_0 + 1] \times \bar{B}(y_0, 1)$. On a donc pour tout $s \in [t_0, t_0 + T]$

$$\|F(\bar{\varphi}_{N_k}(s)) - F(\varphi(s))\| \leq \omega(\|\bar{\varphi}_{N_k}(s) - \varphi(s)\|) \leq \omega(\|\bar{\varphi}_{N_k} - \varphi\|_\infty),$$

et donc

$$\|F \circ \bar{\varphi}_{N_k} - F \circ \varphi\|_\infty \leq \omega(\|\bar{\varphi}_{N_k} - \varphi\|_\infty) \xrightarrow[k\to+\infty]{} 0,$$

ce qui prouve que $F \circ \bar{\varphi}_{N_k}$ converge uniformément vers $F \circ \varphi$. On peut donc, à bon droit passer à la limite dans (I.7) ce qui montre φ vérifie

$$\varphi(t) = y_0 + \int_{t_0}^{t} F(s, \varphi(s))\, ds, \ \ \forall t \in [t_0, t_0 + T],$$

et donc elle est bien solution du problème de Cauchy souhaité.

Au passage, on a donc démontré le

Théorème I.13 (Cauchy-Arzela-Peano)

Soit $F : I \times \mathbb{R}^d \to \mathbb{R}^d$ est continue. Alors pour tout couple $(t_0, y_0) \in I \times \mathbb{R}^d$, il existe **au moins une** *solution du problème de Cauchy associé.*

On a également (presque) démontré le théorème suivant qui donne la convergence de la méthode d'Euler

Théorème I.14

Sous les hypothèses du théorème de Cauchy-Lipschitz, si $y : [t_0, t_0 + T] \to \mathbb{R}^d$ est une solution du problème de Cauchy considéré sur un temps T assez petit et $(y^n)_n$ la suite des itérées obtenues par la méthode d'Euler explicite associée à un pas de temps Δt, nous avons

$$\sup_{0 \leq n \leq T/\Delta t} \|y(t^n) - y^n\| \xrightarrow[\Delta t \to 0]{} 0. \tag{I.8}$$

Nous verrons dans le chapitre précédent que, sous certaines hypothèses supplémentaires, on peut estimer la taille de l'erreur comise entre la solution exacte et la solution approchée.

Preuve :

Rappelons que Δt est relié à N par la formule $\Delta t = T/N$ et que par ailleurs, comme $\varphi_N(t^n) = y^n$, on a

$$\sup_{0 \leq n \leq T/\Delta t} \|y(t^n) - y^n\| \leq \|\varphi_N - y\|_\infty. \tag{I.9}$$

On a vu plus haut que $y = \varphi$, la limite uniforme de la sous-suite $(\varphi_{N_k})_k$. Le résultat sera donc prouvé si on montre que toute la suite $(\varphi_N)_N$ converge vers $\varphi = y$.

Il s'agit d'un raisonnement classique de compacité/unicité qui découle du lemme I.15 que l'on prouvera après.

Lemme I.15

Soit A un espace compact (disons dans un espace métrique pour fixer les idées) et $(x_n)_n$ une suite de points de A. On a alors l'équivalence

$$(x_n)_n \text{ converge dans } A \Leftrightarrow (x_n)_n \text{ possède une unique valeur d'adhérence dans } A.$$

Appliquons ce lemme à la suite $(\varphi_N)_N$, qui est bien contenue dans un compact de $\mathcal{C}^0([t_0, t_0 + T], \mathbb{R}^d)$ d'après le théorème d'Ascoli. On a vu plus haut que toute valeur d'adhérence φ de cette suite est nécessairement solution du problème de Cauchy que nous sommes en train d'étudier. Comme nous avons supposé que le théorème de Cauchy-Lipschitz s'applique, une telle solution est unique (et notée y dans l'énoncé). Il n'y a donc bien qu'une seule valeur d'adhérence de cette suite et le lemme nous donne la convergence uniforme de toute la suite $(\varphi_N)_N$ vers y.

On conclut en utilisant (I.9). ■

Il reste à démontrer le lemme.

Preuve (du Lemme I.15):

L'implication $\Rightarrow$ est immédiate, il nous suffit de montrer l'autre implication. Supposons donc que $(x_n)_n$ a une unique valeur d'adhérence dans A que l'on note x^* et raisonnons pas l'absurde en supposant que $(x_n)_n$ ne converge pas vers x^*.

Cela signifie qu'il existe $\varepsilon > 0$ et une sous-suite $(x_{\varphi(n)})_n$ telle que

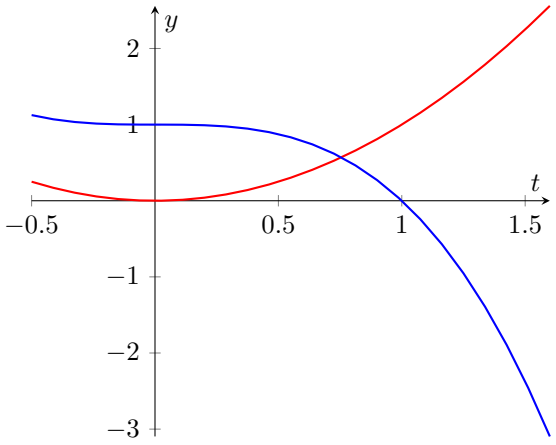
$$d(x^*, x_{\varphi(n)}) \geq \varepsilon, \quad \forall n \geq 0. \tag{I.10}$$

Or, comme A est compact, on peut extraire de $(x_{\varphi(n)})_n$ une sous-suite convergente $(x_{\varphi(\psi(n))})_n$ dont la limite, par hypothèse, ne peut être que x^*. Ceci contredit (IV.8). ■

II.3 Premières conséquences du théorème de Cauchy-Lipschitz

La première conséquence du théorème est bien sûr l'existence et l'unicité d'une solution à un problème de Cauchy qui peut modéliser une situation physique donnée. En réalité, les conséquences du théorème sont bien plus nombreuses. Commençons par quelques exemples utiles.

- **Une solution maximale est forcément définie sur un intervalle ouvert.**
- **Deux trajectoires disctinctes ne peuvent pas se couper :** Soient (J, y_1) et (J, y_2) deux solutions de l'EDO $y' = F(t, y)$ définies sur le même intervalle.



S'il existe $t_0 \in J$ tel que $y_1(t_0) = y_2(t_0)$ alors $y_1 \equiv y_2$.

– **Dans $\mathbb{R}$ les trajectoires sont ordonnées :** On suppose ici que $d = 1$. Avec les mêmes notations que précédemment : S'il existe $t_0 \in J$ tel que $y_1(t_0) < y_2(t_0)$ alors $y_1(t) < y_2(t)$ pour tout $t \in J$. Il suffit de raisonner par l'absurde et d'utiliser le théorème des valeurs intermédiaires pour se ramener à la propriété précédente.

– **Application à la résolution d'une EDO simple :** Par exemple pour résoudre le problème suivant

$$y' = y^2.$$

On dit que, d'après le théorème de Cauchy-Lipschitz, si la solution n'est pas identiquement nulle, elle ne s'annulle jamais ! On peut donc diviser l'équation par y^2 puis intégrer l'équation de part et d'autres afin de la résoudre. L'ensemble de toutes les solutions est constitué de

$$(\mathbb{R}, 0),$$

$$(J_K^-, y_K^-) \text{ avec } J_K^- =]-\infty, K[, \text{ et } y_K^-(t) = \frac{1}{K-t}, \forall t \in J_K^-,$$

$$(J_K^+, y_K^+) \text{ avec } J_K^+ =]K, +\infty[, \text{ et } y_K^+(t) = \frac{1}{K-t}, \forall t \in J_K^+,$$

pour $K \in \mathbb{R}$. Ces dernières solutions ne sont pas globales (i.e. ne sont pas définies sur $\mathbb{R}$ tout entier).

– **Trajectoires périodiques :** On suppose que $I = \mathbb{R}$ et que F est une fonction T-périodique. Alors une solution $(\mathbb{R}, y)$ de l'EDO est T-périodique si et seulement s'il existe un $t_0 \in \mathbb{R}$ tel que

$$y(t_0 + T) = y(t_0).$$

En effet, si ceci est vrai, la fonction $z(t) = y(t+T)$ vérifie la même équation différentielle que y et la même donnée de Cauchy $z(t_0) = y(t_0 + T) = y(t_0)$. Par unicité dans le théorème de Cauchy-Lipschitz, on sait que ces deux solutions sont donc identiques, ce qui prouve le résultat.

II.4 Explosion en temps fini

Comme on l'a vu dans les exemples ci-dessus, il arrive que l'intervalle de définition de la solution maximale d'un problème de Cauchy soit strictement inclus dans I. On va voir pourquoi ce phénomène est appelé **explosion en temps fini**.

Théorème I.16

On suppose F continue et localement Lipschitzienne par rapport à sa variable d'état. Soit (J, y) une solution **maximale** *de* (I.1). *On note $J =]\alpha, \beta[$.*

– *Si $\alpha \in I$, alors*

$$\limsup_{t \to \alpha^+} \|y(t)\| = +\infty. \tag{I.11}$$

– *Si $\beta \in I$, alors*

$$\limsup_{t \to \beta^-} \|y(t)\| = +\infty. \tag{I.12}$$

Preuve :

Supposons que $\alpha \in I$, et que (I.11) soit fausse. Ceci implique que y est bornée au voisinage de α. Comme par ailleurs, y vérifie l'EDO et que F est continue, on constate que y' est aussi bornée au voisinage de α. Par le théorème des accroissements finis et le critère de Cauchy, cela implique que la limite de $y(t)$ quand $t \to \alpha$ existe. On la note y_α.

Le couple (α, y_α) est une donnée de Cauchy admissible pour le problème considéré. D'après le théorème de Cauchy-Lipschitz, il existe une unique solution maximale (J_α, z) de ce problème de Cauchy. Rappelons que J_α est un ouvert.

On note alors $\tilde{J} = J \cup J_\alpha \cap]-\infty, \beta[$ et on définit sur $\tilde{J}$ la fonction suivante

$$\tilde{y}(t) = \begin{cases} y(t), & \text{si } t > \alpha \\ z(t), & \text{si } t \leq \alpha. \end{cases}.$$

On vérifie que $\tilde{y}$ est une fonction dérivable sur $\tilde{J}$ et qu'elle vérifie bien l'équation différentielle (I.1). Ceci contredit la maximalité de la solution (J, y) et montre le théorème. ■

Ce théorème (ou plus exactement sa contraposée) est notamment utile pour démontrer qu'une solution maximale est nécessairement globale.

Exemple I.17

Si la fonction F est bornée sur $I \times \mathbb{R}^d$, toutes les solutions maximales sont globales. En effet, appelons $M > 0$ une borne de $\|F\|$ sur $I \times \mathbb{R}^d$, et (J, y) une solution maximale du problème. On prend $a \in J$ quelconque de sorte que

$$\forall t \in J, t > a,\ \|y(t)\| = \left\| y(a) + \int_a^t F(s, y(s))\, ds \right\| \leq \|y(a)\| + M(t-a),$$

ce qui montre que si la borne supérieure de J est finie, la solution y est bornée au voisinage de $\sup J$ et donc $\sup J \notin I$, d'après le théorème d'explosion en temps fini. La démarche est identique pour montrer que $\inf J \notin I$ et donc $J = I$.

Théorème I.18 (Théorème de Cauchy-Lipschitz global)

Si la fonction F est continue et globalement Lipschitzienne par rapport à y, alors toutes les solutions maximales sont globales. Plus précisément, s'il existe $k : I \mapsto \mathbb{R}$ continue et positive telle que

$$\forall t \in I, \forall y_1, y_2 \in \mathbb{R}^d,\ \ \|F(t, y_1) - F(t, y_2)\| \leq k(t)\|y_1 - y_2\|,$$

alors toutes les solutions maximales sont globales.

Preuve :

Si (J, y) est une solution maximale et $a \in J$, on a

$$y(t) = y(a) + \int_a^t F(s, y(s))\, ds.$$

On prend la norme, pour $t > a$ et on utilise l'hypothèse

$$\|y(t)\| \leq \|y(a)\| + \int_a^t (k(s)\|y(s)\| + \|F(s, 0)\|)\, ds.$$

Supposons que la limite supérieure β de J soit finie et dans I, on a alors

$$\|y(t)\| \leq \left(\|y(a)\| + \int_a^\beta \|F(s, 0)\|\, ds \right) + \int_a^t k(s)\|y(s)\|\, ds,\ \ \forall a \leq t < \beta$$

D'après le lemme de Gronwall (Lemme III.5), on en déduit que

$$\|y(t)\| \leq \left(\|y(a)\| + \int_a^\beta \|F(s, 0)\|\, ds \right) \exp\left(\int_a^\beta k(s)\, ds \right),\ \ \forall a \leq t < \beta.$$

Ceci montre que la fonction y est bornée au voisinage de la borne β, ce qui contredit le théorème d'explosion en temps fini, donc $\beta \notin I$. ■

II.5 Sortie de tout compact

Toute la théorie de Cauchy-Lipchitz (et de Cauchy-Arzela-Peano) développée précédemment s'étend au cas où F est seulement définie sur un ensemble de la forme $I \times \Omega$ où Ω est un ouvert de $\mathbb{R}^d$.

Dans ce cas, le théorème d'explosion en temps fini prend la forme suivante

Théorème I.19 (sortie de tout compact)

Soit $F : I \times \Omega$ une fonction continue et localement Lipschitzienne par rapport à la variable d'état.
Soit (J, y) une solution maximale de l'équation différentielle, on a alors :

- *Si $\beta = \sup J \in I$, alors y sort de tout compact de Ω au voisinage de β, c'est-à-dire que, pour tout compact $K \subset \Omega$, il existe $\varepsilon > 0$ tel que*

$$y(t) \notin K, \ \ \forall t \in]\beta - \varepsilon, \beta[.$$

- *Si $\alpha = \inf J \in I$, alors y sort de tout compact de Ω au voisinage de α, c'est-à-dire que, pour tout compact $K \subset \Omega$, il existe $\varepsilon > 0$ tel que*

$$y(t) \notin K, \ \ \forall t \in]\alpha, \alpha + \varepsilon[.$$

Exemple I.20

On considère l'équation différentielle $x' = -1/x$, c'est-à-dire $F(t, x) = -1/x$, pour $(t, x) \in \mathbb{R} \times]0, +\infty[$, donc $I = \mathbb{R}$ et $\Omega =]0, +\infty[$.
Soit $x_0 \in \Omega$, on veut résoudre le problème de Cauchy pour la donnée $x(0) = x_0$. Pour cela on multiplie par x et on intègre, ce qui donne

$$-xx' = 1,$$

$$-\frac{x^2}{2} + \frac{x_0^2}{2} = t,$$

d'où

$$x(t) = \sqrt{x_0^2 - 2t}, \ \ \text{pour tout } t \in]-\infty, x_0^2/2[.$$

On voit donc que la solution maximale n'est pas globale. Au voisinage du temps d'existence on a

$$\lim_{t \to \frac{x_0^2}{2}} x(t) = 0,$$

et comme $0 \notin \Omega$, cela illustre bien le fait que la solution sort de tout compact de Ω.

II.6 Preuve directe du théorème de Cauchy-Lipschitz global

La preuve donnée plus haut du théorème de Cauchy-Lipschitz global I.18 s'appuie sur le théorème de Cauchy-Lipschitz local et le théorème d'explosion en temps fini. On peut en réalité donner une preuve directe de ce résultat par une méthode de point fixe. Cette preuve est, dans l'esprit, sensiblement la même que celle du théorème local, c'est pourquoi il est instructif de la connaître.

Preuve (du théorème I.18):

On se donne une donnée de Cauchy $(a, y_a) \in \mathbb{R} \times \mathbb{R}^d$ et on veut montrer qu'il existe une unique solution globale au problème. Pour cela, on va démontrer qu'il existe une unique solution définie sur l'intervalle $[a, +\infty[$, le cas de l'intervalle $]-\infty, a]$ se traitant de manière analogue (ou en changeant F en $-F$).

D'après le lemme de Gronwall et les calculs menés plus haut, on voit que **toute éventuelle solution** doit vérifier l'estimation

$$\forall t \in [a, +\infty[, \ \ \|y(t)\| \leq \left(\|y_a\| + \int_a^t \|F(s, 0)\| \, ds \right) \exp\left(\int_a^t k(s) \, ds \right). \tag{I.13}$$

Pour simplifier les écritures, on introduit les fonctions

$$\varphi(t) = \|y_a\| + \int_a^t \|F(s, 0)\| \, ds,$$

et

$$\psi(t) = \int_a^t k(s) \, ds,$$

de sorte que l'estimation (I.13) devient

$$\forall t \in [a, +\infty[, \ \ \|y(t)\| \leq \varphi(t) e^{\psi(t)},$$

celle-ci étant valable *a priori* pour toute solution du problème.

On introduit maintenant l'espace fonctionnel

$$E = \left\{ y \in \mathcal{C}^0([a, +\infty[, \mathbb{R}^d), \ \sup_{t \geq a} \left(\varphi(t)^{-1} e^{-2\psi(t)} \|y(t)\| \right) < +\infty \right\},$$

que l'on munit de la norme $\|y\|_E = \sup_{t \geq a} \left(\varphi(t)^{-1} e^{-2\psi(t)} \|y(t)\| \right)$. On vérifie aisément [1] que cet espace est complet, c'est un espace de Banach. **Remarque :** la présence du facteur 2 dans le terme exponentiel est cruciale [2].

On introduit maintenant l'opérateur $T : E \mapsto E$ défini par

$$\forall y \in E, \forall t \in [a, +\infty[, \ \ Ty(t) = y_a + \int_a^t F(s, y(s)) \, ds.$$

Il est clair maintenant qu'une fonction continue y est solution de notre problème de Cauchy si et seulement si $y \in E$ (estimation *a priori* ci-dessus) et si $Ty = y$ (forme intégrale de l'équation différentielle).

On est donc ramenés à trouver un point fixe de l'opérateur T. Dans un espace de Banach, on dispose du théorème de point fixe de Banach (ou de Picard ...) qui nous dit que si T est contractant alors il existe un unique point fixe. Le résultat souhaité sera donc démontré si on établit que T est contractant.

Soient donc $y, z \in E$, on veut estimer $\|Ty - Tz\|_E$ en fonction de $\|y - z\|_E$. Pour cela, on fixe un instant $t \in [a, +\infty[$ et on majore $\|Ty(t) - Tz(t)\|$ de la façon suivante

$$\begin{aligned} \|Ty(t) - Tz(t)\| &= \left\| \int_a^t F(s, y(s)) - F(s, z(s)) \, dx \right\| \\ &\leq \int_a^t k(s) \|y(s) - z(s)\| \, ds \\ &\leq \|y - z\|_E \int_a^t k(s) \varphi(s) e^{2\psi(s)} \, ds. \end{aligned}$$

On remarque maintenant que $k = \psi'$, de sorte qu'on peut intégrer par parties l'intégrale du membre de droite

$$\int_a^t k(s) \varphi(s) e^{2\psi(s)} \, ds = \int_a^t \varphi(s) \left(\psi'(s) e^{2\psi(s)} \right) \, ds = \frac{1}{2} \varphi(t) e^{2\psi(t)} - \frac{1}{2} \varphi(a) e^{2\psi(a)} - \frac{1}{2} \int_a^t \varphi'(s) e^{2\psi(s)} \, ds.$$

Comme $\varphi'(s) = \|F(s, 0)\| \geq 0$ et $\varphi \geq 0$, les deuxième et troisième termes de cette égalité sont négatifs. *In fine*, on a donc montré

$$\|Ty(t) - Tz(t)\| \leq \|y - z\|_E \frac{1}{2} \varphi(t) e^{2\psi(t)}.$$

Ceci étant vrai pour tout $t \in [a, +\infty[$, on a

$$\|Ty - Tz\|_E = \sup_{t \geq a} \left(\|Ty(t) - Tz(t)\| \varphi(t)^{-1} e^{-2\psi(t)} \right) \leq \frac{1}{2} \|y - z\|_E,$$

ce qui montre que T est $1/2$-Lipschitzienne et donc contractante, d'où le résultat. ■

II.7 Dépendance par rapports aux données

Dans tout ce paragraphe, je suppose que $F : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$ est continue, localement Lipschitzienne par rapport à la variable d'état, et que toutes les solutions maximales de l'équation différentielles $y' = F(t, y)$ sont globales (par exemple si F est globalement Lipschitzienne par rapport à y).

Définition I.21 (Flot associé à l'équation différentielle)

Pour tout $(t_0, y_0) \in \mathbb{R} \times \mathbb{R}^d$ et pour tout $t \in \mathbb{R}$, on note $\varphi(t, t_0, y_0) \in \mathbb{R}^d$ la valeur, à l'instant t, de l'unique solution du problème de Cauchy associé à la donnée (t_0, y_0) et à l'ED considérée.
L'application $\varphi : \mathbb{R} \times \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$ est appellée **le flot** *associé à l'équation différentielle.*

1. c'est un bon exercice !
2. voyez-vous à quel endroit ceci intervient dans la suite ?

Théorème I.22

1. *Pour tout* $t \in \mathbb{R}$, $\varphi(t,t,.) = \mathrm{Id}$.
2. *Pour tout* $t, s, u \in \mathbb{R}$,
$$\varphi(t,s,\varphi(s,u,.)) = \varphi(t,u,.).$$
En particulier, pour tous t, s *l'application* $\varphi(t,s,.)$ *est une bijection de* $\mathbb{R}^d$ *sur lui-même.*
3. *L'application* φ *est continue sur* $\mathbb{R}\times\mathbb{R}\times\mathbb{R}^d$.

Preuve.
La première propriété est triviale. Pour la seconde, on fixe $y_0 \in \mathbb{R}^d$ et on observe que les applications
$$t \mapsto \varphi(t,s,\varphi(s,u,y_0)), \text{ et } t \mapsto \varphi(t,u,y_0),$$
sont deux solutions du même problème de Cauchy (à l'instant $t = s$) donc sont égales.

Montrons la continuité de l'application φ :

– On montre d'abord que pour tout compact $K \subset \mathbb{R}^d$ et tout $t_0 \in \mathbb{R}$, il existe un compact K_1, un $C > 0$, un $\delta > 0$ tels que
$$\varphi(t,t_0,y) \in K_1, \ \ \forall y \in K, \forall |t-t_0| \le \delta,$$
et
$$\|\varphi(t,t_0,y) - y\| \le C|t-t_0|, \ \ \forall y \in K, \forall |t-t_0| \le \delta.$$
Pour cela on note $K_1 = K + \bar{B}(0,1)$ et on pose $M = \sup_{[t_0-1,t_0+1]\times K_1} \|F\|$. On voit que si $\delta = \min(1, 1/K_1)$ alors pour tout $t \in [t_0-\delta, t_0+\delta]$, pour tout $y \in K$, on a $\varphi(t,t_0,y) \in K_1$. En effet, par un raisonnement par l'absurde maintenant usuel si cette propriété n'est pas vraie, on prend le premier instant $t^* \in [t_0, t_0+\delta[$ pour lequel $\varphi(t,t_0,y)$ cesse d'appartenir à K_1 et on écrit
$$\|\varphi(t^*,t_0,y) - y\| = \left\| \int_{t_0}^{t^*} F(s,\varphi(s,t_0,y))\, ds \right\| \le (t^*-t_0)M < \delta M \le 1,$$
ce qui contredit la définition de t^*. On a donc bien, *in fine* la propriété
$$\|\varphi(t,t_0,y) - y\| \le |t-t_0|M,$$
pour tous les $t \in [t_0-\delta, t_0+\delta]$ et tous les $y \in K$.

– On montre maintenant que, t_0 et s_0 étant fixés, l'application $y \in \mathbb{R}^d \mapsto \varphi(t_0,s_0,y)$ est continue.
Il s'agit d'une application du Lemme de Gronwall. Soient $y, z \in \mathbb{R}^d$, on écrit pour tout $t \in [s_0,t_0]$,
$$\varphi(t,s_0,y) = y + \int_{s_0}^{t} F(\tau,\varphi(\tau,s_0,y))\, d\tau,$$
$$\varphi(t,s_0,z) = z + \int_{s_0}^{t} F(\tau,\varphi(\tau,s_0,z))\, d\tau.$$
L'application $\tau \mapsto \varphi(\tau,s_0,y)$ est continue sur $[s_0,t_0]$, elle prend donc ses valeurs dans un compact K. On note $K_1 = K + \bar{B}(0,1)$ et on appelle L la constante de Lipschitz de F par rapport à sa seconde variable sur le compact $[s_0,t_0]\times K_1$.
Ainsi, pour tout $t \in [s_0,t_0]$ tel que $\varphi(\tau,s_0,z) \in K_1$ pour tout $\tau \in [s_0,t]$, on a
$$\begin{aligned}\|\varphi(t,s_0,y) - \varphi(t,s_0,z)\| &\le \|y-z\| + \int_{s_0}^{t} \|F(\tau,\varphi(\tau,s_0,y) - F(\tau,\varphi(\tau,s_0,z))\|\, d\tau \\ &\le \|y-z\| + L\int_{s_0}^{t} \|\varphi(\tau,s_0,y) - \varphi(\tau,s_0,z)\|\, d\tau,\end{aligned}$$
et donc par le lemme de Gronwall, pour de tels t, on a
$$\|\varphi(t,s_0,y) - \varphi(t,s_0,z)\| \le \|y-z\|e^{L|t-s_0|} \le \|y-z\|e^{L|t_0-s_0|}.$$
On voit en particulier que si $\|y-z\| < \frac{1}{e^{L|t_0-s_0|}}$, alors $\varphi(t,s_0,z)$ reste dans le compact K_1 pour tout $t \in [s_0,t_0]$ et donc la formule ci-dessus est valable pour tout $t \in [s_0,t_0]$. En particulier, on trouve
$$\|\varphi(t_0,s_0,y) - \varphi(t_0,s_0,z)\| \le \|y-z\|e^{L|t_0-s_0|}.$$
Ceci montre que $y \mapsto \varphi(t_0,s_0,y)$ est localement Lipschitzienne donc continue.

– Il reste à assembler les morceaux. On écrit

$$\varphi(t,s,y) = \varphi(t,t_0,\varphi(t_0,s_0,\varphi(s_0,s,y))).$$

Pour y dans un voisinage compact K de y_0, et s,t tels que $|s-t| \leq \delta, |s_0 - s| \leq \delta$, on sait que $\varphi(s_0,s,y)$ est dans un compact K_1, et donc que $\varphi(t_0,s_0,\varphi(s_0,s,y))$ est dans un compact K_2 (car $\varphi(t_0,s_0,.)$ est continue). Pour t proche de t_0, on a donc

$$\|\varphi(t,s,y) - \varphi(t_0,s_0,\varphi(s_0,s,y))\| \leq M_1|t-t_0|,$$

puis

$$\|\varphi(t,s,y) - \varphi(t_0,s_0,y)\| \leq M_1|t-t_0| + M_2\|\varphi(s_0,s,y) - y\| \leq M_1|t-t_0| + M_2'|s-s_0|,$$

et enfin

$$\|\varphi(t,s,y) - \varphi(t_0,s_0,y_0)\| \leq M_1|t-t_0| + M_2'|s-s_0| + M_3\|y-y_0\|.$$

■

Sous des hypothèses un peu plus fortes sur F, on peut montrer la différentiabilité du flot.

Théorème I.23 (voir par exemple [?, Th. 5.13, p 148],[?, page 302])

Si la fonction F admet des dérivées partielles par rapport aux y_i continues (par rapport à toutes les variables), alors le flot φ est de classe $\mathcal{C}^1$ et sa différentielle par rapport à y est l'unique solution du problème de Cauchy **linéaire** *suivant (dont l'inconnue est $t \mapsto M(t,s) \in M_d(\mathbb{R})$)*

$$\begin{cases} \dfrac{d}{dt}M(t,s) = \big(d_yF(t,\varphi(t,s,y))\big).M(t,s) \\ M(s,s) = \mathrm{Id}. \end{cases}$$

La démonstration du théorème est assez technique mais il est aisé de retrouver l'équation différentielle vérifiée par la dérivée du flot. Il suffit, formellement, de dériver l'équation de départ par rapport à la variable y en supposant que l'interversion des dérivées est licite dans tous les termes.

III La théorie des équations différentielles linéaires

Une EDO est dite linéaire (homogène) si elle est de la forme

$$y' = A(t)y, \tag{I.14}$$

où $A : I \mapsto M_d(\mathbb{R})$ est une application à valeurs matricielles. Par abus de language, on parle aussi d'équation linéaire non-homogène (ou avec second membre) pour les équations de la forme

$$y' = A(t)y + b(t), \tag{I.15}$$

où $b : I \mapsto \mathbb{R}^d$ est une fonction continue donnée.

III.1 Le cas général

Théorème I.24

Si A et b sont des applications continues, alors le théorème de Cauchy-Lipschitz global s'applique à l'équation (I.15)*. Il y a donc existence et unicité de la solution maximale globale pour tout problème de Cauchy.*
L'ensemble de toutes les solutions de (I.15) *forme un sous-espace affine (vectoriel si $b \equiv 0$) de l'ensemble des fonctions dérivables de I dans $\mathbb{R}^d$ de dimension* **exactement** *d.*
Si $y_1, \ldots, y_d$ sont d solutions indépendantes de l'équation (I.14) *et z une solution quelconque de l'équation* (I.15)*, alors toute solution de* (I.15) *s'écrit* **de manière unique** *sous la forme*

$$y(t) = \sum_{i=1}^{d} \alpha_i y_i(t) + z(t). \tag{I.16}$$

Remarque I.25

Soit $t_0 \in I$. On a l'équivalence suivante

$$\textit{les } (y_i)_i \textit{ sont indépendants dans } \mathcal{C}^0(I,\mathbb{R}^d) \iff \textit{les } (y_i(t_0))_i \textit{ sont indépendants dans } \mathbb{R}^d.$$

Il y a un sens évident et l'autre sens provient de la propriété d'unicité dans Cauchy-Lipschitz.

Preuve :

- Vérifions que la fonction $F(t,y) = A(t)y + b(t)$ est globalement Lipschitzienne. On mettra sur l'ensemble des matrices la norme induite par celle choisie sur l'espace $\mathbb{R}^d$. On a bien alors
$$\|F(t,y_1) - F(t,y_2)\| \leq \|A(t)\|\|y_1 - y_2\|, \ \ \forall y_1, y_2 \in \mathbb{R}^d,$$
avec $t \mapsto \|A(t)\|$ qui est continue.
- On vérifie tout d'abord que toutes les fonctions de la forme (I.16) sont bien solutions du problème. On fixe ensuite $t_0 \in I$. D'après la remarque précédente, les vecteur $(y_i(t_0))_i$ forment une base de $\mathbb{R}^d$ donc, pour tout $\tilde{y} \in \mathbb{R}^d$, il existe une unique famille de réels $(\alpha_i)_i$ tels que
$$\tilde{y} = \sum_{i=1}^{n} \alpha_i y_i(t_0) + z(t_0).$$
Cette famille de réels, fournit bien l'unique solution pour la donnée de Cauchy $(t_0, \tilde{y})$.

■

Malgré le caractère assez simple de ces équations, il n'y a pas de méthode générale de résolution disponible. Par exemple les solutions de l'équation
$$\begin{pmatrix} x \\ y \end{pmatrix}' = \begin{pmatrix} 0 & 1 \\ t & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix},$$
sont des fonctions très complexes qu'on ne peut exprimer simplement à l'aide de fonctions usuelles (voir l'exercice **??**).

Dans le cas scalaire, i.e. avec $d = 1$, on a une formule générale pour l'équation sans second membre
$$y(t) = \exp\left(\int_{t_0}^{t} a(s)\,ds\right) y(t_0).$$
On verra des formules similaires pour l'équation avec second membre plus tard.

III.2 Le cas à coefficients constants, sans second membre

On s'intéresse ici au cas où $t \mapsto A(t)$ ne dépend pas du temps et où le second membre est nul. L'équation considérée est alors
$$y' = Ay. \tag{I.17}$$

Théorème I.26

La solution du problème de Cauchy associé à (I.17) *pour la donnée $(t_0, y_0) \in \mathbb{R} \times \mathbb{R}^d$ est définie sur $\mathbb{R}$ tout entier par la formule*
$$y(t) = e^{(t-t_0)A} y_0.$$

Dans cette formule, on a eu besoin de l'exponentielle de matrices définie par
$$e^M = \sum_{k=0}^{\infty} \frac{M^k}{k!}. \tag{I.18}$$

Preuve.

Considérons l'application $\varphi : t \in \mathbb{R} \mapsto e^{tA} \in M_d(\mathbb{R})$. On a pour tout $t \in \mathbb{R}$ et tout h petit
$$\varphi(t+h) = \sum_{k=0}^{\infty} \frac{(t+h)^k M^k}{k!} = \sum_{k=0}^{\infty} \frac{t^k M^k}{k!} + \sum_{k=0}^{\infty} \frac{k t^{k-1} h M^k}{k!} + O(h^2),$$
Ainsi, on a
$$\lim_{h\to 0} \frac{1}{h}(\varphi(t+h) - \varphi(t)) = \sum_{k=0}^{\infty} \frac{k t^{k-1} M^k}{k!} = \sum_{k=1}^{\infty} \frac{t^{k-1} M^k}{(k-1)!} = M e^{tM} = e^{tM} M.$$
D'où le résultat. ■

On rappelle qu'en général, la formule (I.18) n'est pas pratique pour calculer l'exponentielle et qu'il plus raisonnable d'utiliser des réductions pour ramener le calcul à un cas simple en utilisant le fait que

$$e^{P^{-1}MP} = P^{-1}e^{M}P.$$

– Typiquement, si M est diagonalisable et $P^{-1}MP = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}$ alors

$$e^{tM} = P \begin{pmatrix} e^{t\lambda_1} & & \\ & \ddots & \\ & & e^{t\lambda_n} \end{pmatrix} P^{-1}.$$

Ainsi, les éléments spectraux de la matrice A sont déterminants dans l'étude qualitative du système.

– Si M est nilpotente, la formule se simplifie

$$e^{M} = \sum_{k=0}^{d-1} \frac{M^k}{k!}.$$

– Dans le cas non diagonalisable, on peut utiliser la décomposition de Jordan ou la décomposition de Dunford. Cette dernière (valable dans $\mathbb{C}$) s'écrit $A = D + N$ avec D diagonalisable et N nilpotente **qui commuttent !**. Ainsi, on a

$$e^{tA} = e^{tD}e^{tN},$$

et le calcul de chacun des facteurs peut se faire aisément.

Proposition I.27

Toute matrice $A \in M_d(\mathbb{C})$ (dans un corps quelconque, il suffit que le polynôme caractéristique soit scindé, ce qui est toujours vrai dans $\mathbb{C}$) s'écrit de manière unique sous la forme

$$A = D + N,$$

où D est diagonalisable, N nilpotente et $DN = ND$.
De plus, D et N peuvent s'exprimer sous la forme $D = P_1(A)$ et $N = P_2(A)$ où P_1 et P_2 sont deux polynômes.

Preuve.

– Admettons l'existence d'une telle décomposition $A = D_1 + N_1$. Soit $A = D_2 + N_2$ une autre décomposition (on ne suppose pas que D_2 et N_2 sont des polynômes en A). Comme D_2 et N_2 commuttent avec A, elles commuttent aussi avec $D_1 = P_1(A)$ et avec N_1. Ainsi, on a

$$D_1 - D_2 = N_2 - N_1.$$

Comme D_1 et D_2 commuttent et sont diagonalisables, on sait qu'elles peuvent être diagonalisées dans la même base et que donc leur différence $D_1 - D_2$ est également diagonalisable. De la même façon, la différence $N_2 - N_1$ est nilpotente (toujours grâce au fait qu'elles commuttent).

On a donc montré que la matrice $D_1 - D_2 = N_1 - N_2$ est à la fois diagonalisable et nilpotente, ce qui n'est possible que pour la matrice nulle.

– Montons maintenant l'existence. Soit $P(X) = \prod_{i=1}^{n}(X-\lambda_i)^{\alpha_i}$ le polynôme caractéristique (scindé) de A. On note $P_j(X) = \prod_{i\neq j}(X-\lambda_i)^{\alpha_i}$ et on observe que les $(P_j)_{1\leq j\leq n}$ sont premiers entre eux dans leur ensemble. D'après le théorème de Bezout, il existe donc des polynômes Q_j tels que

$$1 = \sum_{j=1}^{n} Q_j P_j.$$

On a donc, en évaluant cette égalité sur la matrice A :

$$I = \sum_{j=1}^{n} Q_j(A)P_j(A), \tag{I.19}$$

et donc

$$A = \sum_{j=1}^{n} AQ_j(A)P_j(A).$$

On remarque que pour tout $j \neq k$ on a $P_j(A)P_k(A) = P_k(A)P_j(A) = 0$ d'après le théorème de Cayley-Hamilton. En effet, par construction, le polynôme caractéristique de A divise le produit P_jP_k.

Ainsi (I.19) montre que les matrices $Q_j(A)P_j(A)$ sont des projecteurs sur leur image et que la somme directe de celles-ci remplit tout l'espace $\mathbb{C}^d$.

Posons maintenant

$$D = \sum_{j=1}^{n} \lambda_j Q_j(A)P_j(A), \quad N = \sum_{j=1}^{n} (A - \lambda_j I)Q_j(A)P_j(A).$$

- Il est clair que $D + N = A$ et que D et N commuttent avec A.
- D'après la propriété des projecteurs, on a

$$N^d = \sum_{j=1}^{n} Q_j(A)(A - \lambda_j I)^d P_j(A).$$

  Par construction, le polynôme caractéristique P divise $(X - \lambda_j)^d P_j(X)$ et donc par le théorème de Cayley-Hamilton on voit que tous les termes de la somme sont nuls, ce qui donne bien $N^d = 0$ et N est bien nilpotente.
- On pose $\pi(X) = \prod_{j=1}^{n}(X - \lambda_j)$ qui est scindé et à racines simples. On va montrer que π annulle D ce qui prouvera qu'elle est diagonalisable.

  On a tout d'abord

$$(D - \lambda_i I) = \sum_{j=1}^{n} (\lambda_j - \lambda_i)Q_j(A)P_j(A) = \sum_{j \neq i} (\lambda_j - \lambda_i)Q_j(A)P_j(A).$$

  et donc

$$\pi(D) = \prod_{i=1}^{d} \left(\sum_{j \neq i} (\lambda_j - \lambda_i)Q_j(A)P_j(A) \right).$$

  On voit que tous les termes de ce produit, une fois développé contiennent des termes de la forme dont l'un des facteurs est $P_i(A)P_j(A)$ avec $i \neq j$. Or ce facteur est nul comme on l'a vu plus haut (théorème de Cayley-Hamilton).

En réalité les projecteurs $Q_j(A)P_j(A)$ sont les projecteurs sur les sous-espaces caractéristiques de la matrice A.

■

Etudions complètement le cas de la dimension 2, i.e. $A \in M_2(\mathbb{R})$. Trois cas peuvent se produire (voir [**?**, page 290] pour une description plus détaillée) :

- A est diagonalisable dans $\mathbb{R}$. Auquel cas, la formule précédente donne la formule de la solution exacte. De façon plus précise, si λ_1 et λ_2 sont les deux valeurs propres et e_1, e_2 les vecteurs propres associés, les solutions de l'équation sont données par

$$y(t) = \alpha_1 e^{\lambda_1 t} e_1 + \alpha_2 e^{\lambda_2 t} e_2.$$

  Selon le signe des valeurs propres on peut donc tracer les trajectoires (Exercice : trouver les équations cartésiennes des trajectoires).
- A a ses valeurs propres réelles mais n'est pas diagonalisable. Dans ce cas, la valeur propre est nécessairement unique et non semi-simple. D'après le théorème de Jordan (ou la décomposition de Dunford), la matrice A s'écrit

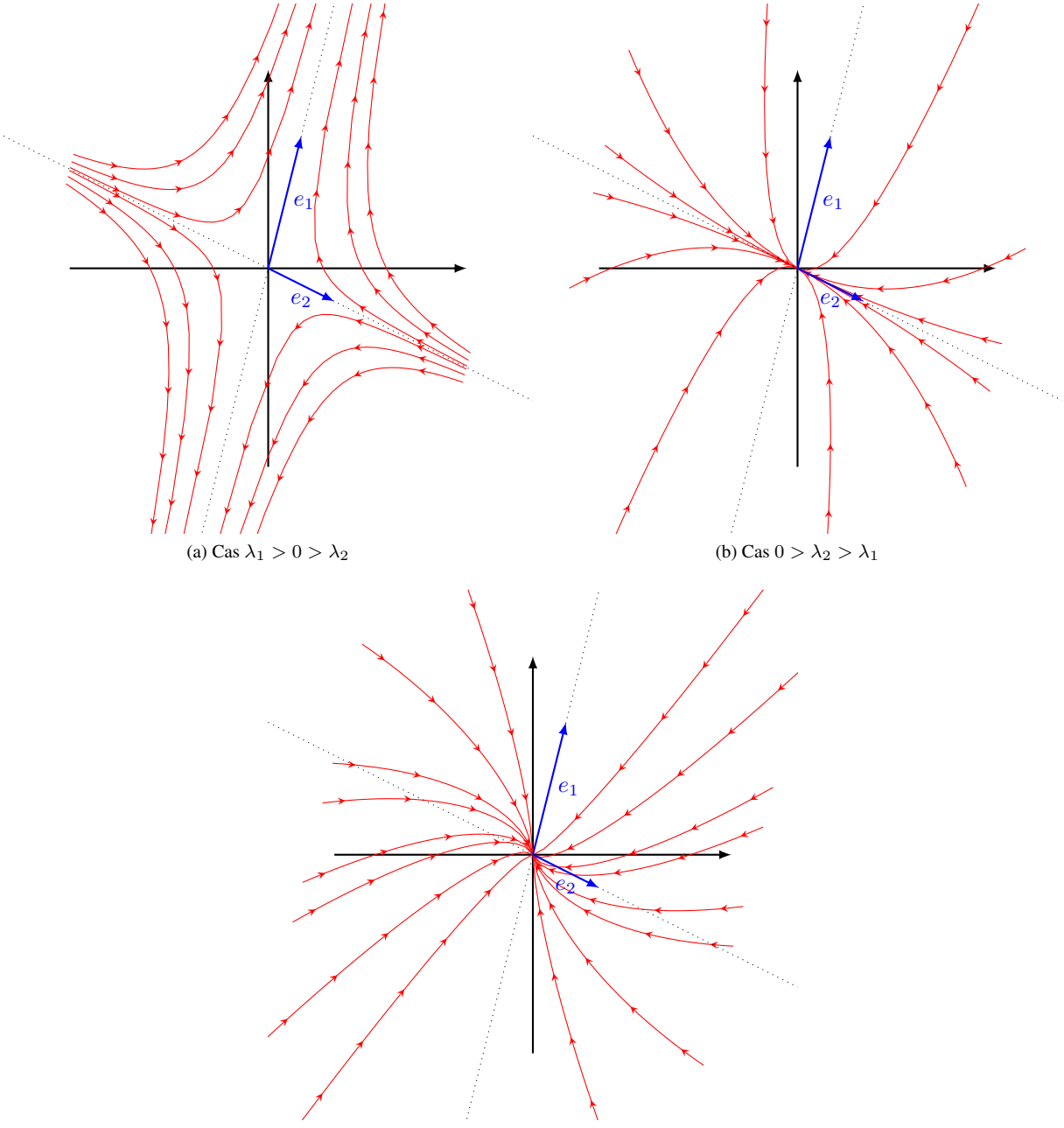
$$P^{-1}AP = \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix},$$

  et dans ce cas, on peut voir que l'exponentielle s'écrit

$$P^{-1}e^{tA}P = \begin{pmatrix} e^{\lambda t} & te^{\lambda t} \\ 0 & e^{\lambda t} \end{pmatrix}.$$

  Si on note e_1 et e_2 les deux colonnes de P (i.e. un vecteur propre de A et un vecteur propre généralisé), on trouve

$$y(t) = (\alpha_1 + \alpha_2 t)e^{\lambda t} e_1 + \alpha_2 e^{\lambda t} e_2.$$

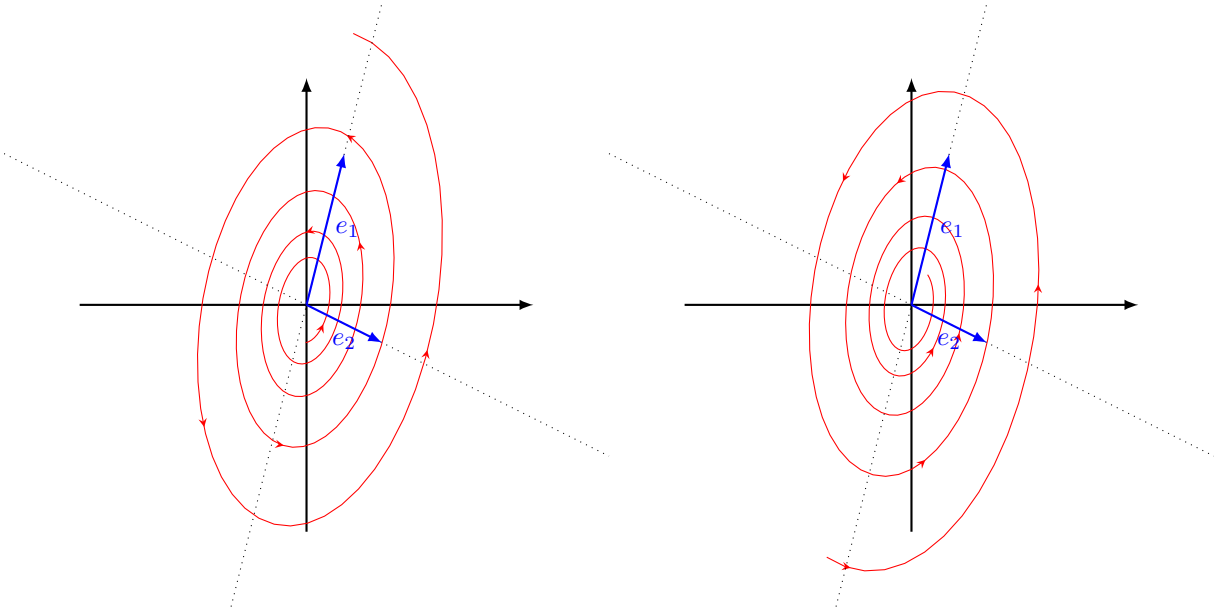


(a) Cas $\lambda_1 > 0 > \lambda_2$

(b) Cas $0 > \lambda_2 > \lambda_1$

(c) Cas $0 > \lambda_2 = \lambda_1$ non diagonalisable

FIGURE I.6 – Trois portraits de phase typiques dans le cas de deux valeurs propres réelles



(a) Cas $\lambda_\pm = \alpha \pm i\beta$, avec $\alpha > 0$ (b) Cas $\lambda_\pm = \alpha \pm i\beta$, avec $\alpha < 0$

FIGURE I.7 – Portraits de phase typiques dans le cas de deux valeurs propres complexes conjuguées

– Enfin, A peut avoir deux valeurs propres complexes conjuguées $\lambda_\pm = a \pm bi$, avec $b \neq 0$. Dans ce cas, on peut vérifier que l'exponentielle de tA est donnée par une formule du genre

$$P^{-1}e^{tA}P = e^{ta}\begin{pmatrix} \cos(bt) & \sin(bt) \\ -\sin(bt) & \cos(bt) \end{pmatrix}.$$

La stabilité du point d'équilibre est alors donnée par la valeur de a : si $a = 0$ les trajectoires sont périodiques, si $a > 0$ le point d'équilibre est instable et si $a < 0$ le point est stable. Qualitativement, les trajectoires s'enroulent autour de l'origine. Le sens de rotation est donné par l'analyse des vecteurs propres de A.

En dimension quelconque, la forme générale des solutions est donc connue :

– Si on travaille dans $\mathbb{C}$: Il existe une famille de solutions fondamentales de la forme

$$e^{\lambda_i t}\left(P_{i,j}(t)\right)_j,$$

où les $P_{i,j}$ sont des polynômes complexes en t. Leur degré maximal dépend de l'indice de nilpotence du bloc de Jordan correspondant.

– Si on travaille dans $\mathbb{R}$: la forme précédente reste valable pour les valeurs propres réelles. Pour les valeurs propres complexes, on profite du fait qu'elles soient nécessairement par paires conjuguées et que la matrice A est réelle pour obtenir

$$e^{\mathcal{R}e(\lambda_i)t}\left(\cos(\mathrm{Im}(\lambda_i)t)P_{i,j}(t) + \sin(\mathrm{Im}(\lambda_i)t)Q_{i,j}(t)\right)_j.$$

III.3 Résolvante. Formule de Duhamel

On considère l'équation linéaire homogène à coefficients variables (I.14).

Définition I.28

Pour tous $s, t \in \mathbb{R}$, on note $R(t,s) \in M_d(\mathbb{R})$ l'unique matrice telle que, pour tout $y_0 \in \mathbb{R}^d$, la solution y de (I.14) *pour la donnée de Cauchy (s, y_0) est donnée, à l'instant t, par*

$$y(t) = R(t,s)y_0.$$

Théorème I.29

- *Pour tout $t \in \mathbb{R}$, on a $R(t,t) = \mathrm{Id}$.*
- *Pour tous t, s, u, on a*
$$R(u,t)R(t,s) = R(u,s).$$
- *Pour tous t, s, on a*
$$R(t,s)R(s,t) = R(t,t) = \mathrm{Id},$$
en particulier la résolvante est inversible pour toutes valeurs de t, s.
- *Pour tous t, s, on a*
$$\partial_t R(t,s) = A(t)R(t,s),$$
et
$$\partial_s R(t,s) = -R(t,s)A(s).$$
- *Si A ne dépend pas du temps, on a*
$$R(t,s) = e^{(t-s)A}.$$

Preuve.
L'équation étant linéaire, le flot $\varphi(t,s,y_0)$ l'est également (par rapport à y_0), ce qui justifie la définition de $R(t,s)$ et prouve la plupart des propriétés ci-dessus.

La seule propriété non triviale est le calcul de la dérivée de R par rapport à s. Pour la vérifier il suffit de dériver (par rapport à s) la formule $R(t,s).R(s,t) = \mathrm{Id}$ (qui par ailleurs nous montre que R est dérivable par rapport à s). ■

Le calcul de la résolvante est quelque chose de complexe en général. Si on connaît d solutions indépendantes $(y_i)_i$ du problème, on peut reconstituer la résolvante (c'est même équivalent car, à s fixé, les d colonnes de la famille de matrices $t \mapsto R(t,s)$ forment une base de solutions du problème considéré). En effet, par définition de la résolvante, on a

$$\forall t \in I, \forall (\alpha_i)_i,\ \sum_{i=1}^{d} \alpha_i y(t) = R(t,s)\left(\sum_{i=1}^{d} \alpha_i y_i(s)\right).$$

Ceci donne

$$R(t,s) = W(t)W(s)^{-1},$$

où W est la matrice suivante, appelée matrice **Wronskienne** des $(y_i)_i$

$$W(t) = \big(y_1(t), \dots, y_d(t)\big).$$

On appelle Wronskien de cette famille de solutions, le déterminant de la matrice $W(t)$, on le note $w(t)$.

Proposition I.30 (Equation du Wronskien)

- *S'il existe $t_0 \in I$ tel que $W(t_0)$ est inversible, alors $W(t)$ est inversible pour tout t.*
- *Le wronskien w vérifie l'équation linéaire*
$$w'(t) = (Tr A(t))w(t),$$
et donc
$$w'(t) = w(t_0)\exp\left(\int_{t_0}^{t} Tr A(s)\, ds\right).$$

Preuve.

- Si les $(y_i(t_0))_i$ sont liés à un instant t_0, il existe donc des $(\alpha_i)_i$ non tous nuls tels que
$$\sum_{i=1}^{d} \alpha_i y_i(t_0) = 0.$$
Alors la fonction $\sum_{i=1}^{d} \alpha_i y_i(t)$ est solution de l'équation et s'annulle en 0, c'est donc la solution identiquement nulle. Ceci prouve bien que pour tout t, les $(y_i(t))_i$ sont liés.

– On note $B(t) = W(t)^{-1}A(t)W(t)$, de sorte que

$$y_i'(t) = A(t)y_i(t) = W(t)B(t)W(t)^{-1}y_i(t) = W(t)B(t)e_i = \sum_{j=1}^{d} b_{ji}(t)y_j(t).$$

Ainsi, en utilisant la multilinéarité du déterminant, on trouve

$$\begin{aligned} w'(t) &= \sum_{i=1}^{d} \det(y_1(t), ..., y_i'(t), ..., y_d(t)) = \sum_{i,j} b_{ji}(t) \det(y_1(t), ..., \underbrace{y_j(t)}_{\text{place } i}, ..., y_d(t)) \\ &= \sum_{i=1}^{d} b_{ii}(t)w(t) = (\mathrm{Tr}B(t))w(t), \end{aligned}$$

ce qui donne le résultat vu que $\mathrm{Tr}A(t) = \mathrm{Tr}B(t)$.

On peut bien sûr aussi utiliser la formule générale de la différentielle du déterminant

$$d(\det)(A).H = (\det A)\mathrm{Tr}(A^{-1}H), \text{ si } A \text{ est inversible,}$$

$$d(\det)(A).H = 0, \text{ si } A \text{ n'est pas inversible.}$$

■

Supposons maintenant connue la résolvante du problème homogène considéré. On va alors pouvoir calculer la solution générale du problème non homogène. Comme on l'a vu, il suffit de trouver une solution particulière du problème.

Fixons $s \in I$ et cherchons cette solution sous la forme

$$y(t) = R(t, s)\alpha(t), \ \ \forall t \in I,$$

avec $\alpha(t) \in \mathbb{R}^d$. Par calcul, on trouve

$$y'(t) = \partial_t R(t, s)\alpha(t) + R(t, s)\alpha'(t) = A(t)y(t) + R(t, s)\alpha'(t).$$

De sorte, que y est solution du problème si et seulement si

$$R(t, s)\alpha'(t) = b(t).$$

On trouve donc (en fixant la constante d'intégration arbitrairement)

$$\alpha(t) = \int_s^t R(s, \tau)b(\tau)\, d\tau,$$

ce qui fournit une solution particulière de la forme

$$y(t) = R(t, s)\alpha(t) = \int_s^t R(t, \tau)b(\tau)\, d\tau.$$

En mettant tous les ingrédients bout à bout, on obtient la formule générale pour la solution du problème de Cauchy pour la donnée (t_0, y_0) de l'équation linéaire non homogène

$$y(t) = R(t, t_0)y_0 + \int_{t_0}^t R(t, s)b(s)\, ds.$$

Cette formule, dite de Duhamel, ou de variation de la constante est très utile.

IV Petit bestiaire

La grande majorité des équations différentielles ne peuvent se résoudre à l'aide de formules explicites. On peut néanmoins, dans quelques cas, tenter de le faire. Je m'inspire ici très fortement de [**?**, Chap. VI]. On pourra aussi consulter [**?**, p 365].

– Le cas linéaire à coefficients constants est bien entendu le plus simple et, dans ce cas, on dispose de formules (modulo le calcul éventuel de primitives). Dans le cas mono-dimensionnel, on peut même résoudre le cas de coefficients variables.

– Equations à variables séparées : Ce sont les équations de la forme

$$y' = f(t)g(y),$$

où $g : \mathbb{R} \to \mathbb{R}$, $f : \mathbb{R} \to \mathbb{R}$. Dans ce cas on peut, en théorie, résoudre complètement le problème dès lors qu'on sait calculer (puis inverser ! !) les primitives de la fonction $1/g$ et les primitives de f. C'est donc le cas par exemple si f est constante et que g est une fraction rationnelle, ou un polynôme trigonométrique, etc ... il n'en demeure pas moins que le calcul est souvent long et pénible.
Exemple :

$$y' = \frac{1-y^2}{1-x^2}.$$

– Systèmes de dimension 2 dont on connait une intégrale première non-triviale.
On considère un système de la forme

$$y' = F(y),$$

avec $F : \mathbb{R}^2 \to \mathbb{R}^2$. On suppose connue une fonction $H : \mathbb{R}^2 \to \mathbb{R}$ non triviale (c'est une notion à définir proprement ...) telle que

$$\frac{d}{dt}H(y(t)) = 0,$$

pour toute solution y de l'équation différentielle.
Ainsi, les trajectoires dans le plan $\mathbb{R}^2$ sont contenues dans les lignes de niveau de H. Si on peut analyser suffisamment précisément la fonction H, il est aisé d'en déduire des propriétés des solutions (voir l'exercice **??**).
Supposons par exemple que l'on puisse appliquer le théorème des fonctions implicites par rapport à y_2 à la fonction H le long de sa ligne de niveau correspondant à la donnée initiale, on peut alors écrire (au moins localement)

$$H(y) = C_0 \Leftrightarrow y_2 = h(y_1),$$

et le système différentiel initial se ramène à un système de dimension 1

$$y_1' = F_1(y_1, h(y_1)).$$

Il est hors de propos de faire une théorie générale de cette approche mais il faut en comprendre l'idée générale et éventuellement être capable de la mettre en oeuvre sur des exemples simples.

– Dans le cas d'une équation scalaire mais non autonome l'approche précédente peut aussi permettre de résoudre complètement le problème. Exemple : pour l'équation

$$y' = \frac{y}{t+y^2},$$

on vérifie que la quantité $t/y - y$ est constante le long des trajectoires. Ceci montre que les solutions sont incluses dans les paraboles

$$t = y^2 + \lambda y,$$

où λ dépend de la donnée initiale. Ceci permet de poursuivre l'étude ...

– Equations de Bernoulli : ce sont les équations de la forme

$$y' = a(t)y + b(t)y^\alpha,$$

avec $\alpha \neq 1$.
  – Dans le cas $\alpha > 1$, l'équation vérifie les hypothèses de Cauchy-Lipschitz et on a existence, unicité de la solution et celle-ci ne peut s'annuler que si elle est identiquement nulle. A l'exception du cas $y \equiv 0$, on peut donc poser $z = y^{1-\alpha}$ et constater que z vérifie une équation **linéaire** non-homogène que l'on peut résoudre par les méthodes usuelles puis revenir à y.
  – Dans le cas $\alpha < 0$, le problème n'est défini que pour $y > 0$, donc la méthode précédente s'applique sans modification. La solution peut tendre vers 0 en temps fini.
  – Dans le cas $\alpha = 0$, il s'agit d'une équation linéaire non-homogène.
  – Dans le cas $0 < \alpha < 1$, on perd l'unicité des solutions et il faut donc faire attention aux différents recollements possibles.

– Equation de Ricatti : ce sont les équations de la forme

$$y' = a(t)y^2 + b(t)y + c(t).$$

Si on connaît une solution particulière $\tilde{y}$ de cette équation (en général une solution évidente), alors toutes les solutions s'obtiennent en écrivant $y = z + \tilde{y}$ et en vérifiant que z est alors solution d'une équation de Bernoulli (dont les coefficients dépendent de a, b, c **ET** de $\tilde{y}$).

– Equations homogènes : ce sont les équations de la forme

$$y' = f(y/x).$$

On effectue le changement de fonction inconnue $z = y/x$ de sorte que

$$z' = \frac{f(z) - z}{x},$$

et on est ramenés à une équation à variables séparées. On peut montrer différentes propriétés géométriques pour ce type d'équation.

– Passage en coordonnées polaires. Dans certains cas, la résolution de certains systèmes peut être simplifiée en passant en coordonnées polaires. Par exemple

$$x' = -y + x(x^2 + y^2),$$
$$y' = x + y(x^2 + y^2).$$

Si on pose $x = \rho cos(\theta)$ et $y = \rho \sin(\theta)$ (justifier cette écriture !) on voit que

$$\rho' = \rho^3,$$

$$\theta' = 1.$$

ce qui permet de calculer complètement la solution.

V Etude des systèmes autonomes

On considère un système différentiel autonome (i.e. le paramètre de temps t n'apparait pas dans l'équation) de la forme

$$y' = F(y),$$

où F est supposée suffisament régulière (disons de classe $\mathcal{C}^1$).

V.1 Portrait de phase

Etablir un portrait de phase d'un système autonome (en général de dimension 1 ou 2 pour des raisons évidentes de commodité de représentation) consiste à tracer les images des trajectoires dans le plan de phase c'est-à-dire, pour toute solution maximale (J, y), on représente l'ensemble

$$\{(y_1(t), y_2(t)) \ \ t \in J\}.$$

Cet ensemble est ou bien réduit à un point, ou bien est une courbe $\mathcal{C}^1$ sans point singulier. L'ensemble de ces courbes forme une partition du plan, il est donc hors de question de les tracer toutes, mais seulement une partie significative d'entre elles afin de bien représenter le comportement du système.

En général, on représente également à l'aide de flèches le sens de parcours de chaque trajectoire (on sait que la trajectoire ne peut pas changer de sens de parcours au cours du temps, boyez-vou pourquoi ?)

N.B. : le portrait de phase ne contient pas explicitement d'information sur la variable de temps. Ainsi le système initial et le système $y' = \alpha F(y)$ ont exactement le même portrait de phases (toutes les solutions de l'un se déduisent des solutions de l'autre par une homothétie en temps de rapport α).

Afin de tracer ce portrait de phases, plusieurs éléments sont utiles à étudier :

– **Points d'équilibre :** Ce sont les points y tels que $F(y) = 0$. Ce sont des points fixe du système au sens où la fonction constante égale à y est solution du système.

– **Trajectoires particulières :** Toute trajectoire explicite "simple" donne des informations précieuses sur le système car toute autre trajectoire ne peut intersecter ces solutions. Par exemple, il peut être fréquent qu'il existe des solutions du système $t \mapsto y(t)$ pour lesquelle l'une des coordonnées de $y(t)$ est constante (Cf. Lotka-Voltera par exemple).

– **Isoclines :** L'isocline $\alpha \in \mathbb{R} \cup \{\infty\}$ est l'ensemble défini par

$$I_\alpha = \left\{ (y_1, y_2) \in \mathbb{R}^2, \ \ \frac{f_2(y_1, y_2)}{f_1(y_1, y_2)} = \alpha \right\}.$$

L'utilité de ces ensembles provient du résultat suivant :

Si C est l'image d'une solution du système dans le plan de phase (y_1, y_2) qui coupe l'isocline I_α en un point y^*, alors la tangente à la courbe C en ce point est de pente α. Les deux cas très utiles en pratique sont :

– l'isocline I_0 : une trajectoire qui coupe l'isocline a une tangente horizontale en ce point.

  - l'isocline I_∞ : une trajectoire qui coupe cette isocline a une tangente verticale en ce point.
- Les isoclines I_0 et I_∞ se "coupent" (pas tout à fait en réalité ...) en les points d'équilibre du système et délimitent des régions du plan ou la trajectoire est monotone :

$$\begin{aligned} Q_{++} &= \{y \in \mathbb{R}^2,\ \ f_1(y) > 0, f_2(y) > 0\},\\ Q_{+-} &= \{y \in \mathbb{R}^2,\ \ f_1(y) > 0, f_2(y) < 0\},\\ Q_{-+} &= \{y \in \mathbb{R}^2,\ \ f_1(y) < 0, f_2(y) > 0\},\\ Q_{--} &= \{y \in \mathbb{R}^2,\ \ f_1(y) < 0, f_2(y) < 0\}. \end{aligned}$$

  Dans chacune de ces régions, les composantes y_1 et y_2 des solutions sont monotones (en particulier on peut déterminer aisément le sens de parcours des solutions).

V.2 Stabilité des points d'équilibre

Soit y^* un point d'équilibre du système, c'est-à-dire un point tel que $F(y^*) = 0$. On sait que que $t \mapsto y^*$ est une solution constante du système. La question de la stabilité consiste à essayer de comprendre le comportement du système pour une donnée initiale au voisinage de y^*.

Définition I.31

On dit que l'équilibre y^* est asymptotiquement stable (sous entendu en temps long*) s'il existe $\varepsilon > 0$ tel que pour toute donnée initiale $y_0 \in B(y^*, \varepsilon)$, on a*

- *La solution y du problème de Cauchy pour la donnée $(0, y_0)$ est définie sur $[0, +\infty[$.*
- *On a $\lim_{t\to+\infty} y(t) = y^*$.*

On peut également définie la notion de **stabilité** (pas asymptotique) qui est un peu plus faible : on demande juste que les solutions $t \mapsto y(t)$ restent proches de y^* si y_0 est assez proche de y^*.

Etudier la stabilité d'un point d'équilibre est un problème difficile en général et donne lieu à de nombreux développements mathématiques très importants. On ne donnera ici qu'un seul exemple de tel résultat (voir l'exercice **??** pour le cas linéaire et l'exercice **??** pour un exemple non linéaire).

Théorème I.32 (Th. de Lyapounov, [?, ?])

Soit y^* un point d'équilibre de F. Si la matrice Jacobienne de F en y^* a toutes ses valeurs propres de parties réelles strictement négatives, alors le point d'équilibre y^* est asymptotiquement stable.

La morale de l'histoire, sous les hypothèses du théorème, c'est que le comportement des solutions au voisinage de y^* est entièrement décrit par le système linéaire obtenu en remplaçant $F(y)$ par son linéarisé $DF(y^*)y$.

Remarque I.33

La condition du théorème est suffisante mais non nécessaire. Dans les cas où cette condition n'est pas vérifiée, l'analyse peut-être beaucoup plus délicate et demander des outils plus puissants.

Pour information, il existe des résultats un peu plus précis (mais plus difficiles !) :

Théorème I.34 (Hartmann-Grobman)

Soit y^* un point d'équilibre de F. On suppose que la matrice jacobienne de F en y^*, notée A, n'a aucune valeur propre imaginaire pure (on dit que A est hyperbolique). Alors il existe un homéomorphisme φ d'un voisinage de 0 dans un voisinage U de y^* tel que les solutions du système $y' = F(y)$ dans U sont données par

$$y(t) = \varphi(e^{tA}\varphi^{-1}(y_0)).$$

Autrement dit, les trajectoires du système non linéaire sont homéomorphes à celles du système linéaire.

En général φ n'est pas un difféomorphisme ! De plus dans le cas où la matrice n'est pas hyperbolique, le résultat tombe en défaut. C'est le cas par exemple du système

$$\begin{cases} x' = \pm x(x^2 + y^2),\\ y' = \pm y(x^2 + y^2). \end{cases}$$

VI Les équations d'ordre supérieur

Une équation différentielle scalaire d'ordre $k \geq 2$ est une équation différentielle de la forme

$$y^{(k)} = F(t, y, y', \dots, y^{(k-1)}), \tag{I.20}$$

ou cette fois on cherche une fonction k fois dérivable sur son intervalle de définition.

Toutes les définitions précédemment introduites (solutions maximales, globales, etc ...) s'adaptent dans changement au cas présent

L'étude de ces équations est assez simple dès lors qu'on constate qu'elles sont équivalentes à des systèmes d'ordre 1 dans $\mathbb{R}^k$.

Proposition I.35

Un couple (J, y) est solution de (I.20) *si et seulement si le couple* $\left(J, Y = \begin{pmatrix} y \\ y' \\ \vdots \\ y^{(k-1)} \end{pmatrix}\right)$ *est solution du système différentiel d'ordre* 1

$$Y' = \mathcal{F}(t, Y),$$

où

$$\forall t \in I, \forall Y \in \mathbb{R}^k, \ \ \mathcal{F}(t, Y) = \begin{pmatrix} Y_2 \\ Y_3 \\ \vdots \\ Y_k \\ F(t, Y_1, Y_2, \dots, Y_k) \end{pmatrix}.$$

On constate alors que $\mathcal{F}$ est continue et localement (resp. globalement) Lipschitzienne par rapport à sa variable d'état (i.e. la deuxième variable) si et seulement si la fonction F est continue et localement (resp. globalement) Lipschitzienne par rapport à ses variables d'état (i.e. ses $k-1$ dernières variables).

On peut ainsi appliquer toute la théorie de Cauchy-Lipschitz et traiter le cas des équations linéaires comme on l'a fait précédemment en utilisant la transformation à un système d'ordre 1 de taille d.

Dans le cas linéaire

$$y^{(n)} = a_{n-1}(t)y^{(n-1)} + ... + a_0(t)y + b(t),$$

le système du premier ordre de taille n équivalent est donné par

$$Y' = \underbrace{\begin{pmatrix} 0 & 1 & 0 & \cdots \\ 0 & 0 & 1 & 0 \cdots \\ \vdots & \vdots & \vdots & \ddots \\ a_0(t) & a_1(t) & \cdots & a_{n-1}(t) \end{pmatrix}}_{=A(t)} Y(t) + \underbrace{\begin{pmatrix} 0 \\ 0 \\ \cdots \\ b(t) \end{pmatrix}}_{=B(t)}.$$

La matrice $A(t)$ a la forme d'une matrice compagnon. L'équation différentielle vérifiée par le Wronskien $w(t)$ d'une famille de solutions est donc donnée par

$$w'(t) = a_{n-1}(t)w(t),$$

soit encore

$$w(t) = w(t_0) \exp\left(\int_{t_0}^{t} a_{n-1}(s)\, ds\right).$$

En particulier, si $a_{n-1} = 0$, le wronskien est constant au cours du temps !

Dans le cas de coefficients a_i constants, la matrice A est exactement la matrice compagnon du polynôme

$$P(X) = X^n - a_{n-1}X^{n-1} - \cdots - a_0.$$

C'est la raison pour laquelle vous avez sans doute appris dans les petites classes à résoudre "l'équation caractéristique" associée à l'équation différentielle (il s'agit juste de trouver les valeurs propres de A) et d'en déduire la forme générale des solutions. En fait, il s'agit de calculer l'exponentielle de tA sans le savoir ...

Si vous n'en avez pas l'habitude, la méthode de la variation de la constante doit **absolument** être faite en passant aux systèmes, sous peine de graves erreurs :

– Soit en utilisant directement la formule de Duhamel si vous vous en souvenez sans erreur ...

– Soit en faisant *varier les constantes* **sous la forme suivante** : si $y_1, ..., y_n$ est une base de solutions de l'équation homogène, on forme les fonctions

$$Y_i = \begin{pmatrix} y_i \\ y_i' \\ \vdots \\ y_i^{(n-1)} \end{pmatrix},$$

et on cherche une solution particulière de l'équation complète sous la forme

$$Y(t) = \alpha_1(t)Y_1(t) + \cdots + \alpha_n(t)Y_n(t),$$

où $\alpha_i : I \to \mathbb{R}$ sont des fonctions à déterminer. On voit alors que Y est solution de l'équation souhaitée si et seulement si

$$\alpha_1'(t)y_1(t) + \cdots + \alpha_n(t)'y_n(t) = 0,$$
$$\alpha_1'(t)y_1'(t) + \cdots + \alpha_n'(t)y_n'(t) = 0,$$

et ainsi de suite jusqu'à

$$\alpha_1'(t)y_1^{(n-2)}(t) + \cdots + \alpha_n'(t)y_n^{(n-2)}(t) = 0,$$
$$\alpha_1'(t)y_1^{(n-1)}(t) + \cdots + \alpha_n'(t)y_n^{(n-1)}(t) = b(t).$$

Les dérivées des α_i vérifient donc n équations, ce qui permet de les déterminer. Sous forme compacte, on écrit

$$W(t)\alpha'(t) = B(t),$$

où $\alpha'(t)$ est le vecteur des $\alpha_i'(t)$. Cette approche est bien sûr complètement équivalente à la formule de Duhamel mais permet de mener les calculs sans se souvenir de la formule générale.

Chapitre II

Les méthodes numériques à un pas

On s'intéresse dans cette section à l'étude de schémas numériques permettant d'obtenir des solutions approchées du problème de Cauchy

$$\begin{cases} y' = F(t,y), \ \text{ sur } [0,T], \\ y(0) = y_0. \end{cases} \tag{II.1}$$

Pour simplifier l'étude, on va supposer que F est globalement lipschitzienne par rapport à la variable d'état, sur $[0,T]\times\mathbb{R}^d$. On notera L sa constante de Lipscthitz, i.e. un nombre tel que

$$\|F(t,y_1) - F(t,y_2)\| \leq L\|y_1 - y_2\|, \ \ \forall t \in [0,T], \forall y_1, y_2 \in \mathbb{R}^d.$$

On supposera également que F est de classe $\mathcal{C}^1$ de sorte que la solution y du problème ci-dessus est globalement bien définie et de classe $\mathcal{C}^2$.

I La méthode d'Euler explicite

I.1 Définition et analyse de l'erreur

On se donne un nombre entier M et on pose $\Delta t = \frac{T}{M}$ et $t^n = n\Delta t$ pour $0 \leq n \leq M$.

La méthode d'Euler explicite revient à approcher la solution aux instants t^n du problème de Cauchy (II.1) par une relation de récurrence donnée par

$$y^{n+1} = y^n + \Delta t F(t^n, y^n), \ \ \forall 0 \leq n \leq M-1,$$

avec $y^0 = y_0$.

Définition II.1

Soit y la solution de (II.1)*. On appelle erreur de consistance (ou erreur de troncature) du schéma d'Euler explicite, la quantité*

$$R^n = \frac{y(t^{n+1}) - y(t^n)}{\Delta t} - F(t^n, y(t^n)), \quad \forall 0 \leq n \leq M-1.$$

L'erreur $e^n = y(t^n) - y^n$ vérifie alors l'équation

$$e^{n+1} = e^n + \Delta t(F(t^n, y(t^n)) - F(t^n, y^n)) + \Delta t R^n.$$

D'après le caractère Lipschitzien de F, on trouve

$$\|e^{n+1}\| \leq (1 + \Delta t L)\|e^n\| + \Delta t\|R^n\|.$$

D'après le lemme de Gronwall discret, et le fait que $e^0 = y^0 - y(0) = 0$, on obtient

$$\sup_{n\Delta t \leq T} \|e^n\| \leq \left(\sum_{n=0}^{M} \Delta t\|R^n\|\right) e^{LT}.$$

D'après les formules de Taylor, et le fait que y est solution du problème considéré, on obtient par calcul

$$\|R^n\| \leq \frac{\Delta t}{2} \sup_{[0,T]} \|y''\|,$$

ce qui prouve que

$$\sum_{n=0}^{M} \Delta t \|R^n\| \leq \Delta t \frac{T}{2} \|y''\|_\infty.$$

In fine, l'estimation d'erreur obtenue est bien d'ordre 1

$$\sup_{n\Delta t \leq T} \|e^n\| \leq \Delta t \frac{T}{2} \|y''\|_\infty e^{LT}.$$

Lemme II.2 (Lemme de Gronwall discret)

Soit $(z^n)_n$ une suite de nombres positifs avec $z^0 = 0$ et $C, L > 0$. On suppose que

$$\forall n \geq 0, z^{n+1} \leq C + L\Delta t \sum_{k=0}^{n} z^k,$$

alors on a

$$\forall n \geq 0, z^n \leq C e^{Ln\Delta t}.$$

Preuve :

On pose $h^n = C + L\Delta t \sum_{k=0}^{n} z^k$, de sorte que

$$h^{n+1} - h^n = L\Delta t z^{n+1} \leq L\Delta t h^n.$$

Il vient

$$h^{n+1} \leq (1 + L\Delta t) h^n \leq e^{L\Delta t} h^n,$$

et ainsi, on trouve

$$h^n \leq h^0 e^{Ln\Delta t} = C e^{Ln\Delta t},$$

d'où le résultat. ■

I.2 A propros de la stabilité

D'après ce qui précède, pour toute EDO raisonnable, la méthode d'Euler explicite converge à l'ordre 1. En particulier, elle est stable pour un pas de temps assez petit (c'est-à-dire que les erreurs ne sont pas amplifiées démesurément, voir le paragraphe III.3). Pour tester plus précisément les propriétés de stabilité de la méthode, on étudie son comportement dans le cas linéaire.

Commençons par le cas $d = 1$. On regarde l'équation $y' = \lambda y$ avec $\lambda \in \mathbb{C}$ de partie réelle négative ou nulle. La solution exacte est donnée par $y(t) = e^{\lambda t} y_0$ et donc on a

$$|y(t)| = e^{\mathcal{R}e\lambda t} |y_0| \leq |y_0|, \ \ \forall t \geq 0.$$

On va voir pour quelles valeurs des paramètres la méthode d'Euler explicite vérifie une propriété analogue. Dans le cas présent, la méthode s'écrit

$$y^{n+1} = y^n + \Delta t \lambda y^n = (1 + \Delta t \lambda) y^n.$$

Elle de dépend donc que du nombre complexe $z = \Delta t \lambda$ et on voit que la norme de y^n décroît si et seulement si

$$|1 + z| \leq 1 \Longleftrightarrow z \in \overline{D}(-1, 1).$$

Le disque fermé $\overline{D}(-1, 1)$ est appelé **zone de A-stabilité de la méthode** [1].

On voit donc que, si λ est fixé, on a une condition sur Δt pour que le schéma se comporte de façon satisfaisante. En particulier, si λ est imaginaire pur, la norme de la solution augmentera au cours du temps, quelque soit le pas de temps choisi ! !

On peut bien sûr faire la même analyse dans le cas vectoriel et regarder ce qui se passe pour l'équation $y' = Ay$ avec $A \in M_d(\mathbb{R})$. Si A est diagonalisable, on voit qu'il faut que toutes les valeurs propres de A multipliées par Δt soient dans la zone de A-stabilité définie plus haut.

1. La **A-stabilité** est parfois aussi appelée **stabilité absolue**

II La méthode d'Euler implicite

II.1 Définition et analyse de l'erreur

Il s'agit maintenant de la méthode numérique suivante

$$y^{n+1} = y^n + \Delta t F(t^{n+1}, y^{n+1}), \ \ \forall 0 \leq n \leq M-1,$$

avec $y^0 = y_0$.

La première difficulté liée à cette nouvelle méthode provient justement de son caractère implicite. En effet, l'existence même d'une suite $(y^n)_n$ satisfaisant ces équations n'est plus évidente.

Cette question peut être résolue par le lemme suivant :

Lemme II.3

Si $L\Delta t < 1$, et $t \in \mathbb{R}$ est fixé, la fonction

$$\Psi_t : y \mapsto y - \Delta t F(t, y),$$

est bijective et son inverse est Lipschitzienne de constante de Lipschitz $(1 - L\Delta t)^{-1}$.

qui lui-même est une simple application du résultat suivant qui sera démontré par la suite :

Lemme II.4

Soit $(E, \|.\|)$ un espace de Banach et $\Phi : E \mapsto E$ une application contractante (i.e. k-Lipschitzienne avec $k < 1$), alors $\mathrm{Id} - \Phi$ est inversible et son inverse est $(1-k)^{-1}$-Lipschitzienne.

On revient à l'analyse de la méthode d'Euler implicite.

Définition II.5

On définit l'erreur de consistance pour ce schéma comme

$$R^n = \frac{y(t^{n+1}) - y(t^n)}{\Delta t} - \Delta t F(t^{n+1}, y(t^{n+1})).$$

Ainsi, l'erreur d'approximation $e^n = y(t^n) - y^n$ vérifie

$$e^{n+1} = e^n + \Delta t (F(t^{n+1}, y(t^{n+1})) - F(t^{n+1}, y^{n+1})) + \Delta t R^n,$$

ce qui s'écrit encore

$$\Psi_{t^{n+1}}(y(t^{n+1})) - \Psi_{t^{n+1}}(y^{n+1}) = e^n + \Delta t R^n.$$

D'après le lemme II.3, on a

$$\|e^{n+1}\| \leq \frac{1}{1 - L\Delta t}(\|e^n\| + \Delta t \|R^n\|) \leq \left(1 + \frac{L\Delta t}{1 - L\Delta t}\right)\|e^n\| + \frac{\Delta t}{1 - L\Delta t}\|R^n\|.$$

En appliquant le lemme de Gronwall discret, on trouve

$$\sup_{n\Delta t \leq T} \|e^n\| \leq \left(\frac{1}{1 - L\Delta t}\sum_{n=0}^{M} \Delta t \|R^n\|\right) e^{\frac{Ln\Delta t}{1 - L\Delta t}}.$$

On peut vérifier que l'estimation de consistance est la même que précédemment

$$\|R^n\| \leq \frac{\Delta t}{2}\|y''\|_\infty.$$

On obtient finalement l'estimation

$$\sup_{n\Delta t \leq T} \|e^n\| \leq \Delta t \frac{T}{2(1 - L\Delta t)}\|y''\|_\infty e^{\frac{LT}{1 - L\Delta t}},$$

qui est donc du même type que celui obtenu pour la méthode explicite. La méthode d'Euler implicite est donc du premier ordre.

Il reste à prouver le résultat admis ci-dessus :

Preuve (du Lemme II.4):

Pour tout $y \in E$, on introduit la fonction $F_y : x \in E \mapsto F_y(x) = y + \Phi(x)$. Par hypothèse, celle-ci est contractante et admet donc un unique point fixe d'après le théorème du point fixe de Banach. Ainsi

$$\forall y \in E,\ \exists! x \in E,\ \ y = x - \Phi(x),$$

ce qui montre bien la bijectivité de $\mathrm{Id} - \Phi$. Si maintenant x_1, x_2, y_1, y_2 vérifient $y_1 = x_1 - \Phi(x_1)$ et $y_2 = x_2 - \Phi(x_2)$, on a

$$x_1 - x_2 = y_1 - y_2 + \Phi(x_1) - \Phi(x_2),$$

et donc

$$\|x_1 - x_2\| \leq \|y_1 - y_2\| + k\|x_1 - x_2\|,$$

ce qui fournit

$$\|x_1 - x_2\| \leq \frac{1}{1-k}\|y_1 - y_2\|,$$

c'est-à-dire le résultat attendu. ■

Remarquons que dans le cas linéaire (i.e. si $\Phi \in L(E)$), ce résultat est souvent appelé le Lemme de Neumann et il dit juste que si Φ est linéaire et de norme $|||\Phi||| < 1$ alors $\mathrm{Id} - \Phi$ est inversible (d'inverse continu) et on a

$$(\mathrm{Id} - \Phi)^{-1} = \sum_{k=0}^{+\infty} \Phi^k,$$

la série étant normalement convergente dans $L(E)$.

II.2 A propos de la stabilité

Refaisons l'analyse de stabilité effectuée précédemment pour la méthode explicite. Sur l'équation $y' = \lambda y$, avec $\lambda \in \mathbb{C}$ de partie réelle négative, la méthode s'écrit

$$y^{n+1} = y^n + \Delta t \lambda y^{n+1},$$

et admet toujours une unique solution (car $\Delta t\lambda$ ne peut jamais être égal à 1) donnée par

$$y^{n+1} = (1 - \Delta t\lambda)^{-1} y^n,$$

et de plus, en posant $z = \Delta t\lambda$ on a

$$|y^{n+1}| = |1 - z|^{-1}|y^n|,$$

dont la méthode est A-stable si et seulement si

$$|1 - z| \geq 1 \longleftrightarrow z \notin D(1,1).$$

Comme $\Delta t > 0$ et $\mathcal{R}e\lambda \leq 0$, on voit que cette condition est toujours remplie, quelque soit la valeur des paramètres.

La méthode d'Euler implicite est donc inconditionnellement A-stable.

II.3 Un petit tour vers l'équation de la chaleur

Quand on discrétise en espace l'équation de la chaleur (ou par une méthode de différences finies, ou par une méthode de Fourier), on est ramenés à résoudre le problème semi-discret suivant

$$y' = A_h y,$$

où $h > 0$ est le paramètre de discrétisation en espace et $y \in \mathbb{R}^N$ avec N grand ($N \sim 1/h$ dans le cas monodimensionel par exemple).

Dans le cas des différences finies 1D sur maillage régulier, la matrice A_h a la forme usuelle suivante

$$\frac{-1}{h^2}\begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & 2 & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & 2 \end{pmatrix}.$$

Ses valeurs propres sont réelles et négatives. La plus grande en module est de l'ordre de $4/h^2$.

Ainsi, si on applique une méthode d'Euler explicite à cette équation, la condition de A-stabilité devient $\Delta t|\lambda_{max}| \leq 2$, c'est-à-dire

$$\Delta t \leq \frac{h^2}{2}.$$

Cette condition est dramatiquement contraignante car elle oblige à prendre un pas de temps beaucoup plus petit que le pas d'espace (ce qui a en particulier pour conséquence d'augmenter énormément les temps de calcul).

A contrario, la méthode d'Euler implicite pour ce problème est inconditionnellement stable. Bien entendu (comme on a rien sans rien), cette méthode demande la résolution d'un gros système linéaire à chaque étape. C'est pourquoi on a aussi besoin de méthodes efficaces pour les systèmes linéaires (Cf. Le cours correspondant de P. Angot).

Remarque II.6

On dit souvent de façon abusive que la méthode d'Euler implicite est plus stable que la méthode explicite. Il faut prendre garde que ceci n'est vrai que dans le cadre de problèmes "dissipatifs" c'est-à-dire ceux pour lesquels la A-stabilité est la bonne notion.
Si d'aventure on s'intéresse à des problèmes du type $y' = \lambda y$ avec $\mathcal{R}e\lambda > 0$, alors c'est exactement l'inverse qui se produit : la méthode explicite est plus stable que la méthode implicite.

III Quelques mots sur la théorie générale des méthodes à un pas

Cette partie a pour vocation de vous introduire brièvement à une théorie générale des schémas présentée notamment dans [**?**, **?**]. Dans l'optique de la préparation à l'option de l'agrégation, il est plus important de comprendre les idées que le détail des preuves et des résultats, d'autant plus que les notations, les définitions, etc ... peuvent changer d'un auteur à un autre.

III.1 Introduction

On considère maintenant une méthode numérique à un pas (constant) écrite sous la forme générale suivante

$$y^{n+1} = y^n + \Delta t \Phi(t^n, y^n, \Delta t), \tag{II.2}$$

où Φ est une fonction continue de ses arguments.

Revenons sur les exemples précédents :

– Euler explicite :

$$\Phi(t^n, y^n, \Delta t) = F(t^n, y^n).$$

– Euler implicite :

$$\Phi(t^n, \cdot, \Delta t) = \frac{1}{\Delta t}\left((\,\mathrm{Id} - \Delta t F(t^n + \Delta t, \cdot))^{-1} - \,\mathrm{Id}\right),$$

à condition bien sûr que cette fonction soit inversible.

On voit dans ce dernier exemple qu'il n'est pas toujours aisé, ni souhaitable, de mettre explicitement un schéma numérique sous la forme générale (II.2).

III.2 Erreur de consistance

Définition II.7

On appelle erreur de consistance d'un schéma sous la forme (II.2) *associée à une solution exacte y de l'équation étudiée, la quantité*

$$R^n = \frac{y(t^{n+1}) - y(t^n)}{\Delta t} - \Phi(t^n, y(t^n), \Delta t).$$

ATTENTION : Cette définition est différente de celle que l'on trouve usuellement dans les livres (notamment dans le bouquin de Demailly). En effet, on trouve parfois la définition suivante

$$y(t^{n+1}) - y(t^n) - \Delta t \Phi(t^n, y(t^n), \Delta t)$$

qui à mon sens est trompeuse car ce terme est petit quelque soit le choix de Φ ! ! Avec la définition de R^n ci-dessus, la consistance peut s'énoncer, au moins intuitivement sous la forme suivante :

Le schéma est consistant avec l'EDO proposée, si pour toute solution exacte de l'équation, l'erreur de consistance R^n, tend vers 0 avec le pas de temps, uniformément pour $0 \leq n \leq T/\Delta t$.

Remarque II.8

Dans le cas de la méthode d'Euler explicite, cette définition coïncide avec celle donnée dans la définition II.1. En revanche, poru la méthode d'Euler explicite, si on note $\tilde{R}^n$ l'erreur de consistance définie dans la Définition II.5 et R^n celle définie ci-dessus, on peut montrer que

$$\|R^n\| \leq (1 - L\Delta t)\|\tilde{R}^n\|,$$

$$\|\tilde{R}^n\| \leq (1 + L\Delta t)\|R^n\|,$$

c'est-à-dire que les deux erreurs de consistance ainsi définies sont du même ordre.
Bien entendu, il faut adapter l'analyse de stabilité à la définition choisie de l'erreur de consistance.

Définition II.9

Le schéma est dit consistant à l'ordre p, si pour toute solution suffisamment régulière de l'EDO $y' = F(t, y)$, on a

$$\sup_{0 \leq n \leq T/\Delta t} \|R^n\| \leq C\Delta t^p.$$

Le schéma est simplement dit consistant si

$$\sup_{0 \leq n \leq T/\Delta t} \|R^n\| \xrightarrow[\Delta t \to 0]{} 0.$$

Proposition II.10

Un schéma (II.2) *est consistant si et seulement si*

$$\Phi(t, y, 0) = F(t, y), \quad \forall t \in [0, T], \forall y \in \mathbb{R}^d. \tag{II.3}$$

Preuve :

– Montrons que la condition est suffisante. On utilise le théorème des accroissements finis

$$R^n = y'(\xi^n) - \Phi(t^n, y(t^n), \Delta t),$$

avec $\xi^n \in]t^n, t^{n+1}[$, puis le fait que y est solution de l'EDO et l'hypothèse (II.3) pour obtenir

$$R^n = F(\xi^n, y(\xi^n)) - \Phi(t^n, y(t^n), \Delta t) = \Phi(\xi^n, y(\xi^n), 0) - \Phi(t^n, y(t^n), \Delta t).$$

Si on note ω, le module d'uniforme continuité de Φ sur le compact $[0, T] \times y([0, T]) \times [0, T]$, on en déduit

$$\|R^n\| \leq \omega(|t^n - \xi^n| + \|y(\xi^n) - y(t^n)\| + \Delta t) \leq \omega(C\Delta t),$$

et ceci est valable pour tout $0 \leq n \leq T/\Delta t$ (car on utilise le fait que $t^n \in [0, T]$).

– Montrons que la condition est nécessaire. On va en fait montrer (II.3) pour $t = 0$. Le résultat général s'en déduit par translation en temps.
Soit $y_0 \in \mathbb{R}^d$ et soit y la solution du problème de Cauchy associé à la donnée $(0, y_0)$ et à l'EDO considérée. On calcule l'erreur de consistance au premier cran

$$R^0 = \frac{y(\Delta t) - y_0}{\Delta t} - \Phi(0, y_0, \Delta t).$$

Par hypothèse, cette quantité tend vers 0 avec Δt. Le premier terme tend vers $y'(0)$ et donc vers $F(0, y_0)$ et le second terme tend vers $\Phi(0, y_0, 0)$ car Φ est continue, ce qui prouve bien que

$$F(0, y_0) = \Phi(0, y_0, 0).$$

■

III.3 Stabilité

Définition II.11

Le schéma (II.2) *est dit stable si le cumul des erreurs (d'approximation, de consistance, d'arrondis, etc ..) au cours du calcul est contrôlé par la taille des erreurs commises.*
Plus précisément cela signifie qu'il doit exister une constante M, indépendante du pas de temps, telle que pour tout pas de temps $\Delta t > 0$ et pour toutes suites $(y^n)_n$, $(\tilde{y}^n)_n$, $(\varepsilon^n)_n$ vérifiant

$$y^{n+1} = y^n + \Delta t \Phi(t^n, y^n, \Delta t),\ \forall n \geq 0,$$

$$\tilde{y}^{n+1} = \tilde{y}^n + \Delta t \Phi(t^n, \tilde{y}^n, \Delta t) + \varepsilon^n,\ \ \forall n \geq 0,$$

on a

$$\sup_{0 \leq n \leq T/\Delta t} \|y^n - \tilde{y}^n\| \leq M \left(\|y^0 - \tilde{y}^0\| + \sum_{0 \leq k \leq T/\Delta t} \|\varepsilon^k\| \right).$$

Théorème II.12

Si Φ est Lipschitzienne par rapport à y (uniformément par rapport à $t \in [0,T]$ et $\Delta t \in [0, \Delta t_{max}]$) alors le schéma est stable.

Preuve :
Il s'agit simplement d'appliquer le lemme de Gronwall discret (Lemme II.2). ■

III.4 Convergence

Théorème II.13

Si le schéma (II.2) *est stable et consistant à l'ordre p alors il est convergent à l'ordre p c'est-à-dire que*

$$\sup_{0 \leq n \leq T/\Delta t} \|y^n - y(t^n)\| \leq C \Delta t^p,$$

où $(y^n)_n$ est la suite d'approximations obtenue par le schéma à partir de la donnée $y^0 = y_0$ et $t \mapsto y(t)$ est la solution exacte du problème de Cauchy.

Preuve.
Par définition de l'erreur de consistance R^n, la suite des valeurs exactes $\tilde{y}^n = y(t^n)$ vérifie

$$\tilde{y}^{n+1} = \tilde{y}^n + \Delta t \Phi(t^n, \tilde{y}^n, \Delta t) + \Delta t R^n.$$

On applique la propriété de stabilité puis la propriété de constance, ce qui donne

$$\sup_{0 \leq n \leq T/\Delta t} \|y^n - \tilde{y}^n\| \leq M \sum_{0 \leq n \leq T/DT} \Delta t \|R^n\| \leq MT \left(\sup_{0 \leq n \leq T/\Delta t} \|R^n\| \right) \leq C \Delta t^p.$$

■

III.5 Exemples d'autres méthodes

– Méthodes de Taylor : L'idée est de construire le schéma à partir du développement de Taylor de y. Notez que l'analyse de consistance est alors immédiate.
Par exemple, à l'ordre 2, on trouve

$$y(t + \Delta t) = y(t) + \Delta t y'(t) + \Delta t^2 / 2 y''(t) + O(\Delta t^3).$$

On utilise alors l'équation pour écrire $y' = F(t,y)$ puis $y'' = \partial_t F(t,y) + D_y F(t,y).F(t,y)$ et ainsi obtenir le schéma d'ordre 2

$$y^{n+1} = y^n + \Delta t F(t^n, y^n) + \Delta t^2/2 \Big(\partial_t F(t^n, y^n) + D_y F(t^n, y^n).F(t^n, y^n) \Big).$$

L'inconvénient majeur de cette méthode est qu'elle nécessite le calcul de dérivées de F.
En guise d'exercice, vous pouvez étudier la A-stabilité de cette méthode, au moins pour des valeurs réelles de λ.

– Méthode de Heun (une méthode de Runge-Kutta d'ordre 2) :

$$y^{n+1} = y^n + \frac{\Delta t}{2}\Big(F(t^n, y^n) + F(t^{n+1}, y^n + \Delta t F(t^n, y^n))\Big).$$

– Méthode du point-milieu

$$y^{n+1} = y^n + \Delta t F\Big(t^{n+\frac{1}{2}}, y^n + \frac{\Delta t}{2} F(t^n, y^n)\Big).$$

Dans le cas linéaire autonome, cette méthode est la même que la précédente. Vous pouvez la-aussi etudier sa A-stabilité.

– Méthode de Runge-Kutta d'ordre 4 (RK4) :

$$\begin{aligned}
k_1 &= F(t^n, y^n),\\
k_2 &= F\left(t^n + \frac{\Delta t}{2}, y^n + \frac{\Delta t}{2} k_1\right),\\
k_3 &= F\left(t^n + \frac{\Delta t}{2}, y^n + \frac{\Delta t}{2} k_2\right),\\
k_4 &= F\left(t^{n+1}, y^n + \Delta t k_3\right),\\
y^{n+1} &= y^n + \frac{\Delta t}{6}\big(k_1 + 2k_2 + 2k_3 + k_4\big).
\end{aligned}$$

Les inconnues intermédiaires k_1, k_2, k_3, k_4 sont des approximations de la pente de la corde entre la solution au temps t^n et celle au temps t^{n+1}. Il faut évidemment les recalculer à chaque pas de temps.

Cette méthode est d'ordre 4 et possède une assez grande zone de A-stabilité (assez compliquée à étudier néanmoins).

IV Compléments

IV.1 Equations d'ordre supérieur

On peut appliquer toutes les idées précédentes aux équations d'ordre supérieur. Les notions de consistance, stabilité et convergence sont adaptées au cas par cas.

Une façon de faire (mais qui n'est pas la seule) est d'appliquer les schémas étudiés précédemment au système d'ordre 1 équivalent à l'équation considérée.

Considérons par exemple l'équation suivante

$$y'' + q(t)y = 0.$$

Si on écrit le système équivalent $Y' = A(t)Y$ et qu'on lui applique le schéma d'Euler explicite, on arrive facilement au schéma suivant

$$\frac{y^{n+1} - 2y^n + y^{n-1}}{\Delta t^2} + q(t^{n-1})y^{n-1} = 0, \ \ \forall n \geq 1,$$

avec les données initiales

$$y^0 = y_0, \ \ y^1 = y_0 + \Delta t \dot{y}_0.$$

On peut y préférer le schéma suivant

$$\frac{y^{n+1} - 2y^n + y^{n-1}}{\Delta t^2} + q(t^n)y^n = 0, \ \ \forall n \geq 1,$$

qui a l'avantage d'être d'ordre 2.

Chapitre III

Les équations de transport

- Quelques outils de la théorie de Fourier sont utilisés dans ce chapitre. Ils sont rappelés dans la section V.
- La section IV n'a pas été traitée en cours. Elle n'est là qu'à titre informatif pour ceux qui voudraient aller un peu plus loin et avoir une idée des difficultés rencontrées dans l'étude des équations hyperboliques non-linéaires.

I Modèles de transport

I.1 Trafic routier

- **Hypothèses de modélisation :** On modélise une route nationale rectiligne, avec une seule voie de circulation. On suppose que la route est infinie (autrement dit, on ne regarde pas les problèmes de bord) et qu'il n'y a pas de dépassements. De plus, on ne regarde pas les problèmes d'entrées et sorties.
  On note $\rho(t,x)$ la densité de véhicules et $v(t,x)$ la vitesse moyenne des véhicules à l'instant t et au point x.
- **Trajectoire d'un véhicule particulier :** Supposons connues ρ et v en tout point et à tout instant.
  On considère un véhicule qui se situerait à la position x_0 et à l'instant t_0. On note $X(t,t_0,x_0)$ sa position à un autre instant t. Par définition du champ de vitesse on a la relation

$$\partial_t X(t,t_0,x_0) = v(t,X(t,t_0,x_0)). \tag{III.1}$$

  On voit donc que, si on connait v et que (v est suffisamment régulier), la trajectoire d'un véhicule est donnée par la solution de l'équation différentielle (III.1) associée à la donnée initiale (t_0,x_0).
- **La loi de conservation des véhicules :**

Proposition III.1

Si elles sont régulières, les deux fonctions ρ et v sont liées par l'équation aux dérivées partielles

$$\partial_t \rho + \partial_x(\rho v) = 0. \tag{III.2}$$

Preuve :
On va donner plusieurs façons de démontrer ce résultat (elles sont bien sûr totalement équivalentes mais c'est plutôt *l'esprit du calcul* qui est diffère)

- La preuve des physiciens : on écrit un bilan de l'évolution du nombre de véhicules entre a et b entre les instants t et $t+\delta$:

$$\int_a^b \rho(t+\delta t,x)\,dx = \int_a^b \rho(t,x)\,dx + \delta t\, v(t,a)\rho(t,a) - \delta t\, v(t,b)\rho(t,b).$$

  La contribution des deux derniers termes s'écrit aussi $-\delta t \int_a^b \partial_x(\rho v)(t,x)\,dx$, il suffit alors de constater que ce bilan doit être vrai pour tout t et tous a et b, ce qui donne le résultat.
- La preuve des matheux : on dérive par rapport à s la relation suivante

$$\int_a^b \rho(t,x)\,dx = \int_{X(t+s,t,a)}^{X(t+s,t,b)} \rho(t+s,x)\,dx,$$

  qui décrit le fait que le nombre de véhicules compris entre deux véhicules donnés (ceux qui étaient en a et en b à l'instant t) reste constant au cours du temps. On utilise alors la définition des caractéristiques et la fin de la preuve est identique à la précédente.

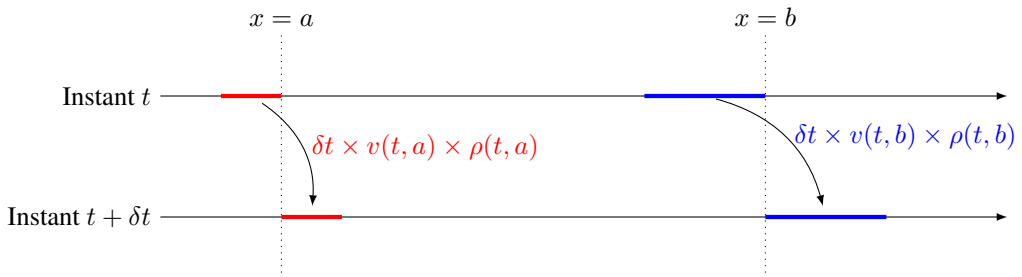


FIGURE III.1 – Preuve des physiciens

■

La loi de conservation (III.2) est valable en toute généralité sous les hypothèses faites au départ.

– **Modélisation du comportement des véhicules :** Il faut trouver un lien entre la densité ρ et la vitesse v.

1. Premier modèle : tous les conducteurs vont à la même vitesse (la vitesse maximale autorisée) quelles que soient les conditions de trafic :
$$v(t,x) = V_{max} = cte.$$
Dans ces conditions, v n'est bien sûr plus une inconnue et il reste l'équation
$$0 = \partial_t \rho + \partial_x(\rho V_{max}) = \partial_t \rho + V_{max}\partial_x \rho,$$
qui est appellée l'équation de transport (ou d'advection) à vitesse constante V_{max}.
2. Second modèle : La vitesse maximale autorisée varie le long de la route selon une loi donnée $V_{max}(x)$, l'équation vérifiée est alors
$$0 = \underbrace{\partial_t \rho + \partial_x(\rho V_{max}(x))}_{\text{forme conservative}} = \underbrace{\partial_t \rho + V_{max}(x)\partial_x \rho + V'_{max}(x)\rho}_{\text{forme non conservative}}.$$
3. Troisième modèle : les conducteurs adaptent leur vitesse au trafic. Plus le trafic est dense, plus la vitesse est réduite, ce qui donne, par exemple, une loi du type
$$v = V_{max}(1-\rho),$$
et donc l'équation aux dérivées partielles **non-linéaire**
$$0 = \partial_t \rho + V_{max}\partial_x(\rho(1-\rho)).$$
On peut aussi l'écrire en fonction de la vitesse
$$0 = \partial_t v + \partial_x(v(v - V_{max})).$$

I.2 Dynamique des gaz

Etudions un contexte simple de mécanique des fluides : on étudie l'évolution isentropique d'un gaz parfait en une dimension d'espace (dans un tuyau de gaz par exemple). Les variables qui décrivent le fluide dans ces conditions sont : la densité $\rho(t,x)$, le champ de vitesse $v(t,x)$ et la pression $p(t,x)$. On néglige la gravité (par exemple parce qu'elle agit orthogonalement au tube).

La première équation qui régit l'évolution du système est l'équation de conservation de la masse. Comme précédemment, on obtient
$$\partial_t \rho + \partial_x(\rho v) = 0.$$

On écrit la quantité de mouvement d'une portion de fluide
$$\frac{d}{dt}\left(\int_{X(t,t_0,a)}^{X(t,t_0,b)} \rho(t,x)v(t,x)\,dx\right) = \text{Somme des forces} = -p(X(t,t_0,b)) + p(X(t,t_0,a)).$$
Ceci étant vrai pour tous réels $a < b$ et tout instant t. On refait le calcul de dérivation comme précédemment et on touve
$$\partial_t(\rho v) + \partial_x(\rho v^2) = -\partial_x p,$$
ce que l'on écrit souvent sous la forme
$$\partial_t(\rho v) + \partial_x(\rho v^2 + p) = 0.$$
Le système de deux équations aux dérivées partielles obtenu est appelé : **équations d'Euler**. Il reste à modéliser le comportement thermodynamique du gaz.

1. Cas numéro 1 : gaz sans pression, ou plus exactement à pression constante. Il s'agit d'une hypothèse très simplificatrice, on trouve alors le système
$$\begin{cases} \partial_t \rho + \partial_x(\rho v) = 0, \\ \partial_t(\rho v) + \partial_x(\rho v^2) = 0 \end{cases}$$
qui est formellement équivalent (pour des solutions régulières telles que $\rho \neq 0$) à l'équation scalaire
$$\partial_t v + v\partial_x v = 0,$$
couplé à l'équation de conservation de la masse.
2. Cas numéro 2 : On va se placer dans un cadre simple dit *isentropique*. Dans ces conditions la pression est liée à la densité par la relation
$$p = p_0 \rho^\gamma,$$
où p_0 et γ sont des constantes positives données ($\gamma = 1.4$ pour un gaz diatomique (O_2, H_2, etc ...).
Le système d'équations devient alors le système des équations d'Euler isentropique :
$$\begin{cases} \partial_t \rho + \partial_x(\rho v) = 0, \\ \partial_t(\rho v) + \partial_x(\rho v^2 + p_0\rho^\gamma) = 0. \end{cases}$$
Ce système est fortement non-linéaire et hautement non trivial à résoudre et à comprendre.
3. Cas numéro 3 : Cadre isotherme. Dans le cas des gaz parfaits, cela revient à dire que la loi liant la pression et la densité est linéaire. On trouve un système de la forme
$$\begin{cases} \partial_t \rho + \partial_x(\rho v) = 0, \\ \partial_t(\rho v) + \partial_x(\rho v^2 + a^2\rho) = 0. \end{cases}$$
On peut récrire ce système en posant $\eta = \log \rho$ comme nouvelle variable, ce qui donne
$$\begin{cases} \partial_t \eta + v\partial_x \eta + \partial_x v = 0, \\ \partial_t v + v\partial_x v + a^2\partial_x \eta = 0. \end{cases}$$
Si on regarde des régimes de petites vitesses (on dit qu'on linéarise autour de $v = 0$) on trouve un système approché
$$\begin{cases} \partial_t \eta + \partial_x v = 0, \\ \partial_t v + a^2\partial_x \eta = 0. \end{cases}$$
Et donc η et v vérifient la même équation des ondes
$$\partial_t^2 \eta - a^2\partial_x^2 \eta = 0,$$
$$\partial_t^2 v - a^2\partial_x^2 v = 0.$$
Le réel a est la vitesse des ondes, dans le cas présent il s'agit de la vitesse du son.

Tous ces systèmes sont dits *hyperboliques* (pour des raisons de géométrie du symbole de l'opérateur) mais ce qu'il faut retenir c'est que cela signifie que localement, on peut toujours trouver des changements de variable grâce auxquels le système ressemble à des équations de type transport.

II Analyse des lois de conservation scalaires linéaires

II.1 L'équation de transport 1D à vitesse constante :

On s'intéresse à la résolution du problème
$$\begin{cases} \partial_t u + c\partial_x u = 0, \\ u(t=0,x) = u_0(x), \end{cases} \tag{III.3}$$
posée dans $\mathbb{R}^+ \times \mathbb{R}$.

Proposition III.2

Si u_0 est une fonction de classe $\mathcal{C}^1$, alors le problème (III.3) *admet une unique solution de classe $\mathcal{C}^1$ donnée par*
$$u(t,x) = u_0(x - ct). \tag{III.4}$$

Preuve :

Il suffit dans un premier temps de vérifier que la formule donnée est bien solution de l'équation. Pour démontrer l'unicité il y a deux façons complémentaires de voir la chose.

– Méthode 1 : on vérifie que si u est une solution alors, pour tout $x_0 \in \mathbb{R}$, la fonction

$$t \mapsto \varphi(t) = u(t, x_0 + ct),$$

est constante, on obtient donc le résultat.

– Méthode 2 : par estimation d'énergie L^2. On regarde l'équation vérifiée par u^2 (c'est formellement la même) et on intègre l'équation sur un domaine de dépendance trapézoidal. On obtient l'estimation

$$\int_{a+ct}^{b+ct} u^2(t,x)\,dx = \int_a^b u^2(0,x)\,dx,$$

valable pour tous $a < b$, $t > 0$. En particulier : si $u(0,\cdot)$ est nulle sur $[a,b]$ alors $u(t,\cdot)$ est nulle sur $[a+ct, b+ct]$. Ceci fournit l'unicité et aussi la notion de vitesse finie de propagation.

■

II.2 Le cas de la vitesse variable

On a obtenu que les solutions (classiques et faibles !) de l'équation d'advection scalaire à vitesse constante sont constantes le long de droites dans le plan (t,x) d'équation $x - ct = x_0$.

On peut donc établir le même genre de propriétés que dans le cas constant. On s'intéresse dans ce paragraphe au problème suivant

$$\begin{cases} \partial_t u + c(t,x)\partial_x u = 0, \\ u(t=0,x) = u_0(x), \end{cases} \tag{III.5}$$

posé dans $\mathbb{R}^+ \times \mathbb{R}$.

On suppose dans toute la suite que la fonction $(t,x) \mapsto c(t,x)$ est donnée, régulière et bornée (hypothèses non optimales mais suffisantes).

Comme dans le modèle de trafic routier, on introduit les trajectoires associées à la vitesse c, qu'on appelle *courbes caractéristiques* et sont solutions de l'équation dite *équation caractéristique* donnée par

$$\begin{cases} \dfrac{d}{dt} X(t,t_0,x_0) = c(t, X(t,t_0,x_0)), \\ X(t_0,t_0,x_0) = x_0. \end{cases} \tag{III.6}$$

– Avec les hypothèses sur la fonction c, d'après le théorème de Cauchy-Lipschitz (Théorème I.7) et le théorème d'explosion en temps fini (Théorème I.16), on sait que ce problème de Cauchy est globalement bien posé pour toutes données t_0, x_0, c'est-à-dire qu'il possède une unique solution définie sur $\mathbb{R}$ tout entier.

– Si $c(x)$ est la fonction constante, on trouve les droites déjà rencontrées

$$X(t,t_0,x_0) = x_0 + c(t - t_0).$$

– Dans le cas général, ces courbes ne sont pas des droites mais elles ne se croisent pas d'après le théorème de Cauchy-Lipschitz.

– Par unicité dans Cauchy-Lipschitz, on a la propriété de semi-groupe

$$X(t,s,X(s,u,x_0)) = X(t,u,x_0),$$

et en particulier

$$X(t,s,X(s,t,x_0)) = x_0, \quad \forall t,s,x_0,$$

autrement dit

$$x \mapsto X(t,s,x)$$

est un difféomorphisme dont l'inverse est $x \mapsto X(s,t,x)$.

Proposition III.3

Toute solution au sens classique du problème (III.5) *est constante le long des courbes caractéristiques :*

$$u(t, X(t,t_0,x_0)) = u(t_0,x_0), \quad \forall t,t_0,x_0.$$

Sous les hypothèses ci-dessus, cette solution existe et est unique et est donnée par (en prenant $t_0 = 0$ et $x = X(t,0,x_0)$.

$$u(t,x) = u_0(X(0,t,x)).$$

Preuve :

Soit u une solution du problème. On fixe $x_0 \in \mathbb{R}$, et on introduit $\varphi(t) = u(t, X(t, t_0, x_0))$. Par calcul, on trouve

$$\varphi'(t) = \partial_t u(t, X(t, t_0, x_0)) + \underbrace{\frac{dX}{dt}(t, t_0, x_0)}_{=c(t,X(t,t_0,x_0))} .\partial_x u(t, X(t, t_0, x_0)) = 0,$$

car u est solution de l'équation de transport à la vitesse c. Ainsi la fonction φ est constante et le résultat est démontré. ■

Ceci redonne les formules obtenues dans le cas des coefficients constants. Par ailleurs, on peut montrer l'unicité des solutions sans utiliser les caractéristiques, par estimation d'énergie L^2, de la façon suivante. C'est surtout la technique de preuve qu'il faut comprendre et retenir :

Théorème III.4

On suppose que c n'est pas constante et on note $\alpha = \sup_{\mathbb{R}} c$ et $\beta = \inf_{\mathbb{R}} c$.
Si u_0 est identiquement nulle sur un intervalle $[a, b]$, alors pour tout $t < \frac{b-a}{\alpha-\beta}$, on a

$$u(t, \cdot) = 0, \quad \textit{sur } [a + \alpha t, b + \beta t].$$

Remarquons que $\alpha > \beta$ et donc l'intervalle en question se *rétrécit* au cours du temps.

Preuve :

On commence par remarquer que u^2 vérifie la même équation que u

$$\partial_t u^2 + c(t, x)\partial_x u^2 = 0. \tag{III.7}$$

Tout est alors basé sur le calcul suivant :

$$\begin{aligned}\frac{d}{dt}\left(\int_{a+\alpha t}^{b+\beta t} u^2(t, x)\, dx\right) &= \beta u^2(t, b + \beta t) - \alpha u^2(t, a + \alpha t) + \int_{a+\alpha t}^{b+\beta t} \underbrace{\partial_t(u^2)}_{=-c(t,x)\partial_x(u^2)}\, dx \\ &= \underbrace{(\beta - c(t, b + \beta t))u^2(t, b + \beta t)}_{\leq 0} + \underbrace{(c(t, a + \alpha t) - \alpha)u^2(t, a + \alpha t)}_{\leq 0} \\ &\quad + \int_{a+\alpha t}^{b+\beta t} \partial_x c(t, x)u^2(t, x)\, dx.\end{aligned}$$

On obtient donc

$$\frac{d}{dt}\left(\int_{a+\alpha t}^{b+\beta t} u^2(t, x)\, dx\right) \leq \int_{a+\alpha t}^{b+\beta t} \partial_x c(t, x)u^2(t, x)\, dx.$$

Si t est petit comme dans l'énoncé du théorème, les bornes de l'intégrale sont dans le bon sens et on a donc la majoration

$$\frac{d}{dt}\left(\int_{a+\alpha t}^{b+\beta t} u^2(t, x)\, dx\right) \leq \|\partial_x c\|_\infty \int_{a+\alpha t}^{b+\beta t} u^2(t, x)\, dx.$$

On peut alors conclure par un lemme de type Gronwall que l'on rappelle ici :

Lemme III.5

Soit $M \in \mathbb{R}$ et $t \mapsto \varphi(t)$ une fonction dérivable qui vérifie

$$\varphi'(t) \leq M\varphi(t), \ \ \forall t,$$

alors on a

$$\varphi(t) \leq \varphi(0)e^{Mt}. \tag{III.8}$$

On applique alors le lemme à la fonction φ définie par

$$\varphi(t) = \int_{a+\alpha t}^{b+\beta t} u^2(t, x)\, dx,$$

ce qui montre que pour tout t convenable, on a

$$\varphi(t) \leq \varphi(0)e^{\|\partial_x c\|_\infty t}.$$

Or, par hypothèse on a $u_0 = 0$ sur $[a, b]$, ce qui implique $\varphi(0) = 0$ et donc $\varphi(t) = 0$ pour tout t, et donc $u(t, \cdot)$ est nulle sur $[a + \alpha t, b + \beta t]$. ■

On peut utiliser ce que l'on vient de faire pour résoudre un problème différent bien que similaire, sous forme conservative :

$$\begin{cases} \partial_t u + \partial_x\big(c(x)u\big) = 0, \\ u(t = 0, x) = u_0(x), \end{cases} \tag{III.9}$$

Théorème III.6

Si c est suffisamment régulière, le problème (III.9) *admet une unique solution donnée par*

$$u(t, x) = u_0(X(0, t, x))\frac{c(X(0, t, x))}{c(x)}, \quad \textit{si } c(x) \neq 0,$$

et

$$u(t, x) = u_0(x)e^{-tc'(x)}, \quad \textit{si } c(x) = 0.$$

Preuve :

Pour le premier cas, il suffit de remarquer que $v = c(x)u$ vérifie le problème non-conservatif et donc v est constante sur les caractéristiques.

Pour le second cas, la caractéristique issue de x étant constante, la fonction $t \mapsto u(t, x)$ vérifie une équation différentielle ordinaire en temps que l'on résout explicitement. ■

III Schémas numériques pour les équations de transport linéaires

On ne s'intéresse pour l'intant qu'au cas du domaine spatial égal à $\mathbb{R}$ tout entier pour éviter les problèmes liés au bord (voir TP).

On souhaite calculer (ou en tout cas approcher) la solution $u(t, x)$ d'une équation aux dérivées partielles. Bien entendu, un ordinateur ne peut pas coder une telle fonction dans son ensemble, il est donc nécessaire de **discrétiser** c'est-à-dire de représenter seulement une partie finie de la solution. (Dans le cas d'un problème sans bord on va traiter un problème dénombrable mais c'est pour l'instant une vue de l'esprit).

La plus grande partie de cette section sera consacrée à la discrétisation de l'équation linéaire à coefficient constant (III.3) avec une vitesse $c > 0$ (le cas $c = 0$ est peu intéressant ; le signe de la vitesse joue un rôle dans ce qui suit mais le cas $c < 0$ se traite de façon symétrique).

III.1 Introduction aux différences finies

La construction de schémas numérique par la méthode *des différences finies* repose sur l'utilisation de développements de Taylor, ce qui suppose (implicitement) que la solution exacte est régulière. Néanmoins, les schémas peuvent être utilisés sur des cas non réguliers avec un succès parfois mitigé.

On se donne donc un pas de discrétisation spatiale Δx (qu'on note très souvent h dans la littérature) et un pas de discrétisation temporelle Δt (aussi souvent noté k dans la littérature). Ces deux pas de discrétisation étant fixés, on considère la famille de points $(x_i = i\Delta x)_{i\in\mathbb{Z}}$ qui discrétisent le domaine spatial $\mathbb{R}$ et la famille de points $(t^n = n\Delta t)_{n\in\mathbb{N}}$ qui discrétisent la variable temporelle.

Ceci étant fait, on peut maintenant espérer représenter dans un ordinateur (dénombrable !) l'ensemble des valeurs

$$(u(t^n, x_i))_{i\in\mathbb{Z}, n\in\mathbb{N}}.$$

Si on trouve une méthode pour calculer de façon exacte ces valeurs nous aurons bien entendu complètement résolu le problème. Malheureusement il est impossible de remplacer toute l'information contenue dans l'EDP (qui est valable en tout instant t et en tout point x) par un nombre dénombrable de relations concernant seulement un *échantillon* de la solution.

Il faut donc trouver un moyen de traduire sur les valeurs discrètes l'équation aux dérivées partielles à laquelle on s'intéresse.

La méthode des différences finies consiste à remarquer que, si la fonction u est suffisamment régulière, alors on peut construire des quotients différentiels à partir des valeurs échantillonnées qui ne sont pas trop loin des dérivées partielles de la fonction en un point (t^n, x_i) de la discrétisation. Ainsi nous avons par exemple

$$\frac{u(t^n, x_{i+1}) - u(t^n, x_i)}{\Delta x} \approx \partial_x u(t^n, x_i),$$

$$\frac{u(t^n,x_{i+1})-u(t^n,x_{i-1})}{2\Delta x}\approx\partial_x u(t^n,x_i),$$
$$\frac{u(t^n,x_i)-u(t^n,x_{i-1})}{\Delta x}\approx\partial_x u(t^n,x_i),$$
$$\frac{u(t^{n+1},x_i)-u(t^n,x_i)}{\Delta t}\approx\partial_t u(t^n,x_i),$$
$$\frac{u(t^n,x_i)-u(t_{n-1},x_i)}{\Delta t}\approx\partial_t u(t^n,x_i),$$
$$\frac{u(t^{n+1},x_i)-u(t_{n-1},x_i)}{2\Delta t}\approx\partial_t u(t^n,x_i).$$

Il sera nécessaire de préciser plus tard ce que l'on entend exactement par le signe $\approx$.

Comme la fonction u que l'on cherche est solution de l'équation aux dérivées partielles au point (t^n,x_i), on obtient (en choisissant l'une des possibilités de combinaison des formules ci-dessus) le résultat suivant

$$\frac{u(t^n,x_i)-u(t_{n-1},x_i)}{\Delta t}+c\frac{u(t^n,x_{i+1})-u(t^n,x_i)}{\Delta x}\approx\partial_t u(t^n,x_i)+c\partial_x u(t^n,x_i)\approx 0.$$

Si ce terme était exactement nul, alors on aurait trouvé des relations liant uniquement les valeurs de u aux points de discrétisation et le tour serait joué. Malheureusement, sauf dans des cas très particuliers, ce calcul ne donne pas exactement 0. Il donne un résultat petit (c'est-à-dire *grosso modo* de la taille des pas Δt et Δx au pire) que l'on appelle **l'erreur de consistance**.

A ce stade, nous n'avons pas écrit de schéma numérique !

Ecrire un schéma numérique par la méthode des différences finies consiste à dire : l'erreur de consistance n'est pas nulle mais est-ce que je peux trouver une famille de nombres réels notée (u_i^n) dont on espère qu'ils seront proches des vraies valeurs inconnues $(u(t^n,x_i))$ qui satisfont **exactement** la relation discrète ci-dessus c'est-à-dire dans l'exemple choisi :

$$\frac{u_i^n-u_i^{n-1}}{\Delta t}+c\frac{u_{i+1}^n-u_i^n}{\Delta x}=0. \tag{III.10}$$

Bien entendu, la donnée initiale du problème doit intervenir quelque part. Elle intervient tout simplement en disant que, puisque je connais toutes les valeurs de u à l'instant 0 (c'est la fonction donnée u_0 !), je vais demander à ce que les u_i^0 soient donnés exactement par les valeurs de la donnée initiale.

$$u_i^0=u(0,x_i)=u_0(x_i),\ \ \forall i\in\mathbb{Z}. \tag{III.11}$$

Les grandes questions de l'analyse numérique des EDP :

- **Schéma bien posé :** Le schéma (III.10) (c'est-à-dire la famille d'équations (III.10)) assorti des conditions initiales (III.11) admet-il une solution ? Est-elle unique ? Sous quelles conditions sur les données du problème et/ou sur les paramètres de discrétisation ?
- **Consistance :** Quel est le sens exact des signes ≈ 0 dans les formules de dérivées numériques plus haut ?
- **Borne - Stabilité :** Supposons que la solution du schéma existe et est unique. Celle-ci dépend bien entendu des paramètres de discrétisation Δt et Δx. Est-ce que la solution discrète obtenue est bornée quand Δt et Δx tendent vers 0 ? Pour quelles normes ?
- **Convergence :** A-t'on convergence de la solution approchée ? Autrement dit a-t'on u_i^n aussi proche que l'on veut de $u(t^n,x_i)$ si Δt et Δx sont assez petits ? Sous quelles conditions sur Δt et Δx ?
- **Estimation de l'erreur :** Peut-on estimer *a priori* l'erreur de convergence c'est-à-dire la taille des nombres $(u_i^n-u(t^n,x_i))_{i,n}$ en fonction de Δt et Δx ?

III.2 Schémas en temps pour le transport

Revenons à nos moutons et à l'équation de transport. Il s'agit d'une équation d'évolution pour laquelle il est naturel de comprendre le schéma comme une relation de récurrence sur la variable de temps discret n. Il est commode d'introduire pour tout n le vecteur $U^n=(u_i^n)_{i\in\mathbb{Z}}$ qui représente *la solution approchée au temps* n.

On distingue alors deux catégories de schémas en temps :

- Ou bien chaque u_i^n s'exprime explicitement en fonction des u_j^k avec $j\in\mathbb{Z}$ et $k<n$, on dit alors que le schéma est **explicite en temps**,
- Ou bien pour trouver les (u_i^n), il est nécessaire de résoudre une ou plusieurs équations (linéaires ou non-linéaires), on dit alors que le schéma est **implicite en temps**.

Dans l'exemple qui nous occupe, dans les différentes formules proposées, on a vu que la dérivée en espace en t^n est approchée par une formule de différences divisées qui ne fait intervenir que des valeurs de u_j^k avec $k = n$, ce terme s'écrit donc formellement sous la forme AU^n ou A est une matrice (ici infinie car on travaille en domaine non borné). Tout dépend donc de la discrétisation de la dérivée en temps.

– La première formule proposée s'écrit en notation vectorielle

$$\frac{U^{n+1} - U^n}{\Delta t} + cAU^n = 0, \tag{III.12}$$

c'est un schéma explicite à 1 pas, on dit que c'est une discrétisation en temps de type **Euler explicite** par analogie avec le schéma correspondant pour les EDO (ici l'EDO sous-jacente est $U' + cAU = 0$).

– La deuxième formule s'écrit

$$\frac{U^n - U^{n-1}}{\Delta t} + cAU^n = 0,$$

il s'agit d'un schéma implicite à 1 pas, on dit que c'est une discrétisation en temps de type **Euler implicite**.

– La dernière formule s'écrit

$$\frac{U^{n+1} - U^{n-1}}{2\Delta t} + cAU^n = 0,$$

il s'agit d'un schéma explicite qui utilise 2 pas de temps, on dit que c'est une discrétisation en temps de type **Saute-mouton**.

Dans la suite de ce chapitre, on va s'intéresser exclusivement au schéma en temps d'Euler explicite pour diverses raisons : tout d'abord ils sont plus aisés à mettre en oeuvre que leurs équivalents implicites (une matrice à inverser) et ont d'assez bonnes propriétés de stabilité qui ne justifient pas, en général (il y a bien sur des contre-exemples), leur utilisation.

Si nécessaire on sera amené à écrire le schéma numérique sous la forme

$$U^{n+1} = SU^n,$$

où $S : \mathbb{R}^{\mathbb{Z}} \mapsto \mathbb{R}^{\mathbb{Z}}$ est un opérateur linéaire. Dans le cas du schéma en temps explicite (III.12) on aura $S = \mathrm{Id} - c\Delta t A$. Il reste bien sûr à choisir l'opérateur A qui représente l'approximation de l'opérateur de dérivation en espace, ce qui est l'objet de ce qui suit.

III.3 Schémas totalement discrets pour le transport

On va maintenant considérer les différents schémas concrètement obtenus par discrétisation en temps d'Euler explicite couplée à l'une des trois formules de discrétisation en espace proposée plus haut. On ne mentionnera plus le fait que l'on utilise Euler explicite en temps, même si en toute rigueur on devrait le faire. Les trois schémas que l'on déduit des formules proposées plus haut s'écrivent de la façon suivante

$$\text{Schéma décentré à gauche :} \quad \frac{u_i^{n+1} - u_i^n}{\Delta t} + c\frac{u_i^n - u_{i-1}^n}{\Delta x} = 0, \tag{SDAG}$$

$$\text{Schéma décentré à droite :} \quad \frac{u_i^{n+1} - u_i^n}{\Delta t} + c\frac{u_{i+1}^n - u_i^n}{\Delta x} = 0, \tag{SDAD}$$

$$\text{Schéma centré :} \quad \frac{u_i^{n+1} - u_i^n}{\Delta t} + c\frac{u_{i+1}^n - u_{i-1}^n}{2\Delta x} = 0, \tag{SC}$$

Dans le cas où $c > 0$, on l'avait fixé au départ et c'est ici que ça intervient, on sait que la solution se propage de la gauche vers la droite c'est pourquoi le schéma décentré à gauche (SDAG) est appelé **schéma décentré amont** alors que le schéma décentré à droite (SDAD) est appelé **schéma décentré aval**. Si $c < 0$ on adopte la dénomination inverse.

De très nombreux autres schémas existent dans la littérature. On pourra par exemple s'intéresser à un schéma un peu exotique, qu'on peut voir comme une variante du schéma centré, appelé **Schéma de Lax-Friedrichs**.

$$\text{Schéma de Lax-Friedrichs :} \quad \frac{u_i^{n+1} - \frac{u_{i+1}^n + u_{i-1}^n}{2}}{\Delta t} + c\frac{u_{i+1}^n - u_{i-1}^n}{2\Delta x} = 0, \tag{LF}$$

III.3.1 Existence et unicité :

Dans le cadre de schémas explicites, l'existence et l'unicité de la solution ne fait aucun doute puisqu'il s'agit juste d'une relation de récurrence (je ne parle pas ici des problèmes - bien entendu fondamentaux - de conditions aux limites).

III.3.2 Consistance des schémas :

Commençons par une définition “avec les mains” :

Définition III.7

Un schéma est dit **consistant avec le problème** (III.3) *si la solution exacte (supposée suffisamment régulière) vérifie le schéma avec un reste qui tend vers* 0 *avec* Δt *et* Δx.

Formalisons cette définition, par exemple pour le schéma décentré à gauche (**cette définition est à adapter en fonction du schéma considéré**).

Définition III.8

On appelle **erreur de consistance** *du schéma décentré à gauche* (SDAG) *les quantités*

$$R_i^n = \frac{u(t^{n+1},x_i)-u(t^n,x_i)}{\Delta t} + c\frac{u(t^n,x_i)-u(t^n,x_{i-1})}{\Delta x},\ \ i\in\mathbb{Z}, n\in\mathbb{N}$$

où u *est une solution exacte de l'équation* (III.3).
On dit que le schéma est d'ordre p *en temps et* q *en espace pour une certaine norme* $\|\cdot\|$ *sur* $\mathbb{R}^{\mathbb{Z}}$ *si, pour tout* $T>0$*, il existe* $M_T>0$ *dépendant seulement de* u *et de* T*, telle que pour tous* Δt *et* Δx *assez petits on a*

$$\sup_{n\le \frac{T}{\Delta t}} \|R^n\| \le M_T(\Delta t^p + \Delta x^q). \tag{III.13}$$

Si on note $\tilde{U}^n = (u(t^n,x_i))_{i\in\mathbb{Z}}$, le vecteur des valeurs exactes de la solution au temps t^n sur les points du maillage, alors l'erreur de consistance R^n s'écrit aussi avec la notation introduite à la fin du précédent paragraphe

$$R^n = \frac{\tilde{U}^{n+1} - S\tilde{U}^n}{\Delta t}.$$

Remarque essentielle : La notion de consistance d'un schéma est relative à un problème donné ! Cette notion dépend de plus de la régularité (supposée) de la solution.

On peut maintenant traiter les exemples. Pour estimer l'erreur de consistance pour un schéma donné, on doit utiliser des développements de Taylor des différents termes en un seul et même point puis utiliser l'équation vérifiée par la fonction u (ou une conséquence de celle-ci). Tous calculs faits on obtient

– Pour les schémas décentrés on trouve

$$R_i^n = c(c\Delta t \pm \Delta x)u_{xx}(t^n,x_i) + O(\Delta t^2+\Delta x^2),$$

donc les schémas sont d'ordre 1 en temps et en espace (sauf si $c\Delta t = \pm\Delta x$).

– Pour le schéma centré on trouve

$$R_i^n = c^2\Delta t u_{xx}(t^n,x_i) + c\frac{\Delta x^2}{6}u_{xxx}(t^n,x_i) + O(\Delta t^2+\Delta x^3),$$

le schéma est donc d'ordre 1 en temps et 2 en espace.

– Pour le schéma de Lax-Friedrichs, on obtient

$$R_i^n = c^2\Delta t u_{xx}(t^n,x_i) + c\frac{\Delta x^2}{6}u_{xxx}(t^n,x_i) + \frac{\Delta x^2}{\Delta t}u_{xx}(t^n,x_i) + O\left(\Delta t^2+\Delta x^3+\frac{\Delta x^4}{\Delta t}\right).$$

On voit donc que ce schéma va bien se comporter seulement si Δt et Δx sont en proportion. Car alors le terme $\frac{\Delta x^2}{\Delta t}$ va se comporter en Δt. On ne peut donc pas prendre Δt trop petit par rapport à Δx ce qui fait que ce schéma se comporte globalement au premier ordre.

Aucun de ces schémas n'est donc d'ordre 2 en temps et en espace. On verra plus loin comment construire un tel schéma.

Cette définition s'adapte à tous les schémas précédents en adaptant la formule de R_i^n. En général on utilise la norme L^∞, la norme L^2 et parfois la norme L^1. De façon plus précise, ces normes sont définies de la façon suivante

$$\|U\|_p = \left(\sum_i \Delta x|u_i|^p\right)^{\frac{1}{p}}, \tag{III.14}$$

le facteur Δx étant là pour tenir compte de la mesure des mailles.

ATTENTION : en général on ne précise pas la dépendance en Δx dans la notation.

Avec cette définition, on obtient le résultat suivant qui justifie le choix particulier de cette norme

Lemme III.9

Pour toute fonction f suffisamment régulière et à support compact, on note $F = (f(x_i))_{i\in\mathbb{Z}} \in \mathbb{R}^{\mathbb{Z}}$. On a alors

$$\|F\|_p \longrightarrow \|f\|_{L^p(\mathbb{R})}, \quad \textit{quand } \Delta x \to 0.$$

Preuve :

On note ΠF la fonction constante par morceaux définie par

$$\Pi F = \sum_{i\in\mathbb{Z}} \mathbf{1}_{M_i} F_i = \sum_{i\in\mathbb{Z}} \mathbf{1}_{M_i} f(x_i).$$

On obtient immédiatement $\|F\|_p = \|\Pi F\|_{L^p(\mathbb{R})}$. Le résultat se déduit donc du théorème de convergence dominée de Lebesgue. ■

III.3.3 Définition générale de la stabilité :

Là encore nous allons commencer par définir la notion de stabilité avec les mains.

Définition III.10

Un schéma pour le problème de transport qu'on notera $S : U^n \mapsto U^{n+1} = SU^n$ est dit stable pour une norme $\|\cdot\|$ si l'amplification des erreurs par le schéma dans le domaine temporel choisi (i.e. dans $[0,T]$) reste bornée quand Δt et Δx tendent vers 0.
Cette propriété peut n'être vraie que sous certaines conditions liant Δt et Δx, on dira alors que le schéma est conditionellement stable.

Et maintenant une définition plus précise :

Définition III.11

Soit $\mathcal{R}$ une partie du quart-de-plan ouvert $\mathbb{R}_*^+ \times \mathbb{R}_*^+$ tel que $(0,0) \in \overline{\mathcal{R}}$.
Un schéma **linéaire** *S est dit stable pour la norme $\|\cdot\|$ dans la région $\mathcal{R}$ si et seulement s'il existe $M_T > 0$ dépendant seulement de $\mathcal{R}$ et des données du problème (i.e. de T, des coeffs de l'équation, etc ...) telle que : pour tout Δx et Δt suffisament petits et tels que $(\Delta t, \Delta x) \in \mathcal{R}$, on ait*

$$\forall U^0 \in \mathbb{R}^{\mathbb{Z}}, \ \ \forall n \leq \frac{T}{\Delta t}, \ \ \|S^n U^0\| \leq M_T \|U^0\|. \tag{III.15}$$

Remarque III.12

- *La condition $(0,0) \in \overline{\mathcal{R}}$ signifie simplement qu'il est possible de faire tendre Δt et Δx vers 0 tout en maintenant le couple $(\Delta t, \Delta x)$ dans la région de stabilité. C'est bien entendu la condition* sine qua non *d'utilisabilité d'un schéma. En effet, le schéma n'est censé approcher correctement la solution exacte que pour des pas de temps et d'espace de plus en plus petits.*
- *Un schéma est d'autant plus stable qu'il a des grandes régions de stabilité. Si $\mathcal{R} = \mathbb{R}_*^+ \times \mathbb{R}_*^+$, on dit que le schéma est inconditionnellement stable.*
- *S'il n'existe pas de telle région de stabilité $\mathcal{R}$, on dit que le schéma est inconditionnellement instable. Dans tous les autres cas, le schéma est dit conditionnellement stable, la condition de stabilité étant la condition d'appartenance du couple $(\Delta t, \Delta x)$ à la région de stabilité $\mathcal{R}$.*

Remarque essentielle : La notion de stabilité dépend de la norme choisie mais ne dépend pas du problème que l'on cherche à résoudre. Il s'agit seulement d'une propriété sur la relation de récurrence qui définit le schéma.

III.3.4 Stabilité L^∞. Monotonie

Définition et Proposition III.13

Soit un schéma linéaire $S : \mathbb{R}^{\mathbb{Z}} \mapsto \mathbb{R}^{\mathbb{Z}}$ fixé. Les propositions suivantes sont équivalentes :

1. *$U \geq 0$ implique $SU \geq 0$.*
2. *$U \geq V$ implique $SU \geq SV$.*

Si ces propriétés sont vraies, on dit que le schéma est monotone (ou positif).

Preuve :

Comme le schéma est linéaire l'équivalence entre 1) et 2) est claire. ■

L'intérêt de cette notion est double :

- Tout d'abord, la positivité de la solution du schéma pour une donnée initiale positive est une propriété vérifiée par la solution exacte d'un problème de transport (on sait que $u(t,x) = u_0(X(0,t,x))$ et donc $u_0 \geq 0 \Rightarrow u \geq 0$).
- Deuxièmement, comme on va le voir, cette propriété implique la stabilité du schéma.

Le premier résultat traite le cas des schémas qui préservent les constantes. On note $D = (1)_{i\in\mathbb{Z}}$ le vecteur constant égal à 1.

Proposition III.14

Soit S un schéma monotone (pour un certain choix de $(\Delta t, \Delta x)$) et qui préserve les constantes, c'est-à-dire tel que $SD = D$. Alors on a

$$\forall U \in \mathbb{R}^{\mathbb{Z}}, \ \|SU\|_\infty \leq \|U\|_\infty,$$

et donc, en particulier,

Un tel schéma L^∞-stable pour ces valeurs de $(\Delta t, \Delta x)$.

Preuve :

Par définition de la norme infinie d'un vecteur on a

$$-\|U\|_\infty D \leq U \leq \|U\|_\infty D,$$

puis on appplique l'opérateur S en utilisant la monotonie, et la préservation des constantes

$$-\|U\|_\infty D \leq SU \leq \|U\|_\infty D,$$

ce qui donne maintenant

$$\|SU\|_\infty \leq \|U\|_\infty.$$

■

On a en fait un résultat un peu plus général qui permet de traiter d'autres équations que la simple équation de transport à vitesse constante.

Proposition III.15

Si S est un schéma monotone, et s'il existe un nombre $M > 0$ tel que $\|SD - D\|_\infty \leq M\Delta t$, alors

$$\forall U \in \mathbb{R}^{\mathbb{Z}}, \quad \sup_{n \leq \frac{T}{\Delta t}} \|S^n U\|_\infty \leq e^{MT} \|U\|_\infty.$$

Si le nombre M est indépendant de Δt et Δx, cela montre bien la stabilité L^∞ d'un tel schéma.

Preuve :

On reprend la démonstration précédente, on obtient

$$-\|U\|_\infty (D + (SD - D)) \leq SU \leq \|U\|_\infty (D + (SD - D)),$$

et donc

$$\|SU\|_\infty \leq (1 + M\Delta t)\|U\|_\infty,$$

ainsi

$$\|S^n U\|_\infty \leq (1 + M\Delta t)^n \|U\|_\infty \leq e^{Mn\Delta t} \|U\|_\infty \leq e^{MT} \|U\|_\infty.$$

■

On peut aisément reconnaître un schéma numérique linéaire monotone qui préserve les constantes de la façon suivante :

Proposition III.16

Soient Δt et Δx donnés. Le schéma S est monotone et préserve les constantes si et seulement si pour tout n et pour tout i, la valeur u_i^{n+1} s'écrit comme combinaison **convexe** *des $(u_j^n)_{j\in\mathbb{Z}}$.*
En particulier un tel couple $(\Delta t, \Delta x)$ appartient à la région de stabilité L^∞ du schéma.

Preuve :

Le sens direct est clair. Pour le sens inverse il suffit de prendre un U particulier qui vaut 0 partout sauf en un point. ■

Appliquons maintenant ces définitions et résultats aux schémas traités ici en exemple.

Théorème III.17

– *Le schéma* (SDAG) *est monotone si et seulement si*

$$0 \leq \frac{c\Delta t}{\Delta x} \leq 1.$$

Sous cette condition, le schéma est donc L^∞ stable.

– *Le schéma* (SDAD) *est monotone si et seulement si*

$$-1 \leq \frac{c\Delta t}{\Delta x} \leq 0.$$

Sous cette condition le schéma est donc L^∞-stable.

– *Le schéma* (SC) *est monotone si et seulement si $c = 0$. Donc dans tous les cas non triviaux ce schéma est inconditionnellement L^∞-instable.*

– *Le schéma* (LF) *est monotone si et seulement si*

$$|c|\frac{\Delta t}{\Delta x} \leq 1.$$

Les conditions ci-dessus sont quasiment optimales pour la stabilité L^∞ des schémas, c'est pourquoi on dira que ce sont **les conditions de stabilité du schéma**. *On les appelle aussi conditions CFL (Courant-Friedrichs-Lewy).*

Preuve :

Il suffit d'écrire u_i^{n+1} comme combinaison linéaire des $(u_j^n)_j$ et de vérifier si tous les coefficients sont positifs.

– Schéma décentré à gauche

$$u_i^{n+1} = \left(1 - c\frac{\Delta t}{\Delta x}\right) u_i^n + c\frac{\Delta t}{\Delta x} u_{i-1}^n.$$

– Schéma décentré à droite

$$u_i^{n+1} = \left(1 + c\frac{\Delta t}{\Delta x}\right) u_i^n - c\frac{\Delta t}{\Delta x} u_{i+1}^n.$$

– Schéma centré

$$u_i^{n+1} = u_i^n - c\frac{\Delta t}{2\Delta x} u_{i+1}^n + c\frac{\Delta t}{2\Delta x} u_{i-1}^n.$$

– Schéma de Lax-Friedrichs

$$u_i^{n+1} = \frac{1}{2}\left(1 - c\frac{\Delta t}{\Delta x}\right) u_{i+1}^n + \frac{1}{2}\left(1 + c\frac{\Delta t}{\Delta x}\right) u_{i-1}^n.$$

■

La constatation essentielle c'est que le schéma centré naïf n'est jamais stable (sauf pour $c = 0$ qui n'a pas grand intérêt) alors que les schémas décentrés le sont. Pour cela il faut que le décentrement ait lieu à gauche pour des vitesses positives et à droite pour des vitesses négatives. On parle alors de **décentrement amont**. Il faut de plus vérifier la condition de stabilité

$$|c|\frac{\Delta t}{\Delta x} \leq 1,$$

appelée condition CFL (pour Courant-Friedrichs-Levy). Cette condition s'interprète géométriquement en termes de caractéristiques.

III.3.5 Stabilité L^2 par la méthode de Von Neumann

Il se trouve que la notion de stabilité L^∞, bien que très utile quand elle est vérifiée, n'est pas suffisante pour étudier certains schémas, surtout pour les schémas qui ne sont pas monotones.

C'est pourquoi on est amenés à aussi considérer la stabilité des schémas au sens de la norme L^2. Pour cela on va voir deux méthodes pour faire cette étude. La première est relativement systématique mais limitée dans son champ d'application. La seconde (paragraphe suivant) est beaucoup plus générale et puissante mais, dans un premier temps, un peu plus délicate à utiliser.

Définition III.18 (Facteur d'amplification)

Soit $S : \mathbb{R}^{\mathbb{Z}} \mapsto \mathbb{R}^{\mathbb{Z}}$ un schéma linéaire qui s'écrit sous la forme

$$(SU)_i = \sum_{k\in\mathbb{Z}} \alpha_k U_{i+k}, \quad \forall i \in \mathbb{Z}, \forall U \in \mathbb{R}^{\mathbb{Z}}, \tag{III.16}$$

où les $(\alpha_k)_k$ ne dépendent pas de i et sont presque tous nuls. En revanche, les α_k dépendent des paramètres du schéma (la vitesse c, les pas Δt et Δx, etc ...).
On appelle facteur d'amplification (à la fréquence $\xi \in \mathbb{R}$), le nombre complexe $a(\xi)$ défini par

$$a(\xi) = \sum_{k\in\mathbb{Z}} \alpha_k e^{k\sqrt{-1}\xi}.$$

Celui-ci dépend bien sûr des paramètres du schéma.

Cette définition du facteur d'amplification peut se justifier par le résultat suivant

Proposition III.19

Soit S donné par (III.16) *et $\xi \in \mathbb{R}$. Si on applique le schéma à un vecteur $U_\xi \in \mathbb{C}^{\mathbb{Z}}$ de la forme $U_\xi = (e^{i\sqrt{-1}\xi})_i$, alors on obtient*

$$SU_\xi = a(\xi)U_\xi.$$

L'action de S sur U_ξ est donc juste une amplification de facteur $a(\xi)$.

Preuve :

C'est un simple calcul

$$(SU_\xi)_i = \sum_{k\in\mathbb{Z}} \alpha_k (U_\xi)_{i+k} = \sum_{k\in\mathbb{Z}} \alpha_k e^{(i+k)\sqrt{-1}\xi} = \left(\sum_{k\in\mathbb{Z}} \alpha_k e^{k\sqrt{-1}\xi}\right) e^{i\sqrt{-1}\xi} = a(\xi)(U_\xi)_i.$$

■

Théorème III.20 (Stabilité au sens de Von Neumann)

Soit un schéma linéaire S sous la forme (III.16)*. Si on a la condition*

$$|a(\xi)| \leq 1, \quad \forall \xi \in \mathbb{R},$$

alors le schéma S est L^2-stable (pour les valeurs considérées des paramètres !). On a même la propriété plus précise

$$\|SU\|_2 \leq \|U\|_2, \quad \forall U \in \mathbb{R}^{\mathbb{Z}}.$$

Preuve :

Soit $U \in l^2(\mathbb{Z})$ que l'on identifie à la fonction constante par morceaux

$$\Pi U = \sum_{i\in\mathbb{Z}} U_i \mathbf{1}_{M_i} \quad \in L^2(\mathbb{R}),$$

où $M_i =]x_i - \Delta x/2, x_i + \Delta x/2[$ est l'intervalle de taille Δx centré en x_i. Par définition de la norme L^2 discrète (formule (III.14)), on a

$$\|U\|_2 = \|\Pi U\|_{L^2}. \tag{III.17}$$

D'après (III.16), on constate que

$$\Pi(SU) = \sum_{i\in\mathbb{Z}} (SU)_i \mathbf{1}_{M_i} = \sum_{i\in\mathbb{Z}}\sum_{k\in\mathbb{Z}} \alpha_k U_{i+k}\mathbf{1}_{M_i} = \sum_{i\in\mathbb{Z}}\sum_{k\in\mathbb{Z}} \alpha_k U_{i+k}\tau_{-k\Delta x}\left[\mathbf{1}_{M_{i+k}}\right]$$
$$= \sum_{k\in\mathbb{Z}} \alpha_k \tau_{-k\Delta x}\left(\sum_{i\in\mathbb{Z}} U_{i+k}\mathbf{1}_{M_{i+k}}\right) = \sum_{k\in\mathbb{Z}} \alpha_k \tau_{-k\Delta x}\Pi U.$$

On rappelle que les opérateurs de translation τ_h sont définis par la formule (III.25).

Appliquons la transformée de Fourier (voir les rappels dans la section V)

$$\mathcal{F}(\Pi(SU)) = \sum_{k\in\mathbb{Z}} \alpha_k \mathcal{F}\left(\tau_{-k\Delta x}\Pi U\right).$$

Il vient, avec la propriété (III.26),

$$\mathcal{F}(\Pi(SU))(\xi) = \left(\sum_{k\in\mathbb{Z}} \alpha_k e^{k\Delta x\sqrt{-1}\xi}\right)\mathcal{F}(\Pi U)(\xi).$$

Si on reprend la définition du coefficient d'amplification, on constate qu'on a en fait montré

$$\mathcal{F}(\Pi(SU))(\xi) = a(\xi\Delta x)\mathcal{F}(\Pi U)(\xi).$$

Appliquons maintenant le théorème de Plancherel (Théorème III.38) qui dit que la transformée de Fourier est une isométrie de L^2, on obtient

$$\|\Pi(SU)\|_{L^2} = \frac{1}{\sqrt{2\pi}}\|\mathcal{F}(\Pi(SU))\|_{L^2} \leq \underbrace{\left(\sup_{\xi\in\mathbb{R}}|a(\xi\Delta x)|\right)}_{=\sup_\xi|a(\xi)|}\frac{\|\mathcal{F}(\Pi U)\|_{L^2}}{\sqrt{2\pi}} = (\sup_\xi|a(\xi)|)\|\Pi U\|_{L^2}.$$

Enfin, avec (III.17), il vient

$$\|SU\|_2 \leq (\sup_\xi|a(\xi)|)\|U\|_2.$$

On trouve bien que si $|a(\xi)| \leq 1$ pour tout $\xi \in \mathbb{R}$, alors l'action du schéma fait décroître la norme L^2 et donc en particulier, le schéma est L^2-stable. ■

En adaptant très légèrement la démonstration ci-dessus, on peut montrer le résultat un peu plus général suivant (à rapprocher de la Proposition III.15).

Corollaire III.21

Soit S de la forme (III.16). *On suppose qu'il existe un nombre $M > 0$ (indépendant de Δt et Δx) tel que*

$$\sup_{\xi\in\mathbb{R}}|a(\xi)| \leq 1 + M\Delta t,$$

alors le schéma S vérifie la propriété de L^2-stabilité suivante

$$\sup_{n\leq\frac{T}{\Delta t}}\|S^n U_0\|_2 \leq e^{MT}\|U_0\|_2, \quad \forall T > 0, \forall U_0 \in \mathbb{R}^{\mathbb{Z}}.$$

Les conditions obtenues ci-dessus sont essentiellement optimales et en pratique c'est celles qu'on utilise. L'approche de Von-Neumann permet de ramener l'analyse de stabilité L^2 à un calcul du coefficient d'amplification et à son étude (qui peut ne pas être triviale).

Son **inconvénient majeur** est qu'elle est strictement réservée au cas des équations à coefficients constants, c'est-à-dire des schémas qui s'écrivent sous la forme (III.16).

Dans le cas qui nous occupe la méthode ne fonctionne absolument pas si la vitesse c dépend de x (ou de t). De même elle ne fonctionne que pour une discrétisation uniforme dans la variable x. C'est donc un bon outil mais très limité dans

son utilisation pratique.

Théorème III.22 (Stabilité L^2 pour le transport)

Pour les trois premiers schémas considérés au théorème III.17, les conditions de stabilité obtenues dans ce théorème sont aussi les conditions de stabilité L^2. Autrement dit :

- *Le schéma décentré à gauche est L^2-stable si $0 \leq c\frac{\Delta t}{\Delta x} \leq 1$.*
- *Le schéma décentré à droite est L^2-stable si $-1 \leq c\frac{\Delta t}{\Delta x} \leq 0$.*
- *Le schéma centré est inconditionnellement L^2-instable.*
- *Le schéma de Lax-Friedrichs est L^2-stable si $|c|\frac{\Delta t}{\Delta x} \leq 1$.*

Preuve :

Je traite ici seulement le cas du schéma décentré à gauche, les autres cas seront traités en TD.

Le schéma (SDAG) s'écrit sous la forme

$$u_i^{n+1} = \left(1 - \frac{c\Delta t}{\Delta x}\right) u_i^n + \frac{c\Delta t}{\Delta x} u_{i-1}^n,$$

soit encore, en introduisant le **nombre CFL** $\nu = \frac{c\Delta t}{\Delta x}$,

$$u_i^{n+1} = (1-\nu)u_i^n + \nu u_{i-1}^n.$$

Autrement dit, le schéma est bien sous la forme (III.16) avec $\alpha_0 = 1-\nu$, $\alpha_{-1} = \nu$ et tous les autres α_k nuls.

On trouve alors

$$a(\xi) = (1-\nu) + \nu e^{-\sqrt{-1}\xi},$$

puis, en multipliant par le conjugué :

$$|a(\xi)|^2 = \left((1-\nu) + \nu e^{-\sqrt{-1}\xi}\right)\left((1-\nu) + \nu e^{\sqrt{-1}\xi}\right) = (1-\nu)^2 + \nu^2 + 2\nu(1-\nu)\cos(\xi).$$

On voit donc que $|a(\xi)|^2$ est maximal pour $\cos(\xi) = \pm 1$, donc

$$\sup_\xi |a(\xi)|^2 = \max\Big((1-\nu)^2 + \nu^2 + 2\nu(1-\nu), (1-\nu)^2 + \nu^2 - 2\nu(1-\nu)\Big),$$

ce qui donne

$$\sup_\xi |a(\xi)|^2 = \max((1-2\nu)^2, 1).$$

Cette quantité est inférieure ou égale à 1 si et seulement si $(1-2\nu)^2 \leq 1$, c'est-à-dire si et seulement si $0 \leq \nu \leq 1$, ce qui montre le résultat. ■

III.3.6 Stabilité L^2 par la méthode d'énergie

A cause des limitations précédemment évoquées de la notion de stabilité au sens de Von Neumann, il est bon de savoir démontrer la stabilité L^2 des différents schémas par une technique dite **d'estimation d'énergie** qui est plus générale mais moins systématique. Elle est donc plus difficile à mettre en oeuvre dans les cas concrets.

On va se contenter de faire le calcul pour le schéma décentré amont pour $c > 0$ (i.e. le schéma décentré à gauche). Pour cela, on va supposer que la donnée initiale du schéma est à support compact (i.e. $U_i^0 = 0$ pour $|i|$ assez grand), ceci a pour effet que, pour tout $n \geq 0$, U^n est aussi à support compact.

La formule algébrique élémentaire (qui en gros revient à faire une intégration par parties discrète) à retenir est la suivante

$$(a-b)a = \frac{1}{2}a^2 - \frac{1}{2}b^2 + \frac{1}{2}(a-b)^2, \tag{III.18}$$

et donc bien sûr on a aussi

$$(b-a)a = \frac{1}{2}b^2 - \frac{1}{2}a^2 - \frac{1}{2}(a-b)^2. \tag{III.19}$$

– On prend le schéma au point i à l'instant n et on le multiplie par u_i^n :

$$(u_i^{n+1} - u_i^n)u_i^n + c\frac{\Delta t}{\Delta x}(u_i^n - u_{i-1}^n)u_i^n = 0.$$

– On utilise les formules algébriques (III.18)-(III.19) pour obtenir

$$\frac{1}{2}(u_i^{n+1})^2 - \frac{1}{2}(u_i^n)^2 - \frac{1}{2}(u_i^{n+1}-u_i^n)^2 + c\frac{\Delta t}{\Delta x}\left(\underbrace{\frac{1}{2}(u_i^n)^2 - \frac{1}{2}(u_{i-1}^n)^2}_{=D_i} + \frac{1}{2}(u_i^n - u_{i-1}^n)^2\right) = 0.$$

– On multiplie par Δx (mesure des mailles) pour faire apparaître des normes L^2 et on somme sur i, à n fixé. La contribution des termes notés D_i est nulle par un phénomène de somme télescopique (et car on a pris des données à support compact). Il reste donc

$$\frac{1}{2}\|U^{n+1}\|_{L^2}^2 - \frac{1}{2}\|U^n\|_{L^2}^2 - \underbrace{\frac{1}{2}\sum_{i\in\mathbb{Z}}\Delta x(u_i^{n+1}-u_i^n)^2}_{=T_1} + \frac{1}{2}c\frac{\Delta t}{\Delta x}\underbrace{\sum_{i\in\mathbb{Z}}\Delta x(u_i^n - u_{i-1}^n)^2}_{=T_2} = 0.$$

– Quelle est la situation ? On souhaite montrer (sous condition CFL) que $\|U^{n+1}\|_{L^2}^2 \leq \|U^n\|_{L^2}^2$. De ce point de vue, le terme T_2 a le bon signe mais le terme T_1 a le mauvais signe. En effet, si il y avait un signe $+$ devant le terme T_1, alors on aurait l'inégalité cherchée et donc le schéma serait L^2-stable sans condition !

  Pour continuer le calcul, il faut se rappeller que les u_i^n sont des solutions du schéma et que donc les termes T_1 et T_2 ne sont pas indépendants. Plus précisément, le schéma s'écrit

$$u_i^{n+1} - u_i^n = -c\frac{\Delta t}{\Delta x}(u_i^n - u_{i-1}^n),$$

  en élevant au carré, en multipliant par h et en sommant, on trouve

$$\sum_{i\in\mathbb{Z}}\Delta x(u_i^{n+1}-u_i^n)^2 = c^2\frac{\Delta t^2}{\Delta x^2}\sum_{i\in\mathbb{Z}}\Delta x(u_i^n - u_{i-1}^n)^2,$$

  autrement dit, on a montré

$$T_1 = c^2\frac{\Delta t^2}{\Delta x^2}T_2.$$

  En reportant ceci dans l'égalité ci-dessus (dans laquelle on a enlevé les coeffs $1/2$ partout), on obtient

$$\|U^{n+1}\|_{L^2}^2 - \|U^n\|_{L^2}^2 + c\frac{\Delta t}{\Delta x}\left(1 - c\frac{\Delta t}{\Delta x}\right)T_2 = 0.$$

  Le terme T_2 étant positif, on voit que l'inégalité d'énergie discrète aura lieu si et seulement si on a

$$c\frac{\Delta t}{\Delta x}\left(1 - c\frac{\Delta t}{\Delta x}\right) \geq 0,$$

  ce qui est exactement équivalent à la condition CFL énoncée plus haut.

Le cas du schéma décentré à droite et strictement identique à un signe près ; en revanche il est instructif de tenter de faire la même preuve pour le schéma centré. On s'apercoit que le terme T_1 dans l'estimation persiste avec un mauvais signe, alors que le terme T_2 est alors identiquement nul ! On a donc aucun espoir d'arriver, même sous condition CFL, à compenser ce mauvais terme. Le schéma est ainsi inexorablement instable !

III.3.7 Convergence et estimation de l'erreur :

Avec les différentes notions vues précédemment, on peut maintenant énoncer et démontrer le résultat de convergence et d'estimation de l'erreur.

Théorème III.23

On considère un schéma numérique linéaire S. On suppose que :

- *Le schéma est consistant à l'ordre p en temps et q en espace pour une certaine norme $\|\cdot\|$ (i.e. on suppose* (III.13) *vérifiée).*
- *Il existe une région de stabilité $\mathcal{R}$ non triviale pour ce schéma pour la même norme $\|\cdot\|$, i.e. on suppose que* (III.15) *est vérifiée.*

Alors, le schéma numérique est convergent à l'ordre p en temps et q en espace pour la norme $\|\cdot\|$, c'est-à-dire que : pour toute donnée initiale u_0 suffisament régulière (et à support compact pour simplifier), si on note u la solution du problème de transport associée et $(u_i^n)_{i\in\mathbb{Z},n\in\mathbb{N}}$ la solution du schéma numérique pour la donnée initiale $U^0 = (u_0(x_i))_{i\in\mathbb{Z}}$ nous avons

$$\sup_{n\leq \frac{T}{\Delta t}} \|E^n\| \leq M_T(\Delta t^p + \Delta x^q), \quad \forall (\Delta t, \Delta x) \in \mathcal{R},$$

où E^n est le vecteur **erreur à l'instant** *n défini par*

$$E^n = (e_i^n)_{i\in\mathbb{Z}}, \quad e_i^n = \widetilde{U}^n - U^n,$$

avec

$$\widetilde{U}^n = (u(t^n, x_i))_{i\in\mathbb{Z}}.$$

Preuve :

La définition de l'erreur de consistance s'écrit avec les notations précédentes

$$R^n = \frac{\widetilde{U}^{n+1} - S\widetilde{U}^n}{\Delta t},$$

on a donc

$$\begin{cases} \widetilde{U}^{n+1} = S\widetilde{U}^n + \Delta t R^n, \ \forall n \geq 0, \\ \widetilde{U}^0 = (u_0(x_i))_{i\in\mathbb{Z}}. \end{cases}$$

Par ailleurs, par définition du schéma on a

$$\begin{cases} U^{n+1} = SU^n, \ \forall n \geq 0, \\ U^0 = (u_0(x_i))_{i\in\mathbb{Z}} = \widetilde{U}^0, \end{cases}$$

et donc par soustraction, nous avons

$$\begin{cases} E^{n+1} = SE^n + \Delta t R^n, \ \forall n \geq 0, \\ E^0 = 0. \end{cases}$$

On démontre maintenant par récurrence la formule

$$E^n = \Delta t \sum_{k=0}^{n-1} S^k R^{n-1-k}.$$

Prenant la norme de cette formule, on trouve

$$\|E^n\| \leq \Delta t \sum_{k=0}^{n-1} \|S^k R^{n-1-k}\|.$$

Supposons $(\Delta t, \Delta x) \in \mathcal{R}$ et utilisons la définition de la stabilité, il vient

$$\|E^n\| \leq \Delta t \sum_{k=0}^{n-1} M_T \|R^{n-1-k}\|,$$

enfin l'estimation de l'erreur de consistance, fournit

$$\|E^n\| \leq \Delta t \sum_{k=0}^{n-1} M_T M_T'(\Delta t^p + \Delta x^q) \leq \left[n\Delta t\right] M_T M_T'(\Delta t^p + \Delta x^q).$$

Il vient donc

$$\sup_{n \leq \frac{T}{\Delta t}} \|E^n\| \leq T M_T M_T'(\Delta t^p + \Delta x^q).$$

■

IV Eléments d'analyse des lois de conservation non-linéaires

Cette partie du chapitre n'a pas été traitée en cours !

On s'intéresse dans ce paragraphe à l'analyse des lois de conservation non-linéaire de la forme

$$\begin{cases} \partial_t u + \partial_x f(u) = 0, \quad \forall t > 0, \forall x \in \mathbb{R}, \\ u(t=0,x) = u_0(x). \end{cases} \tag{III.20}$$

où f est une fonction de classe $\mathcal{C}^1$ et u_0 une donnée initiale bornée.

On renvoie au début du chapitre pour les modèles qui relèvent de ce type de problématique.

L'exemple sur lequel on va s'appuyer en grande partie c'est celui de l'équation de Burgers :

$$\partial_t u + \partial_x(u^2) = 0.$$

Dans le cadre du trafic routier, cette équation est satisfaite par l'inconnue $u = v - \frac{V_{max}}{2}$.

IV.1 Solutions peu régulières du transport linéaire

Question (relativement) naturelle : Peut-on quand même dire que la formule (III.4) ci-dessus donne une solution du problème pour des données peu régulières ? (motivation = le non-linéaire).

Définition III.24 (Formulation faible)

Soit $u_0 \in L^\infty(\mathbb{R})$. On dit qu'une fonction $u \in L^\infty([0,+\infty[\times\mathbb{R})$ est solution **faible** *du problème* (III.3) *si et seulement si on a*

$$\int_0^\infty \int_{\mathbb{R}} u(\partial_t\varphi + c\partial_x\varphi)\, dt\, dx + \int_{\mathbb{R}} u_0(x)\varphi(0,x)\, dx = 0,$$

pour toute fonction $\varphi \in \mathcal{C}_c^\infty([0,+\infty[\times\mathbb{R})$.

Cette définition permet de faire porter toutes les dérivées sur la **fonction-test** φ et a donc parfaitement un sens même pour une fonction u non dérivable. Le lien entre cette définition et la notion de solution "classique" d'une telle équation est le suivant :

Proposition III.25

- *Toute solution régulière (de classe $\mathcal{C}^1$) du problème de transport est aussi une solution faible.*
- *Toute solution faible u qui, de plus, est de classe $\mathcal{C}^1$ est alors solution au sens usuel et vérifie $u(0,x) = u_0(x)$.*

Preuve :

Il s'agit dans les deux cas d'une simple intégration par parties. ■

Remarque : Tout cela fonctionne avec des fonctions L^1_{loc}.

Théorème III.26

Pour toute donnée initiale $u_0 \in L^\infty(\mathbb{R})$, il existe une unique solution faible du problème qui est donnée par

$$u(t,x) = u_0(x - ct).$$

Remarquez que ceci a bien un sens car, même si u_0 et $\tilde{u}_0$ sont égales sauf sur un ensemble de mesure nulle, alors $u = u_0(x-ct)$ et $\tilde{u} = \tilde{u}_0(x-ct)$ sont aussi égales sauf sur un ensemble de mesure nulle. Donc la formule proposée préserve les classes d'équivalence presque partout.

Preuve :

L'existence s'obtient par Fubini, changement de variable et compagnie. Pur montrer l'unicité il suffit de montrer que si u est solution faible avec $u_0 = 0$ alors $u = 0$. **Ceci n'est vrai que parceque le problème est linéaire ! !** (penser à Injectif $\Leftrightarrow$ Noyau trivial).

Pour faire cela, on commence par montrer que pour toute fonction φ régulière à support compact, d'une autre fonction ψ, régulière et à support compact vérifiant

$$\partial_t \psi + c\partial_x \psi = \varphi.$$

Pour cela, on résout par la méthode des caractéristiques et on trouve la formule suivante :

$$\psi(t,x) = -\int_t^\infty \varphi(s, x-c(t-s))\, ds.$$

On vérifie que cette formule définit une fonction $\mathcal{C}^\infty$ à support compact qui vérifie le problème souhaité. Cette formule n'est pas magique ! Elle est obtenue en résolvant une équation différentielle le long des caractéristiques (voir TD).

Une fois que ceci est démontré, on peut prendre la fonction ψ ci-dessus comme fonction test dans la formulation faible (avec $u_0 = 0$ puisqu'on cherche à démontrer l'unicité). On obtient donc la formule

$$\int_{\mathbb{R}^+} \int_{\mathbb{R}} u(t,x)\varphi(t,x)\, dt\, dx = 0. \tag{III.21}$$

Celle-ci est vraie pour toute fonction φ régulière et à support compact. Comme cette famille de fonction est dense dans $L^1(\mathbb{R}^+ \times \mathbb{R})$ et que la quantité qui nous intéresse est continue par rapport à φ pour la topologie de L^1, on en déduit que (III.21) est vraie pour toute fonction φ de L^1. Elle est donc en particulier vraie pour $\varphi = u1_K$ où K est un compact ce qui montre que u est nulle presque partout sur tout compact, elle est donc nulle presque partout. ■

Proposition III.27 (Propriétés immédiates des solutions)

On a les propriétés immédiates suivantes des solutions :

- *Pour tout $t > 0$ on a*
$$\min_{\mathbb{R}} u(t,\cdot) = \min_{\mathbb{R}} u_0, \quad \max_{\mathbb{R}} u(t,\cdot) = \max_{\mathbb{R}} u_0.$$
- *Si u et v sont deux solutions associées à deux données initiales u_0 et v_0, on a*
$$\|u-v\|_{L^\infty(\mathbb{R}^+\times\mathbb{R})} \leq \|u_0 - v_0\|_{L^\infty(\mathbb{R})},$$
mais aussi
$$\sup_t \|u(t,\cdot) - v(t,\cdot)\|_{L^p} \leq \|u_0 - v_0\|_{L^p},$$
dès que u_0 et v_0 sont dans L^p, $1 \leq p < +\infty$.
- *Si u_0 est à support dans un intervalle $[a,b]$ alors pour tout $u(t,\cdot)$ est à support dans $[a+ct, b+ct]$.*

Définition III.28 (Formulation faible)

Soit $u_0 \in L^\infty(\mathbb{R})$. On dit qu'une fonction $u \in L^\infty([0,+\infty[\times\mathbb{R})$ est solution **faible** *du problème* (III.5) *si et seulement si on a*

$$\int_0^\infty \int_{\mathbb{R}} u\big(\partial_t\varphi + \partial_x(c(x)\varphi)\big)\, dt\, dx + \int_{\mathbb{R}} u_0(x)\varphi(0,x)\, dx = 0,$$

pour toute fonction $\varphi \in \mathcal{C}_c^\infty([0,+\infty[\times\mathbb{R})$.

- Bien noter que l'opérateur qui agit sur la fonction-test n'est pas le même que celui qui définit l'équation. Il s'agit de l'opérateur adjoint.
- On a la même propriété que dans le cas précédent : toute solution régulière bornée du problème est solution faible et toute solution faible qui, de plus, est régulière est alors solution au sens usuel.

Théorème III.29

Pour toute donnée initiale $u_0 \in L^\infty(\mathbb{R})$, il existe une unique solution faible du problème (III.5) *qui est donnée par*

$$u(t,x) = u_0(X(0,t,x)).$$

IV.2 Existence et non-existence de solutions régulières

Dans cette partie nous allons supposer qu'il existe une solution **régulière** u de (III.20) définie sur $[0,T]\times\mathbb{R}$ pour un certain temps $T>0$ et nous allons voir ce que nous pouvons en dire.

1. Comme u est régulière et vérifie l'équation nous avons
$$\partial_t u + f'(u)\,\partial_x u = 0.$$
On peut donc maintenant poser :
$$c(t,x) = f'(u(t,x)),\ \ \forall t,x, \tag{III.22}$$
de sorte que la fonction u est donc solution de l'équation de transport à vitesse variable donnée par
$$\partial_t u + c(t,x)\partial_x u = 0.$$
Remarque importante : cette construction est pour l'instant théorique puisqu'à l'évidence la vitesse c dépend de la solution u qu'on ne connaît pas !

2. Les fonctions u et f étant suffisament régulières, la fonction c définie par (III.22) est elle-aussi régulière et on peut donc appliquer le théorème de Cauchy-Lipschitz et en déduire l'existence locale des courbes caractéristiques $t\mapsto X(t,0,x_0)$ associées à ce problème. On rappelle que celles-ci sont définies comme suit
$$\begin{cases} \dfrac{d}{dt}X(t,0,x_0) = c(t,X(t,0,x_0)),\\ X(0,0,x_0) = x_0. \end{cases}$$
D'après ce que l'on a vu concernant les équations de transport linéaires, en dérivant $u(t,X(t,0,x_0))$ on trouve que u est constante le long des courbes caractéristiques, et même de façon plus précise :
$$u(t,X(t,0,x_0)) = u_0(x_0),\ \ \forall 0\le t\le T. \tag{III.23}$$

3. Utilisons maintenant ce que l'on vient de démontrer, i.e. (III.23) ainsi que (III.22) dans la définition des courbes caractéristiques :
$$\frac{d}{dt}X(t,0,x_0) = c(t,X(t,0,x_0)) = f'(u(t,X(t,0,x_0))) = f'(u_0(x_0)) \quad \Leftarrow \text{ne dépend pas de } t.$$
Ceci montre que, *a posteriori*, les courbes caractéristiques sont des **droites** données par
$$X(t,0,x_0) = x_0 + tf'(u_0(x_0)).$$

4. On vient donc de démontrer que la solution est constante le long de droites dont les pentes dépendent de la donnée initiale u_0 et valent $f'(u_0(x_0))$.
Ce raisonnement étant fait, est-ce que cette propriété me définit de façon unique la solution u ? Fixons $t>0$, et étudions la fonction
$$x_0\mapsto \psi_t(x_0) = x_0 + tf'(u_0(x_0)).$$
Comme u_0 est bornée (et pour l'instant continue), on voit que ψ_t est continue et tend vers $\pm\infty$ en $\pm\infty$. Donc, par le théorème des valeurs intermédiaires, pour tout $x\in\mathbb{R}$, il existe au moins un x_0 tel que $\psi_t(x_0)=x$ on aurait donc $u(t,x)=u_0(x_0)$, pour cette valeur de x_0.
Le problème est : peut-il y avoir deux valeurs de x_0 vérifiant $\psi_t(x_0)=x$? Autrement dit, ψ_t est-elle bijective ? On calcule la dérivée de ψ_t :
$$\psi_t'(x_0) = 1 + tu_0'(x_0)f''(u_0(x_0)).$$
Donc pour des valeurs de t assez petites tout va fonctionner. En revanche pour des temps plus long, on peut avoir des problèmes ... selon la convexité de f, le signe de u_0', la valeur de t, ... l'étude du cas général est complexe.
Pour Burgers : $f(u)=u^2$, $f''(u)=2$ donc on trouve
$$\psi_t'(x_0) = 1 + 2tu_0'(x_0).$$
Conclusion pour Burgers :
   - Si u_0 est croissante, $\psi_t'>0$ tout le temps, et ψ_t est un $\mathcal{C}^1$-difféo. Il existe donc une unique solution régulière.
   - S'il existe des valeurs de x_0 en lesquelles u_0 est décroissante, alors ψ_t sera un difféomorphisme si et seulement si
$$t < \frac{-1}{2(\inf u_0')}.$$

5. Que se passe-t'il s'il y a effectivement deux antécédents de x par ψ_t ? Soient $x_1 < x_2$ tels que $x = \psi_t(x_1) = \psi_t(x_2)$. On a donc
$$x_1 - x_2 = -t\,(f'(u_0(x_1)) - f'(u_0(x_2))) \neq 0,$$
et en particulier $u_0(x_1) \neq u_0(x_2)$! Or d'après les raisonnements précédents, on devrait avoir
$$u(t,x) = u_0(x_1), \quad \text{et } \textbf{aussi } \; u(t,x) = u_0(x_2),$$
ce qui n'est bien sûr pas possible.

La conclusion de toute cette affaire, c'est que sous ces hypothèses, une solution régulière ne peut pas exister jusqu'au temps t en question.

Exemples pour Burgers :

– Cas $u_0(x) = x$, auquel cas $\psi_t(x_0) = (1+2t)x_0$ et donc ψ_t est un difféo
$$u(t,x) = u_0(\psi_t^{-1}(x)) = u_0(\frac{x}{1+2t}) = \frac{x}{1+2t}.$$

– Cas $u_0(x) = -x$, auquel cas $\psi_t(x_0) = (1-2t)x_0$ et donc ψ_t n'est un difféo que pour $t < 1/2$ et alors
$$u(t,x) = u_0(\frac{x}{1-2t}) = \frac{x}{1-2t}.$$
On observe la formation de la singularité en $t = 1/2$.

IV.3 Solutions faibles. Conditions de Rankine-Hugoniot

Il est donc naturel d'essayer de voir s'il existe des solutions faibles au problème.

Définition III.30

Une fonction $u \in L^1_{loc}(]0,T[\times\mathbb{R})$ t.q. $f(u) \in L^1_{loc}$ est solution faible du problème (III.20) *si on a*
$$\int_0^T \int_{\mathbb{R}} \Big(u(t,x)\partial_t\varphi(t,x) + f(u(t,x))\partial_x\varphi(t,x) \Big)\, dt\, dx + \int_{\mathbb{R}} u_0(x)\varphi(0,x)\, dx = 0,$$
pour toute fonction test $\varphi \in \mathcal{C}^\infty_c([0,T[\times\mathbb{R})$

Remarque III.31

– *Si u est une solution régulière alors c'est une solution faible.*
– *Si u est une solution faible et si de plus u est régulière, alors c'est une solution au sens classique.*
– *Voir plus loin les conditions de Rankine-Hugoniot.*

On a alors le théorème suivant dont la démonstration est admise.

Théorème III.32

Pour tout $T > 0$ et toute donnée initiale $u_0 \in L^\infty(\mathbb{R})$, il existe **au moins une** *solution faible $u \in L^\infty(]0,T[\times\mathbb{R})$ au problème* (III.20)*.*

Le problème majeur auquel on est alors confrontés c'est la **non-unicité** des solutions faibles.

Proposition III.33

La fonction définie par
$$u(t,x) = \begin{cases} 0 & \text{si } |x| > \sqrt{t}, \\ \frac{x}{2t} & \text{si } |x| \leq \sqrt{t}, \end{cases}$$
est une solution faible non nulle de l'équation de Burgers pour la donnée initiale $u_0 = 0$.
Remarque : *u n'est pas L^∞ mais on a $u \in L^1$ et $f(u) \in L^1$ donc on peut donner un sens à la formulation faible du problème.*

La condition de Rankine-Hugoniot Comment reconnaître qu'une fonction régulière par morceaux est solution faible (ou pas) du problème considéré ?

Théorème III.34 (Condition de Rankine-Hugoniot)

Soit $t \mapsto \alpha(t)$ une courbe régulière. Soit u une fonction de classe $\mathcal{C}^1$, bornée ainsi que ses dérivées, dans $\Omega_- = \{(t,x),\ \ x < \alpha(t)\}$ et de classe $\mathcal{C}^1$ dans $\Omega_+ = \{(t,x),\ \ x > \alpha(t)\}$. Alors u est solution faible du problème (III.20) *si et seulement si :*

- $u(0,\cdot) = u_0$.
- *u vérifie l'équation au sens classique dans Ω_- et Ω_+.*
- *u vérifie de plus la condition de saut suivante*

$$f(u(t,\alpha(t)^+)) - f(u(t,\alpha(t)^-)) = \alpha'(t)(u(t,\alpha(t)^+) - u(t,\alpha(t)^-)),\ \ \forall t \geq 0.$$

Cette condition de saut est souvent notée

$$\frac{[\![f(u)]\!]}{[\![u]\!]} = \alpha'.$$

Preuve :

Il s'agit d'effectuer des intégrations par parties. On suppose $\alpha(0) = 0$ ce qui ne change évidemment rien.

- Cas simple $\alpha(t) = 0$ pour tout t.
- Preuve dans un cas un peu plus général (mais pas commplet) : on suppose $\alpha'(t) > 0$ pour tout t et $\alpha(t) \to +\infty$ quand $t \to +\infty$. Dans ce cas, on peut introduire l'application réciproque de α et il s'agit alors à nouveau d'une intégration par parties, de changement de variables, etc ...

■

IV.4 Problème de Riemann

Afin de comprendre la propagation des discontinuités (chocs) dans une loi de conservation scalaire non-linéaire, on se propose d'étudier un problème simple dont la donnée initiale est une fonction créneau qui prend deux valeurs

$$u_0(x) = \begin{cases} u^+, \text{ si } x > 0, \\ u^-, \text{ si } x < 0. \end{cases}$$

On appelle ce problème particulier : **un problème de Riemann**.

Nous allons construire différentes solutions pour ce problème de Riemann pour l'équation de Burgers ($f(u) = u^2$) :

La solution onde de choc : Il s'agit d'une solution constante par morceaux. D'après la condition de Rankine-Hugoniot, la vitesse du choc vaut $\sigma = \frac{f(u^+)-f(u^-)}{u^+-u^-} = u^+ + u^-$.

La solution détente : Dans le cas où $u^- < u^+$: on construit les caractéristiques associées aux deux vitesses u^- et u^+ et on remplit le secteur angulaire restant par une solution régulière de la forme $w(\frac{x}{t})$ qui doit vérifier l'équation

$$-\frac{x}{t^2}w'(x/t) + \frac{1}{t}f'(w(x/t))w'(x/t) = 0,$$

d'où

$$w'(\xi)(f'(w(\xi)) - \xi) = 0,\ \ \forall \xi \in \mathbb{R}.$$

Peut-on mettre plusieurs chocs à la suite ? La réponse est oui comme on peut le voir sur un exemple.

IV.5 Comment résoudre complètement le problème de la non-unicité des solutions faibles ?

Il faut introduire la notion de solution entropique. On regarde l'équation parabolique

$$\partial_t u_\varepsilon + \partial_x(f(u_\varepsilon)) - \varepsilon\partial_x^2 u_\varepsilon = 0,$$

dont on peut montrer qu'elle admet une unique solution suffisament régulière pour tout $\varepsilon > 0$.

On peut également montrer que sous de bonnes hypothèses, la famille de fonctions $(u_\varepsilon)_\varepsilon$ converge quand ε tend vers 0 vers une fonction u qui est une solution faible particulière de la loi de conservation non-linéaire.

En effet, pour tout fonction convexe η, on démontre que

$$\partial_t \eta(u) + \partial_x q(u) \leq 0,$$

où $q' = f'\eta'$.

Une telle solution faible est dite **entropique** et les fonctions convexes η sont dites **entropies**.

Théorème III.35

Pour toute donnée initiale, il existe une unique solution faible entropique u.

C'est cette solution entropique qu'on va chercher à calculer dorénavant.

Première conséquence pour le problème de Riemann : la solution entropiqe u d'un problème de Riemann est auto-similaire, i.e. ne dépend que de x/t.

Proposition III.36 (Critère de Rankine-Hugoniot entropique)

Une solution faible régulière par morceaux est entropique si et seulement si, le long de toute courbe de discontinuité $t \mapsto \alpha(t)$ on a

$$[\![q(u)]\!] \leq \alpha' [\![\eta(u)]\!].$$

Exemples pour Burgers : On prend $\eta(u) = u^2$, d'où $q(u) = \frac{4}{3}u^3$.

D'après RH on a vu que $\alpha' = u^- + u^+$, et donc

$$[\![q(u)]\!] = \frac{4}{3}[(u^+)^3 - (u^-)^3],$$

$$[\![\eta(u)]\!] = (u^+)^2 - (u^-)^2.$$

La condition de RH entropique devient

$$\frac{4}{3}((u^+)^3 - (u^-)^3) \leq (u^+ + u^-)((u^+)^2 - (u^-)^2),$$

ou encore

$$(u^+)^3 - 3u^+(u^-)^2 + 3u^+(u^-)^2 - (u^-)3 \leq 0,$$

$$(u^+ - u^-)^3 \leq 0,$$

ce qui équivaut à

$$u^+ \leq u^-.$$

Un choc n'est donc entropique que si u diminue à la discontinuité, sinon la solution entropique est donc la solution détente.

IV.6 Schémas

Schéma centré Il s'agit du schéma suivant

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} + \frac{f(u_{i+1}^n) - f(u_{i-1}^n)}{2\Delta x} = 0,$$

qui est d'ordre 1 en temps et 2 en espace, mais qui n'est jamais stable comme d'habitude ...

Schéma de Lax-Friedrichs Il s'agit du schéma suivant

$$\frac{u_i^{n+1} - \frac{u_{i+1}^n + u_{i-1}^n}{2}}{\Delta t} + \frac{f(u_{i+1}^n) - f(u_{i-1}^n)}{2\Delta x} = 0,$$

qui est d'ordre 1 si Δt et Δx sont proportionnels, et monotons et L^2-stable sous la condition CFL

$$\sup_{[\min u_0, \max u_0]} |f'| \frac{\Delta t}{\Delta x} \leq 1.$$

Un schéma décentré amont

$$\frac{u_i^{n+1}-u_i^n}{\Delta t}+\frac{f(u_i^n)-f(u_{i-1}^n)}{\Delta x}=0, \text{ si } f \text{ est croissante,}$$

$$\frac{u_i^{n+1}-u_i^n}{\Delta t}+\frac{f(u_{i+1}^n)-f(u_i^n)}{\Delta x}=0, \text{ si } f \text{ est décroissante.}$$

Ce schéma est consistant d'ordre 1 en temps et en espace mais n'est pas toujours entropique, stable sous CFL.

Si f n'est ni croissante ni décroissante, on peut séparer f en deux parties $f = f_1 + f_2$ avec f_1 croissante et f_2 décroissante, ce qui fournira un nouveau schéma décentré.

Schéma non conservatif Exemple du schéma non conservatif pour Burgers (avec $u_i^n > 0$)

$$\frac{u_i^{n+1}-u_i^n}{\Delta t}+u_i^n\frac{u_i^n-u_{i-1}^n}{\Delta x}=0.$$

Ce schéma ne capture pas la bonne vitesse des chocs car pour des solutions non régulières, les formes conservatives et non conservatives ne sont plus équivalentes.

V Rappels sur la transformée de Fourier

N.B. : Ces très brefs rappels n'ont pour but que de donner les outils utilisés dans ce chapitre.

Pour toute fonction $f \in L^1(\mathbb{R})$, on définit la transformée de Fourier de f comme la fonction (à valeurs complexes)

$$\mathcal{F}f(\xi)=\int_{\mathbb{R}} f(x)e^{-\sqrt{-1}x\xi}\,dx. \tag{III.24}$$

Dans cette formule $\sqrt{-1}$ désigne le nombre complexe imaginaire pur habituellement noté i. Dans le cadre de ce cours, on préfère éviter de le noter i pour ne pas confondre avec les indices.

Pour tout $h \in \mathbb{R}$, on définit l'opérateur de translation $\tau_h : L^1(\mathbb{R}) \mapsto L^1(\mathbb{R})$ (la même définition est valable pour $L^2(\mathbb{R})$ au lieu de $L^1(\mathbb{R})$) par la formule

$$\tau_h f(x) = f(x-h), \ \ \forall x \in \mathbb{R}. \tag{III.25}$$

Proposition III.37

Pour toute fonction $f \in L^1(\mathbb{R})$ et tout $h \in \mathbb{R}$, on a

$$\mathcal{F}(\tau_h f)(\xi) = e^{-\sqrt{-1}h\xi}\mathcal{F}f(\xi), \ \ \forall \xi \in \mathbb{R}. \tag{III.26}$$

Preuve :

Il s'agit d'un simple changement de variable affine $y = x - h$ dans une intégrale

$$\begin{aligned}\mathcal{F}(\tau_h f)(\xi)=\int_{\mathbb{R}}(\tau_h f)(x)e^{-\sqrt{-1}x\xi}\,dx=\int_{\mathbb{R}} f(x-h)e^{-\sqrt{-1}x\xi}\,dx&=\int_{\mathbb{R}} f(y)e^{-\sqrt{-1}(y+h)\xi}\,dy\\ &=e^{-\sqrt{-1}h\xi}\int_{\mathbb{R}} f(y)e^{-\sqrt{-1}y\xi}\,dy=e^{-\sqrt{-1}h\xi}\mathcal{F}f(\xi).\end{aligned}$$

■

Le résultat essentiel que l'on utilisera sur la transformée de Fourier est le résultat suivant (admis).

Théorème III.38 (de Plancherel)

Il existe un unique opérateur $\mathcal{F} : L^2(\mathbb{R}) \to L^2(\mathbb{R})$ qui coïncide avec la transformation de Fourier (formule (III.24)*) sur l'espace $L^1(\mathbb{R}) \cap L^2(\mathbb{R})$. Cet opérateur est encore appelé* **transformation de Fourier**.
De plus, l'opérateur $\frac{1}{\sqrt{2\pi}}\mathcal{F}$ ainsi construit est une bijection isométrique. Plus précisément on a

$$\forall f \in L^2(\mathbb{R}), \quad \|\mathcal{F}f\|_{L^2(\mathbb{R})} = \sqrt{2\pi}\|f\|_{L^2(\mathbb{R})}.$$

Chapitre IV

EDP elliptiques. Equations de Poisson et de Laplace

I Modèles

I.1 Equations de Maxwell

On présente tout d'abord les équations de Maxwell (dites *dans le vide*), en toute généralité. On note ρ la densité volumique de charge électrique et j la densité volumique de courant.

– Equation de Maxwell-Gauss :

$$\text{div } E = \frac{\rho}{\varepsilon_0},$$

elle dit que le flux du champ électrique à travers une surface fermée est égale à la charge totale intérieure divisée par ε_0 (permitivité du vide).

Elle se démontre à partir de la loi de Coulomb qui dit que la force d'attraction ou de répulsion entre deux charges ponctuelles est proportionnelle au produit des charges, inversement proportionnelle au carré de la distance et orientée de sorte que les charges de même signe se repoussent.

– Equation de Maxwell-Flux magnétique :

$$\text{div } B = 0,$$

le flux du champ magnétique à travers toute surface fermée est nul (autrement dit, il n'existe pas de charge électrique).

– L'équation de Maxwell-Faraday :

$$\text{rot } E = -\frac{\partial B}{\partial t},$$

qui dit que la circulation de E le long d'une courbe fermée est l'opposée de la dérivée du flux magnétique à travers la courbe.

– L'équation de Maxwell-Ampère :

$$c^2 \text{rot } B = \frac{j}{\varepsilon_0} + \frac{\partial E}{\partial t},$$

qui dit que c^2 fois la circulation du champ magnétique le long d'une courbe fermée est égale à la somme du courant à travers la courbe divisée par ε_0 plus la dérivée en temps du flux de E à travers la courbe.

Remarques : les équations sont **linéaires** par rapport à ρ, j, E et B.

Proposition IV.1 (Cohérence des équations)

Les équations de Maxwell sont compatibles avec la loi de conservation de la charge :

$$\partial_t \rho + \textit{div } j = 0.$$

Proposition IV.2 (Electrostatique)

Dans une situation stationnaire (i.e. toutes les grandeurs sont indépendantes du temps), on trouve les équations

$$\textit{div}\ E = \frac{\rho}{\varepsilon_0}, \textit{rot}\ E = 0,$$

et donc il existe une fonction V (appellée potentiel électrique) telle que $E = -\nabla V$ et qui est solution de l'équation de **Poisson**

$$-\Delta V = \frac{\rho}{\varepsilon_0}.$$

Si $\rho = 0$, on trouve

$$-\Delta V = 0,$$

qui est appelée équation de **Laplace**.

Proposition IV.3 (Magnétostatique)

Dans une situation stationnaire (i.e. toutes les grandeurs sont indépendantes du temps), on trouve les équations

$$\textit{div}\ B = 0, \textit{rot}\ B = \frac{j}{c^2\varepsilon_0}.$$

On montre alors qu'il existe une fonction **vectorielle** *(appellé potentiel vecteur) notée A telle que $B = \textit{rot}\ A$.*
On peut deplus imposer que $\textit{div}\ A = 0$ (quitte à ajouter ou soustraire un gradient à A).
On en déduit que A est solution de l'équation vectorielle

$$-\Delta A = \frac{j}{c^2\varepsilon_0}.$$

I.2 Equation de la chaleur

On verra plus loin que la température (notée u) dans un matériau Ω satisfait à l'équation (dite équation de la chaleur), issue de la loi de Fourier, qui est la suivante

$$\partial_t u - \underbrace{k\Delta u}_{=\text{div}\ (k(x)\nabla u)} = f(t,x), \ \text{ avec } u(0,x) = u_0(x),$$

où f représente les sources de chaleur ponctuelles, k est une constante qui dépend des caractéristiques du matériau. A cette équation il faut rajouter les conditions aux limites qui décrivent la situation sur le bord du matériau :

- Ou bien la température est imposée (on place une extrémité du matériau dans de l'eau glacée) : $u = \text{quelquechose}$ on parle de conditions aux limites de Dirichlet.
- Ou bien le matériau est isolé et n'échange pas de chaleur avec l'extérieur : $\frac{\partial u}{\partial n} = 0$.
- Ou bien n'importe quelle combinaison des deux ...

A l'état stationnaire, la répartition de la température dans le milieu satisfait donc à l'équation :

$$-\Delta u = f_\infty(x), \quad + \text{CL},$$

où f_∞ est la limite quand $t \to \infty$ du terme source. On retrouve l'équation de Poisson.

I.3 Membrane élastique à l'équilibre

On considère une membrane élastique qui au repos est représentée par une région compacte $\overline{\Omega}$ du plan horizontal $\mathbb{R}^2$, où Ω est un ouvert **connexe** de $\mathbb{R}^2$. On applique une force verticale (petite) notée $f(x)$ en chaque point de la membrane ce qui a pour effet de la déformer.

Tout point $x \in \overline{\Omega}$ est déplacé en un point de $\mathbb{R}^3$ noté $\tilde{u}(x) = (x, u(x))$, où u est une fonction de $\overline{\Omega}$ dans $\mathbb{R}$ à déterminer représentant le déplacement vertical. On se place dans l'hypothèse de petits déplacements. De plus, on va considérée la situation dans laquelle la membrane est attachée par son bord à un référentiel fixe (penser à la peau d'un tambour). Ceci implique que l'on cherchera une fonction u qui est nulle sur le bord $\partial\Omega$.

Faisons le bilan d'énergie potentielle du système :

– Le système acquiert une énergie potentielle virtuelle qui vaut l'opposé du travail des forces extérieures f par rapport au déplacement u, c'est-à-dire :

$$E_1(u) = -\int_\Omega f(x)u(x)\,dx.$$

La présence du signe $-$ est naturelle : si la force est orientée vers le bas ($f \leq 0$) alors les zones de forte énergie potentielle sont les zones les plus hautes (donc pour les $u \geq 0$ grand) (comme pour le champ de gravité par exemple).

– Par ailleurs le système contient de l'énergie potentielle élastique du à la déformation de la membrane. On admet que cette énergie est proportionnelle au changement d'aire de la membrane [1]

$$E_2(u) = k(\text{Aire déformée} - \text{Aire au repos}) = k(|\tilde{u}(\Omega)| - |\Omega|).$$

Par changement de variable, on trouve

$$E_2(u) = k\int_\Omega \left(\sqrt{1+|\nabla u|^2} - 1\right) dx.$$

Dans l'hypothèse des petits déplacements, u est petite ainsi que ses dérivées. Par un développement limité usuel, on approche alors E_2 par l'expression :

$$E_2(u) = \frac{k}{2}\int_\Omega |\nabla u|^2\,dx.$$

L'énergie potentielle totale du système en fonction du déplacement u est donnée par

$$E(u) = E_1(u) + E_2(u) = \frac{k}{2}\int_\Omega |\nabla u|^2\,dx - \int_\Omega fu\,dx.$$

Le principe fondamental de la mécanique Lagrangienne nous dit alors que, sous l'effet du champ de forces f, la membrane va se déformer selon un déplacement u qui va minimiser l'énergie potentielle totale (qu'on devrait plutôt appeler *l'action* selon le vocabulaire *ad hoc* de la mécanique). On s'intéresse donc au problème suivant : trouver $u \in X$ tel que

$$E(u) = \inf_{v\in X} E(v) = \inf_{v\in X}\left(\frac{k}{2}\int_\Omega |\nabla u|^2\,dx - \int_\Omega fu\,dx\right), \tag{IV.1}$$

où, *a priori* X est l'espace fonctionnel suivant :

$$X = \{v : \overline{\Omega} \mapsto \mathbb{R},\ \text{dérivable},\ \ v = 0,\ \ \text{sur } \partial\Omega\}.$$

Nous verrons par la suite que cet espace X n'est pas nécessairement le bon choix.

Les questions mathématiques que l'on veut résoudre : Dans la suite, on va essayer de répondre aux questions suivantes :

1. Le problème (IV.1) admet-il une solution ? En particulier, est-ce que l'infimum de E sur X est fini ?
2. Si une telle solution existe, est-elle unique ?
3. Si une telle solution existe, est-ce qu'on peut la caractériser au moyen d'une équation "simple" que l'on pourra éventuellement résoudre ?

Simplification : A partir de maintenant, et dans toute la suite du chapitre, on va se placer dans le cas plus simple de la dimension 1. Le modèle correspond alors à une corde élastique (et non plus une membrane). Pour simplifier encore, on considère $\Omega =]0,1[$ de sorte que le problème s'écrit de la façon suivante : trouver $u : [0,1] \mapsto \mathbb{R}$ dérivable, nulle en $x = 0$ et $x = 1$ et telle que

$$E(u) = \inf_{v\in X} E(v), \tag{IV.2}$$

où $X = \{v : [0,1] \mapsto \mathbb{R}, \text{dérivable et tq } v(0) = v(1) = 0\}$ et l'énergie s'écrit maintenant

$$E(v) = \frac{k}{2}\int_0^1 |v'(x)|^2\,dx - \int_0^1 f(x)v(x)\,dx.$$

On supposera également que f est, au minimum, une fonction intégrable.

1. Ceci peut se "démontrer", par exemple, en assimilant la membrane à un réseau de petits ressorts

Unicité : On va commencer par traiter le problème de l'unicité, qui est finalement le problème le plus simple. Cette propriété découle naturellement de la **stricte convexité** de la fonctionnelle E (et de la convexité de l'ensemble X).

Supposons données deux fonctions $u_1, u_2 \in X$ solutions du problème (IV.2). On suppose donc, en particulier, que l'infimum de E est fini. On notera sa valeur

$$I_E = \inf_{v\in X} E(v),$$

et on a donc, par hypothèse $E(u_1) = E(u_2) = I_E$.

On pose alors $u = \frac{u_1+u_2}{2}$ et on calcule l'énergie de u, en utilisant l'identité du parallèlogramme [2]

$$\begin{aligned} E(u) &= \frac{k}{2}\int_0^1 \left|\frac{u_1'(x)+u_2'(x)}{2}\right|^2 dx - \int_0^1 f(x)\frac{u_1(x)+u_2(x)}{2}\,dx \\ &= \frac{k}{4}\int_0^1 |u_1'(x)|^2\,dx + \frac{k}{4}\int_0^1 |u_2'(x)|^2\,dx - \frac{k}{2}\int_0^1 \left|\frac{u_1'(x)-u_2'(x)}{2}\right|^2 dx \\ &\qquad - \frac{1}{2}\int_0^1 f(x)u_1(x)\,dx - \frac{1}{2}\int_0^1 f(x)u_2(x)\,dx \\ &= \frac{1}{2}E(u_1) + \frac{1}{2}E(u_2) - \frac{k}{2}\int_0^1 \left|\frac{u_1'(x)-u_2'(x)}{2}\right|^2 dx. \end{aligned}$$

Or, par hypothèse, u_1 et u_2 sont solutions du problème (IV.2) et donc, on a finalement

$$E(u) = I_E - \frac{k}{2}\int_0^1 \left|\frac{u_1'(x)-u_2'(x)}{2}\right|^2 dx.$$

Comme par ailleurs, par définition de l'infimum (et comme $u = \frac{u_1+u_2}{2} \in X$), on a $E(u) \geq I_E$, on déduit que l'on a nécessairement

$$\frac{k}{2}\int_0^1 \left|\frac{u_1'(x)-u_2'(x)}{2}\right|^2 dx = 0.$$

La fonction sous l'intégrale étant positive, on obtient immédiatement que

$$\forall x \in [0,1], u_1'(x) = u_2'(x).$$

Ceci implique que $u_1 - u_2$ est une fonction constante, mais comme $u_1(0) = u_2(0) = 0$ (par définition des conditions au bord dans X), on a finalement montré

$$\forall x \in [0,1],\ \ u_1(x) = u_2(x),$$

ce qui montre bien l'unicité d'une *éventuelle* solution de (IV.2).

Caractérisation de la solution On continue à supposer dans ce paragraphe que la solution $u \in X$ du problème (IV.2) existe. On va montrer que, sous de bonnes hypothèses, on peut la caractériser par une équation aux dérivées partielles (ici en 1D donc avec une seule variable).

La méthode ci-dessous est standard en optimisation et calcul des variations. Il s'agit d'établir les équations **d'Euler-Lagrange** associées au problème de minimisation (IV.2).

La preuve de ce résultat n'est finalement rien d'autre que la traduction en dimension infinie du résultat élémentaire suivant :

Lemme IV.4

Soit $\varphi : \mathbb{R} \mapsto \mathbb{R}$ une fonction dérivable. On suppose qu'il existe $t^* \in \mathbb{R}$ tel que

$$\varphi(t^*) = \inf_{t\in\mathbb{R}} \varphi(t), \tag{IV.3}$$

alors on a

$$\varphi'(t^*) = 0.$$

Il ne semble pas inutile de rappeler la démonstration de ce lemme pour comprendre comment intervient l'hypothèse.

Preuve :

D'après (IV.3), pour tout $h > 0$ (le signe de h joue ici un rôle crucial !), on a

$$\varphi(t^* + h) \geq \varphi(t^*).$$

2. Rappel : dans un Hilbert H, pour tous $a, b \in H$ on a $\|\frac{a+b}{2}\|^2 + \|\frac{a-b}{2}\|^2 = \frac{\|a\|^2}{2} + \frac{\|b\|^2}{2}$.

Comme $h > 0$, on en déduit

$$\frac{\varphi(t^* + h) - \varphi(t^*)}{h} \geq 0.$$

On passe maintenant à la limite quand $h \to 0^+$ dans cette inégalité, ce qui donne par définition du nombre dérivé

$$\varphi'(t^*) \geq 0.$$

Si maintenant on reprend ce calcul avec $h < 0$, on a

$$\varphi(t^* + h) \geq \varphi(t^*),$$

et ainsi (comme $h < 0$!) il vient

$$\frac{\varphi(t^* + h) - \varphi(t^*)}{h} \leq 0.$$

En passant à la limite quand $h \to 0^-$, on trouve

$$\varphi'(t^*) \leq 0,$$

ce qui donne le résultat. ■

Revenons à notre problème de corde élastique. Supposons donc qu'une solution u de (IV.2) existe. On se donne un v quelconque dans X, de sorte que, pour tout $t \in \mathbb{R}$, on a $u + tv \in X$ (car X est un espace vectoriel !). Ainsi, par définition de l'infimum, on a

$$\forall t \in \mathbb{R}, E(u + tv) \geq E(u),$$

ce qui montre que la fonction $\varphi_v : \mathbb{R} \mapsto \mathbb{R}$ définie par $\varphi_v(t) = E(u + tv)$ admet un minimum en $t^* = 0$. Par ailleurs, cette fonction est dérivable (on va même voir ci-dessous que c'est un polynôme de degré 2 dans la variable t). D'après le lemme précédent, on en déduit que nécessairement $\varphi_v'(0) = 0$.

Il reste à calculer $\varphi_v'(0)$. Pour cela, on écrit φ_v sous la forme suivante

$$\varphi_v(t) = E(u) + t\left[k\int_0^1 u'(x)v'(x)\,dx - \int_0^1 f(x)v(x)\,dx\right] + \frac{t^2}{2}k\int_0^1 |v'(x)|^2\,dx,$$

et donc, la relation $\varphi_v'(0) = 0$ devient

$$k\int_0^1 u'(x)v'(x)\,dx = \int_0^1 f(x)v(x)\,dx.$$

Comme ceci est vrai pour toute fonction $v \in X$, on a donc démontré que la solution, si elle existe, du problème (IV.2) vérifie les équations d'Euler-Lagrange suivantes :

$$\forall v \in X, k\int_0^1 u'(x)v'(x)\,dx = \int_0^1 f(x)v(x)\,dx. \tag{IV.4}$$

Il est facile de vérifier la propriété réciproque (à titre d'exercice) :

Proposition IV.5

Si $u \in X$ vérifie (IV.4)*, alors u est solution du problème* (IV.2)*.*

Jusqu'à présent nous n'avons eu besoin d'aucune hypothèse particulière sur la solution u de notre problème. Si on admet que celle-ci est un peu plus régulière que simplement dérivable, alors on peut aller plus loin dans l'analyse.

Théorème IV.6

On suppose que le problème (IV.2) *admet une solution $u \in X$.*
Si on suppose, de plus, que cette solution vérifie $u \in \mathcal{C}^2([0,1])$ et que $f \in \mathcal{C}^0([0,1])$, alors u vérifie le problème de Poisson suivant :

$$\begin{cases} -k\partial_x^2 u = f, \ \ \textit{dans }]0,1[, \\ u(0) = u(1) = 0. \end{cases} \tag{IV.5}$$

La réciproque est également vraie (voir l'exercice ci-dessous).

On retrouve donc bien le problème de Poisson (en dimension 1 bien sûr). Démontrons ce théorème.

Preuve :

Il s'agit, dans un premier temps, d'une simple intégration par parties. Pour tout $v \in X$, comme u est de classe $\mathcal{C}^2$, on peut en effet intégrer par parties l'équation (IV.4) et obtenir

$$\int_0^1 (k\partial_x^2 u(x) + f(x))v(x)\,dx - [k(\partial_x u)v]_0^1 = 0.$$

On ne sait rien de la valeur de $\partial_x u$ en $x = 0$ et $x = 1$, par contre on a $v(0) = v(1) = 0$, ce qui montre que le dernier terme de cette formule est nul.

Si on pose $G(x) = k\partial_x^2 u(x) + f(x)$, on a donc obtenu

$$\forall v \in X,\ \int_0^1 G(x)v(x)\,dx = 0. \tag{IV.6}$$

On voudrait déduire de cela que la fonction G est identiquement nulle, ce qui montrera bien (IV.5). Il faut tout d'abord remarquer que l'on ne peut pas simplement prendre $v = G$ dans la formule car la fonction G n'est pas dans l'espace X (on ne sait pas, a priori, que $G(0) = G(1) = 0$). Il faut donc raisonner autrement. On propose ici deux façons de faire (la première étant bien plus générale car elle n'utilise pas réellement le fait que G est continue).

Preuve 1 : On utilise la propriété suivante (voir le cours d'intégration, ou d'analyse fonctionelle) :

L'ensemble des fonctions $\mathcal{C}^\infty$ à support compact dans $]0,1[$ est dense dans $L^1(]0,1[)$.

Cela signifie que toute fonction de L^1 peut-être approchée par une suite de fonctions de classe $\mathcal{C}^\infty$ à support compact (pour la topologie de L^1 bien sûr). Comme $\mathcal{C}_c^\infty(]0,1[) \subset X$, on en déduit que X est dense dans L^1. Comme G est continue (donc dans L^1), il existe une suite $(v_n)_n$ d'éléments de X telle que $\|v_n - G\|_{L^1} \to 0$.

On a alors

$$\int_0^1 |G(x)|^2\,dx = \int_0^1 G(x)(G(x) - v_n(x))\,dx + \int_0^1 G(x)v_n(x)\,dx.$$

Comme $v_n \in X$, le second terme est nul d'après (IV.6), et on peut majorer le premier terme comme suit

$$\int_0^1 |G(x)|^2\,dx \leq \|G\|_{L^\infty}\|G - v_n\|_{L^1} \xrightarrow[n\to\infty]{} 0.$$

Ainsi, $|G|^2$ est une fonction continue positive et d'intégrale nulle, elle est donc bien nulle.

Preuve 2 : On raisonne par l'absurde en supposant que G est non identiquement nulle. Comme G est continue, et quitte à changer G en $-G$, il existe une constante $C > 0$ et un intervalle d'intérieur non vide $[\alpha, \beta] \subset [0,1]$ tels que

$$\forall x \in [\alpha,\beta], G(x) \geq C.$$

On construit alors une fonction v de classe $\mathcal{C}^1$, positive, identiquement nulle en dehors de $[\alpha,\beta]$ et telle que $\int_\alpha^\beta v(x)\,dx > 0$. La construction d'une telle fonction est immédiate mais on n'a pas besoin de connaître la formule exacte pour faire la démonstration.

D'après (IV.6) et les propriétés de G et v, on a

$$0 = \int_0^1 G(x)v(x)\,dx = \int_\alpha^\beta G(x)v(x)\,dx \geq C\int_\alpha^\beta v(x)\,dx > 0,$$

ce qui constitue une contradiction manifeste.

■

Exercice IV.1

- *Démontrer directement (i.e. sans utiliser tout ce qui précède), que le problème* (IV.5) *admet, au plus, une solution de classe* $\mathcal{C}^2$.
- *Démontrer que toute solution du problème* (IV.5) *est solution des équations d'Euler-Lagrange* (IV.4). *Ceci prouve que les trois problèmes considérés précédemment (problème de minimisation, équations d'Euler-Lagrange et équation de Poisson) sont équivalents, au moins pour des fonctions suffisament régulières.*

Remarque IV.7

Tous les calculs précédents peuvent être effectués en dimension quelconque. L'équation aux dérivées partielles que l'on obtient alors à la place de (IV.5) *est*

$$\begin{cases} -k\Delta u = f, \text{ dans } \Omega, \\ u = 0, \text{ sur } \partial\Omega. \end{cases}$$

Remarque IV.8

Si on reprend toute l'analyse précédente en supposante que la tension k de la corde/membrane dépend du point x où l'on se place, alors on aboutit aux équations suivantes

$$-div\,(k(x)\nabla u) = f, \ \ dans\ \Omega,$$

en dimension 2 *et*

$$-\partial_x(k(x)\partial_x u) = f(x), \ \ dans\ ]0,1[,$$

en dimension 1*, toujours assortie des conditions aux limites.*

Pour l'instant nous avons démontré que si le problème de minimisation de E sur X admet une solution, alors elle est unique et que de plus si celle-ci est régulière, alors elle est solution de l'équation de Poisson. La question de démontrer l'existence du minimiseur n'est pas encore résolue, elle le sera un peu plus loin, dans la section II. En attendant, il est instructif de savoir résoudre explicitement le problème de Poisson en dimension 1 d'espace.

Résolution explicite du problème de Dirichlet homogène en 1D (voir aussi le TD) Soit à résoudre explicitement le problème

$$-\partial_x(k(x)\partial_x u) = f(x), \ \ u(0) = 0, \ u(1) = 0,$$

où f et k sont continues sur $[0,1]$ avec $\inf_{[0,1]} k > 0$.

– On commence par primitiver une première fois : on obtient

$$-\partial_x u(x) = \frac{C}{k(x)} + \frac{1}{k(x)} \int_0^x f(t)\,dt,$$

puis on primitive une seconde fois (en notant $F(x) = \int_0^x f(t)\,dt$) en utilisant le fait que $u(0) = 0$:

$$u(x) = -C \int_0^x \frac{1}{k(y)}\,dy - \int_0^x \frac{F(y)}{k(y)}\,dy.$$

– On détermine maintenant C pour que $u(1)$ soit égal à 0. On voit immédiatement qu'il n'y a qu'une seule solution possible

$$0 = -C \int_0^1 \frac{1}{k(y)}\,dy - \int_0^1 \frac{F(y)}{k(y)}\,dy,$$

soit

$$C = -\frac{1}{\int_0^1 \frac{1}{k}} \int_0^1 \frac{F(y)}{k(y)}\,dy.$$

– On note dorénavant $h = 1/k$ puis on reporte dans la formule qui donne u et on trouve

$$u(x) = \frac{1}{\int_0^1 h} \left[\left(\int_0^1 F(t)h(t)\,dt \right) \left(\int_0^x h(t)\,dt \right) - \left(\int_0^x F(t)h(t)\,dt \right) \left(\int_0^1 h(t)\,dt \right) \right].$$

Essayons d'arranger cette expression en utilisant le théorème de Fubini

$$\int_0^x F(t)h(t)\,dt = \int_0^x \left(\int_0^t f(y)\,dy \right) h(t)\,dt = \int_0^x \left(\int_y^x h(t)\,dt \right) f(y)\,dy.$$

Il vient donc

$$u(x) = \frac{1}{\int_0^1 h}\Big[\left(\int_0^1 \left(\int_y^1 h(t)\,dt\right) f(y)\,dy\right)\left(\int_0^x h(t)\,dt\right) \\ - \left(\int_0^x \underbrace{\left(\int_y^x h(t)\,dt\right)}_{=T} f(y)\,dy\right)\left(\int_0^1 h(t)\,dt\right)\Big].$$

Ecrivons le terme T sous la forme

$$T = \int_y^1 h(t)\,dt - \int_x^1 h(t)\,dt,$$

il vient

$$u(x) = \frac{1}{\int_0^1 h}\Big[\left(\int_0^1 \left(\int_y^1 h(t)\,dt\right) f(y)\,dy\right)\left(\int_0^x h(t)\,dt\right) \\ - \left(\int_0^x \left(\int_y^1 h(t)\,dt\right) f(y)\,dy\right)\underbrace{\left(\int_0^1 h(t)\,dt\right)}_{=T'} \\ + \left(\int_0^x f(y)\,dy\right)\left(\int_x^1 h(t)\,dt\right)\left(\int_0^1 h(t)\,dt\right)\Big].$$

Ecrivons enfin le terme T' sous la forme

$$T' = \int_0^x h(t)\,dt + \int_x^1 h(t)\,dt,$$

de sorte que la première partie du terme correspondant se combien avec le premier terme et la seconde partie avec le troisième terme. Il reste

$$u(x) = \frac{1}{\int_0^1 h}\Big[\left(\int_x^1 \left(\int_y^1 h(t)\,dt\right) f(y)\,dy\right)\left(\int_0^x h(t)\,dt\right) \\ + \left(\int_0^x \left(\int_0^y h(t)\,dt\right) f(y)\,dy\right)\left(\int_x^1 h(t)\,dt\right)\Big].$$

On a donc obtenu une formule qui donne u sous la forme

$$u(x) = \int_0^1 G(x,y) f(y)\,dy,$$

où la fonction G est donnée par

$$G(x,y) = \begin{cases} \dfrac{1}{\int_0^1 h}\left(\displaystyle\int_0^y h(t)\,dt\right)\left(\displaystyle\int_x^1 h(t)\,dt\right), & \text{si } 0 \le y \le x, \\ \dfrac{1}{\int_0^1 h}\left(\displaystyle\int_y^1 h(t)\,dt\right)\left(\displaystyle\int_0^x h(t)\,dt\right), & \text{si } x \le y \le 1. \end{cases}$$

Celle-ci a comme propriétés :

- G est symétrique.
- G est positive.
- G est continue sur $[0,1]^2$, dérivable en dehors de la diagonale $\{x = y\}$.

La fonction G s'appelle la **fonction de Green** associé au problème de Poisson avec condition aux limites de Dirichlet homogènes.

On a donc montré le théorème suivant :

Théorème IV.9

Pour toute fonction continue f, le problème de Poisson avec conditions aux limites de Dirichlet homogènes a une unique solution u de classe $\mathcal{C}^2$ donné par

$$u(x) = \int_0^1 G(x,y) f(y)\, dy,$$

où G est définie ci-dessus.
En particulier, on a la propriété fondamentale suivante appelée **principe du maximum** *:*

$$f \geq 0 \Longrightarrow u \geq 0,$$

et si de plus u s'annulle en un point à l'intérieur de $]0,1[$, alors u est identiquement nulle.

Les propriétés ci-dessus sont très générales, et on peut prouver l'existence de la fonction de Green pour tout opérateur elliptique raisonnable. De même, le principe du maximum est valable pour une très large classe de problèmes elliptiques, en dimension quelconque, mais cela dépasse le cadre de ce cours. Cette propriété a d'innombrables conséquences qui permettent d'établir de nombreuses propriétés très fines des solutions de ces équations.

II Eléments d'analyse avancée

Le but de ce deuxième paragraphe est de montrer l'existence d'une solution au problème de minimisation (IV.2). C'est en effet, le seul point que l'on a pas encore abordé dans l'analyse du problème. On rappelle qu'on a déjà établi l'unicité d'une éventuelle solution, qu'on a caractérisé cette éventuelle solution par le biais des équations d'Euler-Lagrange et qu'enfin, par intégration par parties, on a montré que cette fonction devait vérifier une équation aux dérivées partielles (équation de Poisson), dès lors qu'elle est suffisamment régulière suffisamment régulière.

II.1 Comment montrer l'existence d'un minimiseur ?

Le principe général de la preuve de l'existence d'un minimiseur pour une fonctionnelle E sur un espace X est le suivant :

1. On démontre que $\inf_X E > -\infty$. Si ceci est faux, le problème de chercher un minimiseur n'a évidemment pas de sens, c'est donc une étape indispensable.
2. On considère une *suite minimisante*, c'est-à-dire une suite $(u_n)_n \subset X$ qui vérifie $\lim_{n\to\infty} E(u_n) = \inf_X E$. Une telle suite existe **toujours**, c'est juste la définition de l'infimum qui nous la fournit.
3. On essaie de démontrer que cette suite (ou l'une de ses sous-suites) converge (en un sens à préciser : idéalement pour la norme de X). On note u la limite obtenue.
4. On essaie de vérifier que u réalise bien le minimum recherché.

Pour faire fonctionner ce programme de travail, le choix du bon espace X et de la bonne topologie sur cet espace sont cruciales. En effet, l'espace X étant, en général, de dimension infinie, le choix d'une topologie sur X (même une topologie d'e.v.n.) n'est pas du tout trivial.

II.2 Pourquoi l'espace X défini précédemment ne convient pas ?

De façon générale, pour réaliser la troisième étape du programme ci-dessus, on va essayer de montrer que la suite minimisante est de Cauchy, donc convergente, **si l'espace X est complet**. C'est pourquoi la complétude de l'espace est une notion centrale.

Une première idée serait donc de changer un tout petit peu l'espace X considéré en choisissant

$$X = \{v : [0,1] \to \mathbb{R}, \text{ de classe } \mathcal{C}^1 \text{ et tq } v(0) = v(1) = 0\},$$

muni de la norme $\|v\|_X = \|v\|_{L^\infty} + \|\partial_x v\|_{L^\infty}$. On sait en effet que c'est espace est un Banach.

Finitude de l'infimum Commençons par vérifier que l'infimum est fini. Pour cela, on utilise le résultat suivant

Lemme IV.10

Si $v \in X$, on a

$$\|v\|_{L^\infty} \leq \|\partial_x v\|_{L^2}.$$

Preuve :

On écrit, pour tout $x \in [0,1]$,

$$v(x) = \underbrace{v(0)}_{=0, \text{car } v \in X} + \int_0^x v'(t)\,dt,$$

puis on majore l'intégrale par l'inégalité de Cauchy-Schwarz. ■

Ainsi, pour tout $v \in X$,

$$E(v) = \frac{k}{2}\|\partial_x v\|_{L^2}^2 - \int_0^1 fv\,dx \geq \frac{k}{2}\|\partial_x v\|_{L^2}^2 - \|f\|_{L^1}\|v\|_{L^\infty} \geq \frac{k}{2}\|\partial_x v\|_{L^2}^2 - \|f\|_{L^1}\|\partial_x v\|_{L^2}.$$

Or la fonction (polynômiale) $y \mapsto \frac{k}{2}y^2 - \|f\|_{L^1} y$ est minorée sur $\mathbb{R}$ par $-\frac{\|f\|_{L^1}^2}{2k}$.

On a donc montré le résultat attendu

$$\inf_X E \geq -\frac{\|f\|_{L^1}^2}{2k}.$$

Etude d'une suite minimisante Comment peut-on alors montrer la convergence d'une suite minimisante dans cet espace X ?

Une des seules choses que l'on sait sur la suite minimisante, c'est que la suite de nombre réels $(E(u_n))_n$ est bornée et converge vers l'infimum. Est-ce qu'on peut en déduire que la suite $(u_n)_n$ est bornée dans X ?

La réponse à cette question est, en général, non ! On peut par exemple facilement construire une suite $(u_n)_n$ de fonctions de X qui ne soit pas bornée et telle que $(E(u_n))_n$ soit bornée.

Exercice IV.2

On suppose (par exemple) que $f = 0$ sur un intervalle $[\alpha, \beta] \subset [0,1]$. Construire alors une telle suite sous la forme $u_n(x) = \frac{1}{\sqrt{n}}\,\varphi\,(n(x - x_0))$ avec une fonction φ et un point $x_0 \in [0,1]$ bien choisis.

Correction :

On prend $x_0 = \frac{\alpha+\beta}{2}$, φ une fonction $\mathcal{C}_c^\infty(\mathbb{R})$ à support dans $]-1,1[$ non identiquement nulle. Ainsi la fonction u_n est bien nulle en 0 et 1 pour tout n assez grand.

- On a immédiatement, pour n assez grand, $\|u_n\|_{L^\infty} = \frac{\|\varphi\|_{L^\infty}}{\sqrt{n}}$ et $\|u_n'\|_{L^\infty} = \sqrt{n}\|\varphi'\|_\infty$. On a donc bien $\|u_n\|_X \to \infty$ quand $n \to \infty$.
- Calculons l'énergie de u_n

  $$E(u_n) = \frac{k}{2}\int_0^1 |u_n'(x)|^2\,dx - \int_0^1 f(x)u_n(x)\,dx.$$

  Par hypothèse sur f, et par construction de u_n, les supports de f et de u_n sont disjoints et donc le second terme est nul. Il nous reste donc à évaluer la première intégrale. On utilise la définition de u_n et un changement de variable $y = n(x - x_0)$

  $$E(u_n) = \frac{k}{2}\int_{\mathbb{R}} n|\varphi'(n(x-x_0))|^2\,dx = \frac{k}{2}\int_{\mathbb{R}} |\varphi'(y)|^2\,dy.$$

  Ainsi la suite $(E(u_n))_n$ est bien bornée (elle est même constante sur cet exemple ! !).

■

On voit donc qu'il n'est pas du tout évident de tirer une information utile sur la suite $(u_n)_n$ uniquement à partir de la connaissance de la suite des énergies $(E(u_n))_n$.

En théorie de l'optimisation, on dit que la fonctionnelle E n'est pas **coercive** sur $(X, \|\cdot\|_X)$.

On pourrait envisager de changer la norme sur X pour mieux correspondre au problème étudié. Ainsi, si on munit X de la norme suivante

$$\|u\|_{H^1} = \sqrt{\|u\|_{L^2}^2 + \|\nabla u\|_{L^2}^2},$$

alors il est immédiat de voir que l'on récupère une forme de coercivité de E sur $(X, \|\cdot\|_{H^1})$.

Proposition IV.11

Si $(u_n)_n$ est une suite d'éléments de X telle que $(E(u_n))_n$ est bornée, alors $(u_n)_n$ est bornée dans $(X, \|\cdot\|_{H^1})$.

Preuve :

On note $C = \sup_n E(u_n) < +\infty$. On utilise ensuite la définition de l'énergie

$$\frac{k}{2}\|\partial_x u_n\|_{L^2}^2 = E(u_n) + \int_0^1 f(x)u_n(x)\,dx \leq C + \|f\|_{L^1}\|u_n\|_{L^\infty} \leq C + \|f\|_{L^1}\|\partial_x u_n\|_{L^2}.$$

On utilise ensuite l'inégalité de Young suivante (immédiate à vérifier)

$$\forall \varepsilon > 0, \forall a, b \in \mathbb{R}, \quad ab \leq \varepsilon a^2 + \frac{1}{4\varepsilon}b^2.$$

pour majorer le dernier terme de la façon suivante

$$\frac{k}{2}\|\partial_x u_n\|_{L^2}^2 \leq C + \underbrace{\frac{k}{4}}_{=\varepsilon}\|\partial_x u_n\|_{L^2}^2 + \frac{\|f\|_{L^1}^2}{k}.$$

On en déduit

$$\frac{k}{4}\|\partial_x u_n\|_{L^2}^2 \leq C + \frac{\|f\|_{L^1}^2}{k},$$

ce qui donne une borne sur $\|\partial_x u_n\|_{L^2}$. Il nous faut maintenant une borne sur $\|u_n\|_{L^2}$ mais celle-ci s'obtient immédiatement avec le lemme IV.10

$$\|u_n\|_{L^2}^2 = \int_0^1 |u_n|^2\,dx \leq \|u_n\|_{L^\infty}^2 \leq \|\partial_x u_n\|_{L^2}^2.$$

■

Ainsi, avec cette nouvelle norme, on a la coercivité de E et l'on peut en déduire que toute suite minimisante est bornée pour la norme $\|\cdot\|_{H^1}$. C'est un premier pas vers la convergence de la suite. On peut même montrer que cette suite est de Cauchy toujours pour la norme $\|\cdot\|_{H^1}$ (voir plus loin la démonstration).

Malheureusement, en changeant la norme sur X, on a perdu une propriété essentielle : la complétude. On ne peut donc rien déduire du fait que la suite minimisante est de Cauchy.

Exercice IV.3

Vérifier que la suite $(u_n)_n$ définie par

$$u_n(x) = \sqrt{\left(x - \frac{1}{2}\right)^2 + \frac{1}{n}} - \sqrt{\frac{1}{4} + \frac{1}{n}},$$

est une suite de Cauchy non convergente dans $(X, \|\cdot\|_{H^1})$. On pourra utiliser le fait que, d'après le lemme IV.10, la convergence dans cet espace, implique la convergence uniforme sur $[0, 1]$.

Correction :

- Commençons par montrer que la suite $(u_n)_n$ n'est pas convergente dans $(X, \|\cdot\|_{H^1})$. Pour cela, on suppose qu'elle converge vers un certain $u \in X$. D'après le lemme IV.10, ceci implique que, u_n converge uniformément vers u sur $[0, 1]$. En particulier, on aurait la convergence simple de u_n vers u sur $[0, 1]$.
  On voit alors immédiatement que cela implique $u(x) = \left|x - \frac{1}{2}\right| - \frac{1}{2}$. Or, cette fonction u n'est pas de classe $\mathcal{C}^1$ sur $[0, 1]$ à cause de la singularité en $x = \frac{1}{2}$. Ceci établit donc une contradiction.
- Montrons maintenant que $(u_n)_n$ est de Cauchy dans H^1. Pour cela on commence par utiliser le lemme IV.10 pour établir que, u_n et u_{n+p} étant dans X, on a

$$\|u_n - u_{n+p}\|_{H^1}^2 = \|u_n - u_{n+p}\|_{L^2}^2 + \|u_n' - u_{n+p}'\|_{L^2}^2 \leq 2\|u_n' - u_{n+p}'\|_{L^2}^2.$$

Ainsi, pour montrer le critère de Cauchy dans H^1 pour $(u_n)_n$ il suffit de vérifier que (u_n') est de Cauchy dans L^2. Calculons donc $u_n' - u_{n+p}'$

$$\begin{aligned} u_n'(x) - u_{n+p}'(x) &= \frac{x-1/2}{\sqrt{(x-1/2)^2+\frac{1}{n}}} - \frac{x-1/2}{\sqrt{(x-1/2)^2+\frac{1}{n+p}}} \\ &= (x-1/2)\frac{\sqrt{(x-1/2)^2+\frac{1}{n+p}} - \sqrt{(x-1/2)^2+\frac{1}{n}}}{\sqrt{(x-1/2)^2+\frac{1}{n}}\sqrt{(x-1/2)^2+\frac{1}{n+p}}} \\ &= \frac{(x-1/2)\left(\frac{1}{n+p}-\frac{1}{n}\right)}{\sqrt{(x-1/2)^2+\frac{1}{n}}\sqrt{(x-1/2)^2+\frac{1}{n+p}}\left(\sqrt{(x-1/2)^2+\frac{1}{n}}+\sqrt{(x-1/2)^2+\frac{1}{n+p}}\right)}. \end{aligned}$$

On prend la valeur absolue de cette expression et on essaie de majorer intelligemment les différents termes. Comme on sait qu'il n'y a pas convergence uniforme de cette suite de fonctions (car sinon la limite u' serait continue ...), la majoration qu'on doit obtenir doit encore dépendre de x !

De façon plus précise, on utilise les inégalités

$$\sqrt{(x-1/2)^2+1/(n+p)} \geq |x-1/2|$$

et

$$\sqrt{(x-1/2)^2+1/(n+p)} + \sqrt{(x-1/2)^2+1/n} \geq \sqrt{(x-1/2)^2+1/n}.$$

Il vient ainsi

$$\begin{aligned} |u_n'(x) - u_{n+p}'(x)| &\leq \frac{\frac{1}{n}}{(x-1/2)^2+\frac{1}{n}} \\ &= \frac{1}{n}\left(\frac{1}{(x-1/2)^2+\frac{1}{n}}\right)^{\frac{7}{8}}\left(\frac{1}{(x-1/2)^2+\frac{1}{n}}\right)^{\frac{1}{8}} \\ &\leq \frac{1}{n}\left(\frac{1}{\frac{1}{n}}\right)^{\frac{7}{8}}\frac{1}{|x-1/2|^{\frac{1}{4}}} = \frac{1}{n^{1/8}}\frac{1}{|x-1/2|^{\frac{1}{4}}}. \end{aligned}$$

Si maintenant on élève l'inégalité au carré et qu'on l'intègre entre 0 et 1, il vient

$$\int_0^1 |u_n' - u_{n+p}'|^2\,dx \leq \frac{1}{n^{1/4}}\int_0^1 \frac{1}{|x-1/2|^{\frac{1}{2}}}\,dx.$$

Comme l'intégrale qui apparaît dans le membre de droite est convergente (c'est une intégrale de Riemann), on a bien montré

$$\|u_n' - u_{n+p}'\|_{L^2} \leq \frac{C}{n^{\frac{1}{8}}},$$

ce qui prouve bien que la suite $(u_n')_n$ est de Cauchy dans L^2, et donc que $(u_n)_n$ est de Cauchy dans $(X, \|\cdot\|_{H^1})$ d'après la remarque initiale.

■

Bilan : Un choix pour lequel la démarche va fonctionner sera de remplacer X par son **complété** pour la norme $\|\cdot\|_{H^1}$. Cet espace est, *a priori*, un espace abstrait. On va voir dans la suite du chapitre, qu'on peut en fait le construire de façon relativement explicite et travailler dans cet espace de façon finalement naturelle.

II.3 L'espace de Sobolev $H^1(]a,b[)$

Bien qu'on se consacre ici au cas monodimensionnel à fin de comprendre les idées mises en place, il faut retenir que toutes les méthodes se généralisent au cas multi-D.

ATTENTION quand meme : les résultats théoriques sur les espaces de Sobolev peuvent varier selon la dimension de l'espace.

On se place donc en dimension 1 et on travaille sur un intervalle ouvert borné $I =]a, b[$.

Définition et Proposition IV.12

Soit $u \in L^2(I)$ et $g \in L^2(I)$. On dit que g est une **dérivée faible** *de u dans $L^2(I)$ si, pour toute fonction test $\varphi \in \mathcal{C}_c^\infty(I)$, on a la formule*

$$\int_a^b u(x)\varphi'(x)\,dx = -\int_a^b g(x)\varphi(x)\,dx.$$

Si une telle dérivée faible existe, elle est unique (au sens presque partout*) et on la note ∇u ou u' ou $\partial_x u$.*
On appelle **espace de Sobolev** *$H^1(I)$, l'ensemble des fonctions u de $L^2(I)$ qui admettent une dérivée faible dans $L^2(I)$:*

$$H^1(I) = \{u \in L^2(I),\ \exists \nabla u \in L^2(I)\},$$

et on le munit de la norme

$$\|u\|_{H^1(I)} = \sqrt{\|u\|_{L^2}^2 + \|\nabla u\|_{L^2}^2}.$$

Preuve :

La seule chose à démontrer ici, c'est l'unicité de la dérivée faible d'une fonction u. Pour cela, on constate que si $g_1, g_2 \in L^2(I)$ sont deux dérivées faibles de u, on a par définition

$$\int_a^b (g_1(x) - g_2(x))\varphi(x)\,dx = 0,\ \ \forall \varphi \in \mathcal{C}_c^\infty(I).$$

Par densité de $\mathcal{C}_c^\infty(I)$ dans $L^2(I)$, la formule précédente reste valable pour $\varphi \in L^2(I)$, en particulier on peut prendre $\varphi = g_1 - g_2$ ce qui donne

$$\int_a^b |g_1 - g_2|^2\,dx = 0,$$

et donc $g_1 = g_2$. ■

Proposition IV.13

Si u est une fonction de classe $\mathcal{C}^1$ sur $\overline{I} = [a, b]$, alors la dérivée de u au sens classique est aussi l'unique dérivée faible de u, et donc $u \in H^1(I)$. On a donc l'inclusion algébrique

$$\mathcal{C}^1(\overline{I}) \subset H^1(I),$$

mais on a aussi continuité de cette injection, c'est-à-dire que

$$\exists C > 0,\ \ \forall u \in \mathcal{C}^1(\overline{I}),\ \ \|u\|_{H^1} \leq C\|u\|_{\mathcal{C}^1}.$$

Théorème IV.14 (Principales propriétés de $H^1(I)$)

1. *L'espace $H^1(I)$ défini plus haut est un espace de Hilbert.*
2. *Si $u \in H^1(I)$ a une dérivée faible $\partial_x u$ nulle presque partout, alors u est une constante.*
3. *Toute fonction $u \in H^1(I)$ admet un représentant (au sens presque partout) continu sur $\overline{I}$, qu'on note toujours u, et on a*
$$u(y) - u(x) = \int_x^y (\partial_x u)\,dx,\ \ \forall x, y \in \overline{I}.$$
4. *L'ensemble $\mathcal{C}^\infty(\overline{I})$ est dense dans $H^1(I)$.*

Preuve :

1. Il est clair que la norme $\|\cdot\|_{H^1}$ est une norme Hilbertienne associée au produit scalaire H^1 défini par
$$(u, v)_{H^1} = (u, v)_{L^2} + (\partial_x u, \partial_x v)_{L^2} = \int_I uv\,dx + \int_I \partial_x u \partial_x v\,dx.$$
Il reste juste à vérifier la complétude de $H^1(I)$. Soit donc une suite de Cauchy $(u_n)_n$ dans $H^1(I)$. Par définition de la norme H^1, on déduit que $(u_n)_n$ et $(\partial_x u_n)_n$ sont de Cauchy dans $L^2(I)$. Comme $L^2(I)$ est complet, il existe $u, g \in L^2(I)$ telles que
$$u_n \to u,\ \text{ et }\ \partial_x u_n \to g.$$

Pour toute fonction test φ, et pour tout n, on a

$$\int_I u_n(x)\varphi'(x)\,dx = -\int_I \partial_x u_n(x)\varphi(x)\,dx,$$

et il est clair que les convergences établies plus haut permettent de passer à la limite dans cette formule (la fonction test φ étant fixée !). Ceci prouve que $u \in H^1(I)$ et que $\partial_x u = g$.
La convergence de $(u_n)_n$ vers u pour la norme H^1 est alors immédiate.

2. Soit donc une fonction $u \in L^2(\Omega)$ telle que pour toute fonction test φ on a :

$$\int_I u(x)\varphi'(x)\,dx = 0. \tag{IV.7}$$

On fixe une fonction test θ telle que $\int_I \theta(t)\,dt = 1$. Maintenant pour toute fonction-test $\psi \in \mathcal{C}_c^\infty(I)$ on pose

$$\varphi(x) = \int_a^x \psi(t)\,dt - \left(\int_I \psi\right)\left(\int_a^x \theta(t)\,dt\right).$$

Cette fonction est bien sûr de classe $\mathcal{C}^\infty$, et on vérifie qu'elle est également à support compact. De plus on a

$$\varphi'(x) = \psi(x) - (\int_I \psi)\theta(x).$$

Appliquons (IV.7) à la fonction φ ainsi construite, on trouve

$$\int_I u(x)\psi(x)\,dx = \left(\int_I \psi(x)\,dx\right)\underbrace{\left(\int_I u(x)\theta(x)\,dx\right)}_{=m_{u,\theta}}.$$

On a donc obtenu que, pour toute fonction test ψ, on a

$$\int_I \left(u(x) - m_{u,\theta}\right)\psi(x)\,dx = 0.$$

Par densité dans L^2 des fonctions régulières ceci implique que $u - m_{u,\theta}$ est une fonction nulle presque partout, ce qui prouve bien que u est constante.

3. Soit v la fonction définie sur I par

$$v(x) = \int_a^x \partial_x u(t)\,dt.$$

D'après le théorème de convergence dominée de Lebesgue, cette fonction est continue sur $\overline{I}$. Vérifions que $v \in H^1(I)$ et que $\partial_x v = \partial_x u$. Pour cela, on choisit une fonction-test φ et on calcule, en utilisant le théorème de Fubini

$$\int_a^b \left(\int_a^x \partial_x u(t)\,dt\right)\varphi'(x)\,dx = \int_a^b \left(\int_t^b \varphi'(x)\,dx\right)\partial_x u(t)\,dt = -\int_a^b \varphi(t)\partial_x u(t)\,dt.$$

Donc, $u - v$ est une fonction de $H^1(I)$ à dérivée nulle, il existe donc une constante C telle que $u = C + v$ presque partout.
Ceci montre que $C + v$ est un représentant continu de u. En identifiant u à ce représentant continu et en utilisant la formule qui définit v on obtient bien la formule annoncée.

4. Ce résultat est admis et se démontre de la même façon que la densité des fonctions régulières dans $L^2(I)$.

■

Exemples : On prend $I =]-1, 1[$.

- La fonction $x \mapsto u(x) = |x|$ est dans $H^1(I)$ et $\partial_x u$ est la fonction de Heaviside $H \in L^2(I)$ qui vaut 1 sur $]0, 1[$ et -1 sur $]-1, 0[$.
- La fonction de Heaviside H elle-même n'appartient pas à $H^1(I)$ car elle n'admet pas de représentant continu. De fait, on peut vérifier que l'on a

$$\int_{-1}^1 H(x)\varphi'(x)\,dx = -2\varphi(0),$$

cette dernière quantité ne pouvant pas s'écrire comme un élément de $L^2(I)$.

– La fonction $x \mapsto u(x) = |x|^\alpha$ est dans $H^1(I)$ si et seulement si $\alpha > 1/2$ et on a alors $\partial_x u = \alpha |x|^{\alpha-2}x$, pour $x \neq 0$.

Corollaire IV.15

L'injection canonique de $H^1(I)$ dans $\mathcal{C}^0(\overline{I})$ est continue.

Preuve :

D'après l'inégalité de Cauchy-Schwarz on a pour tout x, y

$$|u(y)| \leq |u(x)| + |b-a|^{\frac{1}{2}} \|\partial_x u\|_{L^2}.$$

Si on intègre ceci par rapport à x, on trouve

$$|b-a||u(y)| \leq \int_I |u(x)|\, dx + |b-a|^{\frac{3}{2}} \|\partial_x u\|_{L^2}.$$

Utilisant à nouveau l'inégalité de Cauchy-Schwarz on obtient

$$|b-a||u(y)| \leq |b-a|^{\frac{1}{2}} \|u\|_{L^2} + |b-a|^{\frac{3}{2}} \|\partial_x u\|_{L^2}.$$

En prenant finalement le $\sup$ par rapport à y on trouve

$$\|u\|_{\mathcal{C}^0(\overline{I})} \leq |b-a|^{-\frac{1}{2}} \|u\|_{L^2} + |b-a|^{\frac{1}{2}} \|\partial_x u\|_{L^2}.$$

Ceci prouve bien la continuité de l'application identité. ■

II.4 L'espace $H^1_0(I)$

Définition IV.16

On appelle $H^1_0(I)$, le sous-espace fermé de $H^1(I)$ constitué des fonctions de $H^1(I)$ dont les valeurs sont nulles au bord.

Cette définition a bien un sens car on a vu que les fonctions de $H^1(I)$ ont un unique représentant continu sur $\overline{I}$, ce qui légitime la notion de "valeur au bord" pour des fonctions *a priori* définies seulement presque partout.

Cette espace est un fermé car c'est le noyau de l'application **continue** (d'après le corollaire IV.15) définie par

$$u \in H^1(I) \longmapsto \begin{pmatrix} u(a) \\ u(b) \end{pmatrix} \in \mathbb{R}^2.$$

Le résultat suivant est fondamental dans la théorie. Il est à rapprocher du Lemme IV.10.

Proposition IV.17 (Inégalité de Poincaré)

Pour tout $u \in H^1_0(I)$, on a l'inégalité suivante

$$\|u\|_{L^2} \leq |b-a| \|\partial_x u\|_{L^2}.$$

Corollaire IV.18

L'application $u \mapsto \|\partial_x u\|_{L^2}$ est donc une norme sur $H^1_0(I)$ équivalente à la norme de $H^1(I)$. La structure de Hilbert correspondante est équivalente à la structure héritée de H^1.

La démonstration du corollaire est immédiate. Montrons l'inégalité de Poincaré :

Preuve :

Comme $u(a) = u(b) = 0$. Pour tout $x \in I$, on a

$$|u(x)| \leq \int_a^x |\partial_x u|\, dt \leq |b-a|^{\frac{1}{2}} \|\partial_x u\|_{L^2}.$$

En élevant cette inégalité au carré, puis en l'intégrant sur I, on trouve

$$\int_a^b |u(x)|^2\,dx \leq |b-a|^2 \|\partial_x u\|_{L^2}^2,$$

ce qui donne le résultat. ■

Remarque IV.19

La constante $|b-a|$ qui apparaît dans l'inégalité de Poincaré ci-dessus n'est pas optimale. On peut montrer que la valeur optimale (i.e. la plus petite) de cette constante vaut $\frac{|b-a|}{\pi}$ (voir le TD).

On a enfin le résultat suivant qui est admis.

Théorème IV.20

L'ensemble des fonctions $\mathcal{C}_c^\infty(I)$ est dense dans $H_0^1(I)$.

Ce résultat démontre que $H_0^1(I)$ est bien **le** complété de l'espace $X = \{v : [0,1] \to \mathbb{R}, v(0) = v(1)\}$ introduit plus haut pour la norme H^1. En effet, on a les inclusions

$$\mathcal{C}_c^\infty(I) \subset X \subset H_0^1(I),$$

et donc la proposition ci-dessus établit la densité de X dans $H_0^1(I)$ pour la norme H^1. Cet espace étant complet, on a bien affaire à l'unique complété (à isomorphisme près) de X pour cette norme.

L'espace $H_0^1(I)$ semble donc un bon candidat pour notre analyse.

II.5 Résolution du problème variationnel pour la corde élastique

Revenons au problème qui a motivé toute cette théorie. On a maintenant compris qu'il faut travailler dans l'espace $X = H_0^1(I)$ muni de sa norme. Le problème est donc reformulé de la façon suivante :

Soit $f \in L^2(I)$, trouver $u \in H_0^1(I)$ vérifiant

$$E(u) = \inf_{v \in H_0^1(I)} E(v).$$

L'espace $H_0^1(I)$ est un espace vectoriel et dans ces conditions la preuve de l'unicité d'un éventuel minimum, qu'on a déjà effectuée est encore valable.

Démontrons maintenant l'existence d'un minimum. Pour cela, on va commencer par démontrer que la fonctionnelle E est minorée sur $H_0^1(I)$. En effet pour tout $v \in H_0^1(I)$, nous avons

$$\begin{aligned}\int_I fv\,dx &\leq \|f\|_{L^2}\|v\|_{L^2}, \text{ par Cauchy-Schwarz}\\ &\leq |b-a|\|f\|_{L^2}\|\partial_x v\|_{L^2}, \text{ par Poincaré, car } v \in H_0^1(I)\\ &\leq \frac{1}{2k}|b-a|^2\|f\|_{L^2}^2 + \frac{k}{2}\|\partial_x v\|_{L^2}^2, \text{ par Young.}\end{aligned}$$

Tout ceci montre que

$$E(v) \geq -\frac{|b-a|^2}{2k}\|f\|_{L^2}^2.$$

Comme E est minorée, son infimum est fini et par définition de celui-ci, il existe une suite minimisante, c'est-à-dire une suite de fonctions $(u_n)_n$ dans $H_0^1(I)$ telle que la suite de nombre réels $(E(u_n))_n$ converge vers $\inf_{H_0^1} E$.

Il faut maintenant démontrer que la suite $(u_n)_n$ converge. On va en proposer deux démonstrations. La première est relativement abstraite (mais plus générale !) que la seconde, elle utilise la notion de convergence faible.

Preuve 1 : On commence par démontrer que $(u_n)_n$ est nécessairement bornée dans H_0^1. En effet, comme la suite des énergies $(E(u_n))_n$ converge, il existe $M > 0$ telle que $E(u_n) \leq M$ pour tout n. On a donc

$$\frac{k}{2}\int_I |\partial_x u_n|^2\,dx \leq M + \int_I f u_n\,dx,$$

et par des manipulations semblables aux précédentes, on montre que

$$\frac{k}{4}\int_I |\partial_x u_n|^2\,dx \leq M + \frac{1}{k}|b-a|^2\|f\|_{L^2}^2.$$

Ceci montre bien que la norme $H_0^1(I)$ de $(u_n)_n$ est bornée (on a utilisé l'inégalité de Poincaré).

Comme H_0^1 est un espace de Hilbert, on sait que de toute suite bornée on peut extraire une sous-suite notée $(u_{n_k})_k$ **faiblement convergente** vers un certain $u \in H_0^1$ et que de plus on a

$$\|u\|_{H_0^1} \leq \liminf_{k\to\infty} \|u_{n_k}\|_{H_0^1}.$$

L'injection canonique de $H_0^1(I)$ dans $L^2(I)$ étant continue, la suite $(u_{n_k})_k$ converge aussi faiblement vers u dans $L^2(I)$.

In fine, on a

$$\frac{k}{2}\int_I |\partial_x u|^2\,dx \leq \liminf_{k\to\infty} \frac{k}{2}\int_I |\partial_x u_{n_k}|^2\,dx = \liminf_{k\to\infty}\left(E(u_{n_k}) + \int_I f u_{n_k}\,dx\right).$$

Or, $E(u_{n_k})$ converge vers $\inf_{H_0^1} E$ par définition de la suite minimisante $(u_n)_n$, et l'intégrale $\int_I f u_{n_k}\,dx$ converge vers $\int_I f u\,dx$ par convergence faible dans $L^2(I)$. On a donc démontré que

$$\frac{k}{2}\int_I |\partial_x u|^2\,dx \leq \inf_{H_0^1} E + \int_I f u\,dx,$$

ce qui s'écrit encore

$$E(u) \leq \inf_{H_0^1} E,$$

et donc l'égalité $E(u) = \inf_{H_0^1} E$ est vérifiée, ce qui prouve bien que E admet un (nécessairement unique !) minimiseur sur $H_0^1(I)$.

Preuve 2 : On va directement montrer que $(u_n)_n$ est de Cauchy dans $H_0^1(I)$.

On reprend la démonstration de la propriété d'unicité (avec l'identité du parallèlogramme), pour montrer

$$E\left(\frac{u_n + u_{n+p}}{2}\right) = \frac{1}{2}E(u_n) + \frac{1}{2}E(u_{n+p}) - \frac{k}{8}\|\partial_x u_n - \partial_x u_{n+p}\|_{L^2}^2,$$

d'où l'on déduit

$$\frac{k}{8}\|\partial_x u_n - \partial_x u_{n+p}\|_{L^2}^2 \leq \frac{1}{2}E(u_n) + \frac{1}{2}E(u_{n+p}) - \inf_{v\in H_0^1} E(v). \tag{IV.8}$$

Soit $\varepsilon > 0$, comme $E(u_n)$ tend vers l'infimum de E sur $H_0^1(I)$, il existe $n_0 \geq 0$ tel que

$$\forall n \geq n_0, \quad \left|E(u_n) - \inf_{H_0^1} E\right| \leq \varepsilon.$$

Ainsi, pour $n \geq n_0$ et $p \geq 0$, l'inégalité (IV.8) fournit

$$\frac{k}{8}\|\partial_x u_n - \partial_x u_{n+p}\|_{L^2}^2 \leq \varepsilon.$$

Ceci montre bien que $(u_n)_n$ est de Cauchy dans $H_0^1(I)$, donc elle converge vers une certain u.

On montre ensuite que $E(u) = \lim_{n\to\infty} E(u_n)$ en passant à la limite dans tous les termes de $E(u_n)$ (dans le premier c'est une conséquence de la convergence H^1 et dans le second c'est la convergence L^2 qui donne le résultat). Finalement, comme $(u_n)_n$ est une suite minimisante, on a bien $E(u) = \inf_{H_0^1} E$.

On peut maintenant reformuler les équations d'Euler-Lagrange associées à ce problème de minimisation

$$u \in H_0^1(I) \text{ et vérifie } \forall v \in H_0^1(I), \quad \int_I k\partial_x u \partial_x v\,dx = \int_I f v\,dx. \tag{IV.9}$$

Notons que cette formulation variationnelle admet, elle-aussi, une unique solution. En effet, si u_1 et u_2 sont deux telles solutions, alors leur différence $\bar{u} \in H_0^1(I)$ vérifie la même formulation avec terme source nul. En prenant $v = \bar{u}$ dans la formulation, on obtient

$$k\int_I |\partial_x \bar{u}|^2\,dx = 0,$$

ce qui prouve que $\bar{u}$ est une constante qui ne peut être que 0 car $\bar{u}$ est nulle au bord.

Régularité de la solution : Dans le cadre de la dimension 1, il est facile maintenant de démontrer que la solution $u \in H_0^1(I)$ obtenue est suffisament régulière et vérifie l'EDP de Poisson.

En effet, comme $\mathcal{C}_c^\infty(I) \subset H_0^1(I)$, les équations d'Euler-Lagrange (IV.9) impliquent que

$$\forall \varphi \in \mathcal{C}_c^\infty(I), \ \int_0^1 \partial_x u \partial_x \varphi \, dx = \int_0^1 \frac{f}{k} \varphi \, dx.$$

Par définition, ceci montre que $\partial_x u$ est elle-même une fonction de l'espace de Sobolev $H^1(I)$ qui admet $-\frac{1}{k} f$ comme dérivée faible dans L^2. On a donc, avec des notations évidentes,

$$\partial_x^2 u = \partial_x(\partial_x u) = -\frac{f}{k},$$

ce qui montre bien que u vérifie l'équation aux dérivées partielles $-k\partial_x^2 u = f$ et bien sûr les conditions aux limites (car $u \in H_0^1(I)$).

On peut en dire davantage sur la régularité de u. En effet, comme $\partial_x u \in H^1(I)$, le théorème IV.14 nous dit que $\partial_x u \in \mathcal{C}^0(\overline{I})$ (ou en tout cas, admet un représentant continu). Par ailleurs, ce même théorème nous dit que

$$u(x) = \underbrace{u(0)}_{=0} + \int_0^x u'(t) \, dt,$$

et donc u est la primitive d'une fonction continue, c'est donc bien une fonction de classe $\mathcal{C}^1$ sur tout l'intervalle $\overline{I}$. On remarque, en particulier, que la solution u ainsi obtenue appartient à l'espace X défini au début de l'analyse, et comme $X \subset H_0^1(I)$, la fonction u réalise aussi le minimum de E sur X :

$$E(u) = \inf_{v \in X} E(v).$$

N.B. : Bien que u réalise aussi le minimum de E sur X, il faut bien comprendre que l'on ne pouvait pas établir directement cette propriété en travaillant dans X et que l'introduction des espaces de Sobolev pour résoudre ce problème est absolument cruciale !

Jusqu'à présent, on a seulement utilisé le fait que $f \in L^2(I)$. Si maintenant, on fait l'hypothèse supplémentaire que le terme source f est une fonction continue sur $\overline{I}$, on peut aller loin car on a alors

$$\partial_x^2 u = -\frac{f}{k} \in \mathcal{C}^0(\overline{I}),$$

et ainsi $\partial_x u$ est la primitive d'une fonction continue, c'est donc une fonction $\mathcal{C}^1$ et donc u (qui est une primitive de $\partial_x u$) est une fonction $\mathcal{C}^2$. Ainsi, la fonction u résout bien le problème aux dérivées partielles suivant, au sens classique :

$$\begin{cases} -k\partial_x^2 u = f, \ \text{ dans } I \\ u(0) = u(1) = 0. \end{cases}$$

II.6 Cadre général : théorème de Lax-Milgram

L'analyse précédente montre que le principe général de minimisation de l'énergie d'un problème physique amène naturellement à introduire un cadre fonctionnel (en l'ocurrence Hilbertien) dans lequel on peut écrire une formulation variationnelle (on parle aussi de formulation faible) du problème physique qui mène, sous des hypothèses de régularité de la solution, à une équation aux dérivées partielles avec ces conditions aux limites.

Le processus de raisonnement est donc schématiquement :

Pb physique $\Rightarrow$ Pb variationnel (minimisation de E) $\Rightarrow$ Existence et unicité d'un extremum
$\Rightarrow$ Caractérisation de l'extremum (Euler-Lagrange) $\Rightarrow$ Pb aux dérivées partielles.

D'un point de vue de la théorie des équations aux dérivées partielles (linéaires), ce cheminement intellectuel a été en quelque sorte inversé. En effet, on procède maintenant de la façon suivante : étant donné un problème au bord à résoudre, on commence par poser le bon cadre fonctionnel puis on écrit directement la formulation faible de l'équation (en utilisant des fonctions tests et des intégrations par parties formelles). Ceci étant fait, on démontre l'existence et l'unicité d'une solution à ca problème puis on essaie de prouver la régularité de la solution obtenue afin de pouvoir (parfois) remonter à des solutions classiques du problème de départ.

De fait, la notion "d'énergie" ou de problème de minimisation a essentiellement disparu de ce processus. En effet, l'existence et l'unicité de la solution à la formulation faible peuvent s'obtenir directement grâce à un résultat général qui est le suivant.

Théorème IV.21 (Lax-Milgram)

Soit H un espace de Hilbert. On se donne une forme **bilinéaire continue** $a : H \times H \mapsto \mathbb{R}$. *On suppose de plus que a est coercive, c'est-à-dire qu'il existe $\alpha > 0$ telle que*

$$\forall u \in H, \;\; a(u,u) \geq \alpha \|u\|_H^2.$$

Alors, pour toute forme linéaire continue $L : H \mapsto \mathbb{R}$, il existe un unique $u \in H$ tel que

$$a(u,v) = L(v), \quad \forall v \in H. \tag{IV.10}$$

Si deplus, a est une forme symétrique, alors u est l'unique élément de H réalisant le minimum de la fonctionnelle E définie par

$$E(v) = \frac{1}{2} a(v,v) - L(v), \;\; \forall v \in H.$$

Preuve :

- Cas symétrique : C'est le cas le plus facile et on va en donner deux démonstrations :
  - Démonstration 1, donnant également le dernier point du théorème. On démontre tout d'abord que l'énergie E est minorée sur H, puis que toute suite minimisante $(u_n)_n$ est de Cauchy par l'identité du parallèlogramme. On en déduit la convergence de la suite $(u_n)_n$ vers un certain u qui réalise le minimum de E. On moontre alors comme on l'a fait précédemment sur un exemple que ce minimum vérifie bien (IV.10).
  - Démonstration 2 : Comme a est continue, symétrique et coercive, elle définit un produit scalaire sur H, équivalent au produit scalaire ambiant. Le résultat découle donc immédiatement du théorème de représentation de Riesz dans le Hilbert H muni de ce nouveau produit scalaire.
- Cas non-symétrique : Ce cas est un petit peu plus délicat et nous ne le traiterons pas en cours (Cf. Brézis). Il faut quand même retenir que le résultat est vrai et que la preuve utilise le théorème du point fixe de Banach.

■

Traitons juste un exemple de problème non-symétrique ou le théorème de Lax-Milgram s'applique, bien que le système ne possède pas d'énergie associée. On cherche à résoudre le problème au bord

$$-k\partial_x^2 u + c(x)\partial_x u = f, \quad u(0) = u(1) = 0,$$

où c est une fonction donnée de classe $\mathcal{C}^1$. La formulation faible proposée est

$$\text{Trouver } u \in H_0^1(I) \text{ tel que : } \forall v \in H_0^1(I) \quad \int_I k\partial_x u \partial_x c \, dx + \int_I c(x)(\partial_x u) v \, dx = \int_I f v \, dx. \tag{IV.11}$$

Si on suppose que u est suffisament régulière, on vérifie bien par intégration par parties que le problème initial est effectivement résolu.

Posons donc maintant

$$a(u,v) = \int_I k\partial_x u \partial_x v \, dx + \int_I c(x)(\partial_x u) v \, dx,$$

et

$$L(v) = \int_I f v \, dx.$$

Le fait que L soit une forme linéaire continue sur $H = H_0^1$ ne fait aucun doute (en utilisant l'inégalité de Poincaré par exemple). Vérifions que a est bien une forme bilinéaire continue :

$$|a(u,v)| \leq k\|\partial_x u\|_{L^2}\|\partial_x v\|_{L^2} + \|c\|_\infty \|\partial_x u\|_{L^2}\|v\|_{L^2} \leq (k + \|c\|_\infty C_P)\|u\|_H \|v\|_H,$$

où C_P est la constante de Poincaré du domaine sur lequel on travaille.

Etudions maintenant la coercivité de la forme a en calculant

$$a(u,u) = k\|\partial_x u\|_{L^2}^2 + \int_I c(x)\partial_x u u \, dx.$$

Il nous faut pouvoir estimer le deuxième terme. Plusieurs stratégies sont possibles (donnant des résultats différents mais bien sûr pas contradictoires). En voici une : on aimerait écrire que $(\partial_x u)u = \frac{1}{2}\partial_x(u^2)$ puis intégrer ce terme par parties.

Le problème est : a-t'on le droit de faire un tel calcul pour une fonction u qui est seulement dans $H_0^1(I)$. Il n'est en effet pas évident d'après la définition que les formules de dérivation usuelles d'un produit de deux fonctions soit encore valables.

Lemme IV.22

Pour toute fonction $u \in H_0^1(I)$, on a

$$\int_I c(x)\partial_x u u\, dx = -\frac{1}{2}\int_I (\partial_x c)u^2\, dx. \tag{IV.12}$$

Preuve :

On remarque tout d'abord que si $u \in \mathcal{C}_c^\infty(I)$, alors la formule (IV.12) est parfaitement correcte en intégrant par parties comme suggéré plus haut.

Supposons maintenant $u \in H_0^1(I)$. On utilise la propriété de densité de $\mathcal{C}_c^\infty(I)$ dans $H_0^1(I)$ (Théorème IV.20) qui nous donne l'existence d'une suite $(u_n)_n$ de fonctions de $\mathcal{C}_c^\infty(I)$ qui converge vers u dans $H_0^1(I)$. Pour chaque n fixé, la formule (IV.12) est donc bien valable pour u_n.

$$\forall n \geq 0,\ \int_I c(x)\partial_x u_n u_n\, dx = -\frac{1}{2}\int_I (\partial_x c)u_n^2\, dx.$$

Il reste donc à justifier le passage à la limite dans les deux termes de la formule.

Pour le premier terme, on utilise l'inégalité de Cauchy-Schwarz pour écrire

$$\begin{aligned}\left|\int_I c(x)(\partial_x u_n)u_n\, dx - \int_I c(x)(\partial_x u)u\, dx\right| &\leq \int_I |c(x)||\partial_x u_n - \partial_x u||u_n|\, dx + \int_I |c(x)||\partial_x u||u_n - u|\, dx \\ &\leq \|c\|_{L^\infty}\|\partial_x u_n - \partial_x u\|_{L^2}\|u_n\|_{L^2} + \|c\|_{L^\infty}\|\partial_x u\|_{L^2}\|u_n - u\|_{L^2} \leq M\|u_n - u\|_{H^1},\end{aligned}$$

où M est une constante dépendant de c, de u, et de la borne L^2 de la suite $(u_n)_n$. Par construction de la suite $(u_n)_n$ on a bien convergence vers 0 de ce membre de droite, ce qui justifie le passage à la limite dans le premier terme de l'égalité ci-dessus.

Le second terme se traite de façon tout à fait analogue. ■

La conséquence de ce lemme est donc que, si $\partial_x c \leq 0$ (en particulier si c est constante) on obtient immédiatement la coercivité de a car

$$a(u,u) = k\|\partial_x u\|_{L^2}^2 - \frac{1}{2}\int_I (\partial_x c)u^2\, dx \geq k\|\partial_x u\|_{L^2}^2 \geq C\|u\|_{H^1}^2,$$

la toute dernière inégalité étant une conséquence de l'inégalité de Poincaré (Corollaire IV.18).

Exercice IV.4

Vérifier que a est également coercive sur $H_0^1(I)$, sous l'hypothèse (plus générale) que

$$\sup_I(\partial_x c) < \frac{k}{C_P^2},$$

où C_P désigne la meilleure constante possible dans l'inégalité de Poincaré dans $H_0^1(I)$.

Quoi qu'il soit, sous les différentes hypothèses sur c précédentes, le théorème de Lax-Milgram s'applique et fournit l'existence et l'unicité d'une solution au problème variationnel (IV.11). Il est alors facile de voir que la solution ainsi obtenue vérifie $\partial_x u \in H^1(I)$ et que l'on a

$$\partial_x^2 u = -\frac{f(x) + c(x)\partial_x u}{k},$$

ainsi que les conditions au bord $u(0) = u(1) = 0$.

En particulier, u est une fonction de classe $\mathcal{C}^1$ sur $\overline{I}$. Si on suppose de plus que f est continue, alors u est de classe $\mathcal{C}^2$ et elle vérifie l'équation aux dérivées partielles au sens usuel.

III Schémas numériques en 1D

III.1 Schémas aux différences finies pour l'équation de Poisson

Dans cette section, nous allons nous intéresser aux schémas aux différences finies pour l'équation de Poisson en dimension 1 d'espace, assortie de conditions aux limites de Dirichlet homogènes :

$$\begin{cases} -u''(x) = f(x), \ \ \forall x \in]0,1[, \\ \quad u(0) = u(1) = 0. \end{cases} \tag{IV.13}$$

III.1.1 Construction du schéma

On considère alors un "maillage" de l'intervalle $]0,1[$ c'est-à-dire une subdivision de celui-ci

$$0 = x_0 < x_1 < \cdots < x_n < x_{n+1} = 1.$$

Il y a donc $n+2$ points en comptant les bornes et $n+1$ intervalles. On note $h_{i+1/2} = x_{i+1} - x_i$ qui joue le rôle d'un pas d'espace variable Δx dans le cadre d'un maillage non-uniforme.

Exemple : Le maillage uniforme consiste à poser $\Delta x = \frac{1}{n+1}$ et ensuite $x_i = i \times \Delta x$, ce qui fait que $h_{i+1/2} = \Delta x$ pour tout i.

Le schéma numérique a pour ambition de calculer des valeurs approchées de la solution u du problème (IV.13) aux points de la discrétisation c'est-à-dire de la famille de nombres réels $(u(x_i))_{0\leq i\leq n+1}$. Remarquons de suite que les valeurs de u aux bornes de l'intervalle sont connues car elles sont données par les conditions aux limites $u(x_0) = u(x_{n+1}) = 0$. En réalité, nous avons donc seulement n valeurs inconnues à approcher.

Le principe des méthodes de différences finies consiste à écrire des formules de Taylor en un point x_i, $1 \leq i \leq n$

$$u(x_{i+1}) = u(x_i) + h_{i+1/2}u'(x_i) + \frac{1}{2}h_{i+1/2}^2 u''(x_i) + \cdots ....,$$

$$u(x_{i-1}) = u(x_i) - h_{i-1/2}u'(x_i) + \frac{1}{2}h_{i-1/2}^2 u''(x_i) + \cdots .....,$$

desquelles on déduit une relation approchée

$$\frac{\frac{u(x_{i+1})-u(x_i)}{h_{i+1/2}} - \frac{u(x_i)-u(x_{i-1})}{h_{i-1/2}}}{\frac{1}{2}(h_{i+1/2}+h_{i-1/2})} \approx u''(x_i) = -f(x_i),$$

la dernière inégalité étant vérifie si u est la solution du problème (IV.13).

Le principe de la méthode numérique consiste donc à chercher des nombres réels $(u_i)_{0\leq i\leq n+1}$ qui vérifient d'une part les conditions aux limites $u_0 = u_{n+1} = 0$, et d'autre part les équations

$$-\frac{\frac{u_{i+1}-u_i}{h_{i+1/2}} - \frac{u_i-u_{i-1}}{h_{i-1/2}}}{\frac{1}{2}(h_{i+1/2}+h_{i-1/2})} = f_i, \ \ \forall i \in \{1,\ldots,n\}, \tag{IV.14}$$

où, pour simplifier, on a noté $f_i = f(x_i)$ (c'est une donnée du problème !).

Cas particulier du maillage uniforme : Si le maillage est uniforme, alors tous les $h_{i+1/2}$ sont égaux à Δx et on obtient le schéma suivant

$$-\frac{u_{i+1} - 2u_i + u_{i-1}}{\Delta x^2} = f_i, \quad \forall i \in \{1,\ldots,n\}.$$

Ce schéma est très classique et c'est celui que vous devez connaître par coeur.

III.1.2 Existence et unicité de la solution du schéma. Principe du maximum discret.

Le schéma numérique que nous avons écrit est une famille de n équations linéaires mettant en jeu n inconnues réelles : les $(u_i)_{1\leq i\leq n}$ car u_0 et u_{n+1} sont égaux à 0. Si on note $U = (u_i)_{1\leq i\leq n}$ le vecteur des inconnues et $F = (f_i)_{1\leq i\leq n}$ le vecteur contenant le terme source, on voit que le schéma aux différences finies s'écrit sous la forme

$$AU = F,$$

où A est une matrice carrée de taille n que l'on peut écrire explicitement. Cette matrice s'appelle la matrice du schéma et on va en étudier les principales propriétés.

Théorème IV.23

1. *A est une matrice tridiagonale.*
2. *Soient $v, b \in \mathbb{R}^n$ vérifiant $Av = b$, alors on a :*
$$b \geq 0 \Longrightarrow v \geq 0.$$
En conséquence, A st inversible et la matrice A^{-1} a tous ses coefficients positifs.
3. *La matrice A est symétrique et définie positive pour le produit scalaire sur $\mathbb{R}^n$ défini par*
$$(u, v) = \sum_{i=1}^{n} \frac{1}{2}(h_{i-1/2} + h_{i+1/2}) u_i v_i = \sum_{i=1}^{n} \frac{x_{i+1} - x_{i-1}}{2} u_i v_i. \tag{IV.15}$$

Preuve :

1. Il est clair que la i-ième ligne de la matrice n'a des coefficients non nuls que pour les colonnes $i-1$, i, $i+1$ ce qui montre que A est tridiagonale.
2. Soient v et b comme dans l'énoncé. On pose $v_0 = v_{n+1} = 0$ de sorte que v et b vérifient les équations
$$-\frac{\frac{v_{i+1}-v_i}{h_{i+1/2}} - \frac{v_i - v_{i-1}}{h_{i-1/2}}}{\frac{1}{2}(h_{i+1/2} + h_{i-1/2})} = b_i, \ \ \forall i \in \{1, \dots, n\}.$$
Supposons, par l'absurde, que v ne soit pas positif. Il existe donc un plus petit indice $i_0 \in \{1, \dots, n\}$ tel que $v_{i_0} = \inf_j(v_j) < 0$. Ecrivons l'équation pour l'indice i_0, on trouve
$$\frac{\frac{v_{i_0}-v_{i_0+1}}{h_{i_0+1/2}} + \frac{v_{i_0}-v_{i_0-1}}{h_{i_0-1/2}}}{\frac{1}{2}(h_{i_0+1/2} + h_{i_0-1/2})} = b_{i_0} \geq 0.$$
Ceci n'est pas possible car le membre de gauche est strictement négatif d'après le choix de l'indice i_0.
On a donc démontré que si $Av \geq 0$, alors $v \geq 0$. Ceci implique que le noyau de A est trivial car si v est un vecteur vérifiant $Av = 0$, on a à la fois $Av \geq 0$ et $A(-v) \geq 0$ et donc $v \geq 0$ et $-v \geq 0$, ce qui prouve que $v = 0$. La matrice A est donc inversible. Chaque colonne de A^{-1} est la solution du système $Av = b$ avec le second membre b égal à l'un des vecteurs de la base canonique. Ces vecteurs étant positifs, on en déduit que les vecteurs colonnes de A^{-1} sont également positifs.
3. Effectuons le calcul de (Au, v) :
$$\begin{aligned}
(Au, v) &= \sum_{i=1}^{n} \left(\frac{u_i - u_{i+1}}{h_{i+1/2}} + \frac{u_i - u_{i-1}}{h_{i-1/2}} \right) v_i \\
&= \sum_{i=0}^{n} \frac{u_i - u_{i+1}}{h_{i+1/2}} v_i + \sum_{i=1}^{n+1} \frac{u_i - u_{i-1}}{h_{i-1/2}} v_i \\
&= \sum_{i=0}^{n} \frac{u_i - u_{i+1}}{h_{i+1/2}} v_i + \sum_{i=0}^{n} \frac{u_{i+1} - u_i}{h_{i+1/2}} v_{i+1} \\
&= \sum_{i=0}^{n} \frac{(u_{i+1} - u_i)(v_{i+1} - v_i)}{h_{i+1/2}}.
\end{aligned}$$
Ceci montre immédiatement que $(Au, v) = (Av, u)$, et que de plus
$$(Au, u) = \sum_{i=0}^{n} \frac{(u_{i+1} - u_i)^2}{h_{i+1/2}} = \sum_{i=0}^{n} h_{i+1/2} \left(\frac{u_{i+1} - u_i}{h_{i+1/2}} \right)^2,$$
donc A est positive et si $(Au, u) = 0$, on trouve que u est une constante qui est forcément nulle car $u_0 = u_{n+1} = 0$.

■

Remarque IV.24

La matrice du Laplacien en 1D sur un maillage régulier s'écrit

$$A = \frac{1}{h^2}\begin{pmatrix} 2 & -1 & 0 & \cdots & \cdots \\ -1 & 2 & -1 & 0 & \vdots \\ 0 & -1 & 2 & -1 & 0 \\ \vdots & 0 & -1 & 2 & -1 \\ \cdots & \cdots & 0 & -1 & 2 \end{pmatrix}$$

Cette matrice est particulièrement célèbre et il faut savoir la reconnaître et donc connaître ces propriétés. On peut en particulier calculer explicitement ces éléments propres ...

Notons que dans tous les cas, la matrice du schéma obtenue est dite **creuse**, c'est-à-dire qu'elle a peu de coefficients non nuls. En effet, il s'agit d'une matrice de taille $n \times n$ (donc *a priori* n^2 coefficients) mais dont seulement $3n-2$ coefficients sont non nuls. Cette propriété est cruciale d'un point de vue de la mise en oeuvre des schémas comme nous le verrons en TP.

III.2 Consistance. Stabilité. Convergence et estimation d'erreur.

III.2.1 Consistance.

On peut définir comme on l'a fait pour les schémas numériques pour le transport la notion de consistance du schéma :

Définition IV.25

L'erreur de consistance $R = (R_i)_{1\leq i\leq n}$ du schéma aux différences finies est définie par

$$R_i = -\frac{\frac{u(x_{i+1})-u(x_i)}{h_{i+1/2}} - \frac{u(x_i)-u(x_{i-1})}{h_{i-1/2}}}{\frac{1}{2}(h_{i+1/2}+h_{i-1/2})} + f(x_i),$$

où u est la solution du problème étudié.
On dit que la consistance est d'ordre p en norme $\|\cdot\|$ si on a

$$\|R\| \leq C\Delta x^p,$$

pour tout maillage vérifiant $\sup_i h_{i+1/2} \leq \Delta x$, où la constante C dépend de la solution u, du second membre f mais pas du maillage.

Proposition IV.26

Le schéma aux différences finies proposé plus haut est consistant à l'ordre 1 en norme infinie en général, et à l'ordre 2 si on se restreint aux maillages uniformes. plus précisément on a

$$\|R\|_\infty \leq C\Delta x \|u^{(3)}\|_\infty, \quad \textit{dans le cas général},$$

$$\|R\|_\infty \leq C\Delta x^2 \|u^{(4)}\|_\infty, \quad \textit{dans le cas uniforme}.$$

Corollaire IV.27

Dans le cas général, le schéma est exact si u est un polynôme de degré 2 (autrement dit si f est constante).
Dans le cas uniforme, le schéma est exact si u est un polynôme de degré 3 (autrement dit si f est affine).

III.2.2 Stabilité.

Là encore, comme dans le chapitre précédent, on définit "formellement" la stabilité d'un schéma comme le fait que la solution du schéma soit bornée indépendamment du paramètre Δx en fonction de la donnée.

Définition IV.28

On dit que le schéma aux différences finies est stable pour une norme $\|\cdot\|$ s'il existe une constante $C > 0$, indépendante du maillage telle que

$$\forall b \in \mathbb{R}^n, \ \ \|A^{-1}b\| \leq C\|b\|.$$

On peut maintenant démontrer la stabilité L^∞ du schéma ci-dessus.

Proposition IV.29

Le schéma aux différences finies étudié plus haut est L^∞ stable. Plus précisément on a

$$\|A^{-1}b\|_\infty \leq \frac{1}{8}\|b\|_\infty, \quad \forall b \in \mathbb{R}^n.$$

Preuve :

On a vu que pour tout maillage, le schéma est exact sur les polynômes de degré 2. Considérons donc la fonction $v(x) = x(1-x)$ qui vérifie bien les conditions aux limites, qui est un polynôme de degré 2 et tel que

$$-v''(x) = 2, \ \ \forall x \in]0,1[.$$

En conséquence, le schéma est exact sur cette fonction, ce qui signifie que si on pose $V = (v(x_i))_{1\leq i\leq n}$ et $D = (1)_{1\leq i\leq n}$ le vecteur constant égal à 1, nous avons la relation

$$AV = 2D.$$

Nous avons maintenant pour tout $b \in \mathbb{R}^n$, les inégalités entre vecteurs

$$-\|b\|_\infty D \leq b \leq \|b\|_\infty D.$$

Comme on a démontré que A^{-1} est une matrice ayant tous ses coefficients positifs, on peut appliquer la matrice A à cette inégalité, et on obtient

$$-\frac{\|b\|_\infty}{2}V \leq A^{-1}b \leq \frac{\|b\|_\infty}{2}V.$$

Or on a $\|V\|_\infty \leq \frac{1}{4}$ et donc

$$\|A^{-1}b\|_\infty \leq \frac{1}{8}\|b\|_\infty.$$

■

De même, on peut vérifier la stabilité L^2. Pour cela, on définit la norme L^2 de la façon suivante

$$\|U\|_2 = \left(\sum_{i=1}^n \frac{1}{2}(h_{i+1/2} + h_{i-1/2})|u_i|^2\right)^{\frac{1}{2}},$$

qui est la norme associée au produit scalaire défini par la formule (IV.15).

Proposition IV.30

Le schéma aux différences finies étudié plus haut est L^2-stable. Plus précisément on a

$$\forall b \in \mathbb{R}^n, \ \ \|A^{-1}b\|_2 \leq \|b\|_2.$$

Preuve :

D'après les calculs effectués dans la démonstration du théorème IV.23, si on pose $U = A^{-1}b$, on a

$$\sum_{i=0}^n h_{i+1/2}\left(\frac{u_{i+1}-u_i}{h_{i+1/2}}\right)^2 = (AU,U) = (b,U) \leq \|b\|_2\|U\|_2.$$

Par ailleurs, pour tout $1 \leq i \leq n$, on a

$$u_i = \sum_{j=0}^{i-1}(u_{j+1}-u_j) = \sum_{j=0}^{i-1} h_{j+1/2}\frac{u_{j+1}-u_j}{h_{j+1/2}},$$

et donc

$$|u_i| \leq \sum_{j=0}^{n} h_{j+1/2}\left|\frac{u_{j+1}-u_j}{h_{j+1/2}}\right|.$$

En utilisant l'inégalité de Cauchy-Schwarz, on en déduit

$$|u_i| \leq \left(\sum_{j=0}^{n} h_{j+1/2}\left(\frac{u_{j+1}-u_j}{h_{j+1/2}}\right)^2\right)^{\frac{1}{2}} \underbrace{\left(\sum_{j=0}^{n} h_{j+1/2}\right)^{\frac{1}{2}}}_{=1} \leq \|b\|_2^{\frac{1}{2}}\|U\|_2^{\frac{1}{2}}.$$

Il vient

$$\|U\|_2^2 = \sum_{i=1}^{n}\frac{1}{2}(h_{i+1/2}+h_{i-1/2})|u_i|^2 \leq \underbrace{\left(\sum_{i=1}^{n}\frac{1}{2}(h_{i+1/2}+h_{i-1/2})\right)}_{\leq 1}\|b\|_2\|U\|_2,$$

et donc

$$\|U\|_2 \leq \|b\|_2.$$

■

III.2.3 Estimation d'erreur.

De façon similaire à l'étude effectuée pour le transport, on note $\overline{U} = (u(x_i))_{1\leq i\leq n}$ le vecteur constitué des valeurs de la solution exacte u du problème étudié sur les points du maillage. On note U le vecteur contenant les valeurs approchées et obtenu en résolvant le système linéaire $AU = F$. On note enfin $E = \overline{U} - U$ le vecteur **erreur**.

Par définition de l'erreur de consistance, on a la relation

$$A\overline{U} = F + R.$$

En soustrayant cette identité à la définition de U, on trouve

$$AE = R.$$

D'après l'estimation de stabilité, on a

$$\|E\|_\infty = \|A^{-1}R\|_\infty \leq \frac{1}{8}\|R\|_\infty.$$

Enfin, en utilisant l'estimation de l'erreur de consistance, on a

$$\|E\|_\infty \leq C\Delta x\|u^{(3)}\|_\infty, \text{ dans le cas général,}$$

$$\|E\|_\infty \leq C\Delta x^2\|u^{(4)}\|_\infty, \text{ pour un maillage uniforme.}$$

On a donc démontré la convergence du schéma en norme L^∞, celle-ci étant d'ordre 1 en général et d'ordre 2 sur des maillages uniformes.

III.3 Exemple de non-stabilité

Si on essaie d'approcher le problème $-u'' - \pi^2 u = f$ sur $]0,1[$ avec des CL de Dirichlet homogène (qui est **mal posé !** comme on l'a vu en TD) on peut être amenés naturellement à considérer le schéma numérique suivant

$$AU - \pi^2 U = F,$$

où A est la matrice du Laplacien sur maillage uniforme par exemple (i.e. celle décrite dans le paragraphe précédent).

Il se trouve que l'on connait parfaitement les éléments propres de A qui sont donnés par

$$\lambda_k = \frac{4}{\Delta x^2}\sin^2\left(\frac{k\pi\Delta x}{2}\right),\ U_k = \big(\sin(k\pi i\Delta x)\big)_i,$$

de sorte qu'on a

$$AU_k = \lambda_k U_k.$$

En particulier, π^2 n'est pas une valeur propre de A et donc $A - \pi^2 \,\mathrm{Id}$ est une matrice inversible, ce qui prouve que le schéma précédent est bien posé pour toute valeur de Δx. De plus, on vérifie qu'il est bien entendu consistant à l'ordre 2 en norme infinie (pour peu que la solution existe, ce qui n'est pas toujours le cas).

Montrons néanmoins que ce schéma n'est pas L^∞-stable. En effet, considérons le vecteur $U = U_1$ qui est bien non nul

Nous avons

$$AU - \pi^2 U = \left(\frac{4}{\Delta x^2} \sin^2 \left(\frac{\pi \Delta x}{2} \right) - \pi^2 \right) U,$$

et donc

$$(A - \pi^2 \,\mathrm{Id})^{-1} U = \frac{1}{\left(\frac{4}{\Delta x^2} \sin^2 \left(\frac{\pi \Delta x}{2} \right) - \pi^2 \right)} U.$$

On en déduit que

$$\|(A - \pi^2 \,\mathrm{Id})^{-1}\|_\infty \geq \frac{\|(A - \pi^2 \,\mathrm{Id})^{-1} U\|_\infty}{\|U\|_\infty} = \frac{1}{\left| \frac{4}{\Delta x^2} \sin^2 \left(\frac{\pi \Delta x}{2} \right) - \pi^2 \right|} \underset{\Delta x \to 0}{\sim} \frac{24}{\pi^4 \Delta x^2},$$

et donc le schéma n'est évidemment pas stable car cette norme explose quand on raffine le maillage.