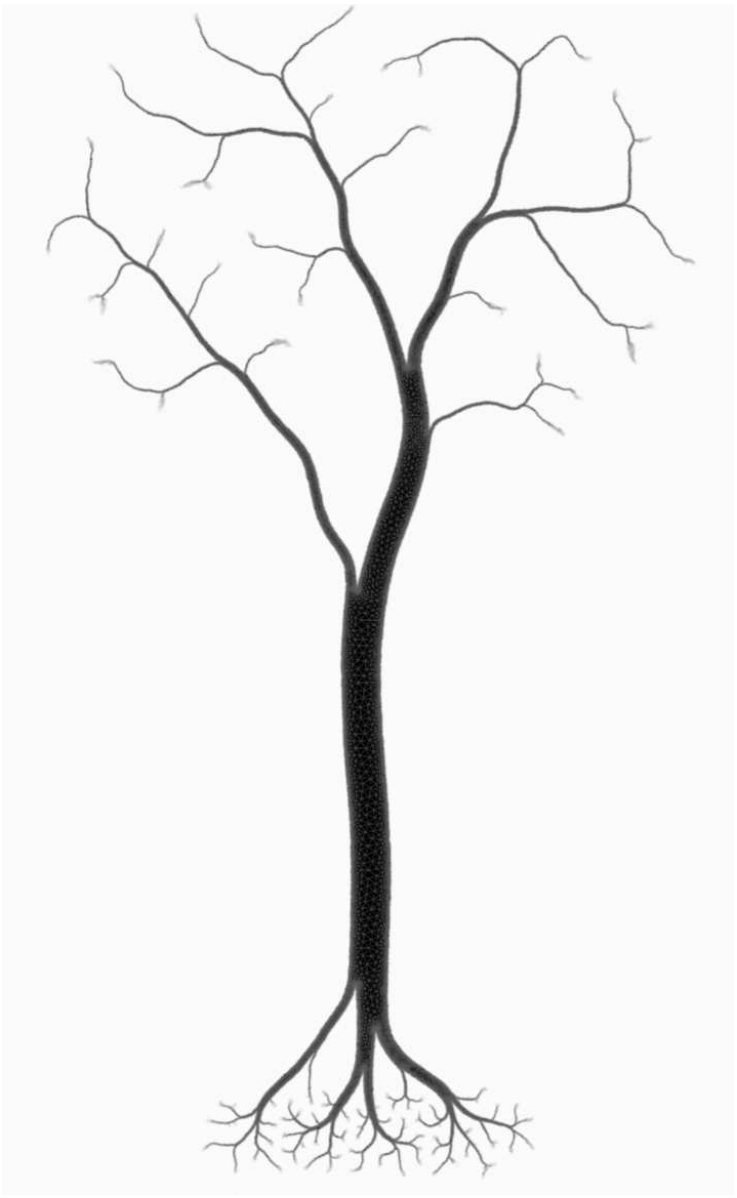
Modélisation



B. Maury

.

Avertissement

Ce document a été réalisé en accompagnement d'un cours d'une cinquantaine d'heures donné à l'École Normale Supérieure au printemps 2016. Il s'agit d'une première version, écrite au fil des cours, qui présente sans aucun doute de multiples imperfections ou coquilles.

Les modèles particuliers abordés reflètent de façon évidente les activités de recherche passées et présentes de l'auteur, mais nous espérons que leur étude peut permettre d'acquérir des connaissances et principes généraux qui pourront être mis en œuvre de façon féconde dans d'autres contextes.

La structure, si tant est que l'on puisse parler de structure, est en revanche volontaire est assumée.

La partie IV regroupe des éléments théoriques classiques qui sont utilisés dans le reste de l'ouvrage.

La partie III présente différentes méthodologies [1] liées à la résolution numérique d'Équations aux Dérivées Partielles ou de problèmes d'optimisation avec et sans contraintes. .

La partie II traite, de façon plus ou moins formalisée selon le sujet, de notions générales en modélisation mathématique, et d'interprétations de concepts théoriques dans un contexte de modélisation.

La partie I développe un certain nombre de modèles, essentiellement autour des phénomènes de transport.

Au delà de ce regroupement en quatre parties, les sections ne sont pas ordonnées autrement qu'alphabétiquement.

1. Différences finies et éléments finis, qui pourraient être complétés dans l'avenir par une section sur les méthodes de volumes finis

Table des matières

Première partie

Modèles

1 Conservation, transport, et diffusion

1.1 Vecteur flux, équation de conservation

On s'intéresse ici à la description de la distribution d'une substance dans l'espace au cours du temps, décrite par sa densité $\rho(x,t)$.

Definition 1.1. *(Vecteur flux)*
Soit x un point du domaine occupé par la substane, n un vecteur unitaire, et $D_\varepsilon(n)$ un disque (ou un segment s'il s'agit de la dimension 2) centré en x, d'aire ε (de longueur ε en dimension 2), et normal à n. On note $J(\varepsilon,n)$ la quantité de substance qui traverse D_ε par unité de temps, comptée positivement dans le sens n. S'il existe un vecteur J tel que, pour tout n, la quantité $J(\varepsilon,n)/\varepsilon$ tende vers une limite quand ε tend vers 0, et que cette limite s'écrive $J\cdot n$, on appelle $J = J(x)$ le vecteur flux en x.

Cette définition formelle, à la base de toutes les équations aux dérivées partielles qui expriment la conservation d'une certaine quantité, n'a en fait pas un sens très clair. En premier lieu, pour tous les phénomènes réels impliquant des *particules*[2], elle n'a de sens que si le diamètre du disque n'est pas trop petit vis à vis des tailles caractéristiques du phénomène microscopique étudié[3]. La notion n'a en particulier par de sens si $\sqrt{\varepsilon}$ ($\approx$ diamètre du disque $D_\varepsilon(n)$) est de l'ordre de la distance interparticulaire, ou plus petit. Par ailleurs, l'expression *par unité de temps* sous-entend que l'on fait le bilan sur un intervalle de temps petit, mais suffisamment grand pour laisser passer un nombre significatif d'entités. Pour que cette notion ait un sens, il faut par ailleurs que ε et le temps d'intégration ne soient pas trop grands. Si en divisant par exemple ε par deux, on trouve une valeur significativement différente, c'est que la fenêtre d'observation est trop grande. De façon générale, cette notion n'aura de sens que pour des plages de tailles et temps caractéristiques adaptées au problème considéré. Ces plages peuvent être très étroites dans le cas par exemple du trafic routier ou piétons ; le rapport entre l'échelle macroscopique (taille caractéristique du domaine étudié, tronçon de route ou couloir dans un bâtiment), et l'échelle microscopique (taille des entités considérées, et / ou des distances entre elles) n'est pas très grand, de l'ordre de 10^2 dans certains cas. La situation est évidemment plus favorable pour des systèmes de particules du type gaz, avec une échelle macroscopique de l'ordre du mètre, et microscopique de l'ordre de 10^{-10} m (taille des molécules) ou 5×10^{-9} m (distance entre molécules).

Remarque 1.2. *On peut se demander quelle est la nature de l'objet mathématique qui résulterait de l'application* à la lettre *de la définition 1.1, dans le cas où l'on a un nombre fini de particules, de masses m_i et vitesses $u_i(t)$, $i = 1,\ldots,N$. En dimension 1, considérons le cas d'une particule de masse m parcourant la trajectoire $t\mapsto X(t)$, et donc animée d'une vitesse*

3. L'aire ε tend vers 0, mais *pas trop* ...

3. Particules dans un sens très large : il peut s'agir de particules physiques de type molécules, ou d'entités de taille plus importante comme des cellules, des voitures pour les équations du trafic routier, ou des piétons.

$V(t) = \dot{X}(t)$. *On peut approcher cette particule par une particule de taille finie, de densité uniforme* m/ε *sur* $]X(t), X(t)+\varepsilon[$. *Le flux est alors défini en* (x,t) *par*

$$J_\varepsilon(x,t) = V\frac{m}{\varepsilon}\mathbb{1}_{]X(t),X(t)+\varepsilon[}.$$

A t *fixé,* J_ε *converge donc (au sens des mesures, ou au sens des distributions) vers* $mV\delta_{X(t)}$. *Si l'on se fixe un intervalle en temps, on peut aussi voir* J *comme une mesure en espace-temps, qui converge vers une mesure singulière supportée par la trajectoire, avec*

$$\langle J\,,\varphi\rangle = \int_0^T m\varphi(X(t),t)\,dt,$$

pour $\varphi \in C_c^\infty([0,T],\mathbb{R})$, *où, si l'on introduit l'abscisse curviligne sur la trajectoire* Σ, *comme une mesure singulière de densité linéique* $mV/\sqrt{1+V^2}$. *En dimension supérieure, on pourra de la même manière identifier le vecteur flux à une mesure vectorielle singulière suportée par la trajectoire, avec une même expression pour la densité linéique (où* V *est maintenant un vecteur de* $\mathbb{R}^d$*). Le vecteur flux pour une collection de particules peut ainsi se voir comme une somme de mesure singulières portées par les trajectoires dans l'espace-temps.*

Remarque 1.3. *On peut, d'une certain manière, rendre statique le problème d'évolution en le considérant comme un problème posé sur l'espace-temps. Toute entité viellit à la vitesse de* 1 *(sans unité : il s'agit de* secondes par seconde*). Une solution de l'équation de conservation peut alors se voir comme une densité* $\rho(x,t)$ *telle que le champ* $F = (\rho\times 1, J)$ *est à divergence nulle en espace temps :*

$$\nabla_{t,x}\cdot F = \partial_t\rho + \nabla_x\cdot J = 0.$$

Nous privilégierons néanmoins dans ce qui suit l'approche consistant à distinguer la variable de temps, de telle sorte que $\nabla\cdot$ *représentera bien la divergence vis-à-vis de la variable d'espace.*

Équation de conservation. On considère une substance qui se propage selon le vecteur flux J. On écrit que la dérivée en temps de la quantité de substance N_ω contenue dans un sous-domaine ω immobile est égal au bilan instantané des flux à travers la frontière.

$$\frac{dN_\omega}{dt} = \frac{d}{dt}\int_\omega \rho(x,t)\,\mathrm{dx} = -\int_{\partial\omega} J\cdot n = -\int_\omega \nabla\cdot J.$$

Cette identité étant vérifiée pour tout ω, on en déduit l'équation

$$\frac{\partial\rho}{\partial t} + \nabla\cdot J = 0. \tag{1.1}$$

Terme source. On peut intégrer à ce modèle des termes-source (ou termes-puits si l'on enlève de la matière), en considérant une quantité f de matière injectée par unité de temps et par unité de volume. Le bilan instantané de matière sur un volume ω s'écrit alors

$$\frac{d}{dt}\int_\omega \rho = -\int_{\partial\omega} J\cdot n + \int_\omega f,$$

ce qui conduit à l'équation

$$\frac{\partial\rho}{\partial t} + \nabla\cdot J = f.$$

1.2 Transport

Modèle 1.4. *(Équation de continuité)*
On considère une substance décrite par sa densité $\rho(x,t)$, et convectée par un champ de vitesse u. Le vecteur flux s'écrit $J = \rho u$, et l'équation correspondante est

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho u) = f.$$

Cette équation est parfois appelée équation *de transport conservatif.*

Remarque 1.5. *Dans le cas où le champ convectant est à divergence nulle, l'équation s'écrit*

$$\frac{\partial \rho}{\partial t} + u \cdot \nabla \rho = 0,$$

c'est cette dernière équation qui est le plus couramment appelé équation de transport. On prendra garde cependant au fait qu'elle correspond (dans cas où le champ n'est pas à divergence nulle) au transport d'une quantité non extensive. *Elle n'exprime ainsi pas le transport d'une quantité de matière, mais d'une variable de type intensif, comme un signal, une caractéristique intrinsèque d'entité transportée, un label, une information, typiquement des variables qui ne se* somment *pas.*

Flot d'un champ de vecteur et équation de transport conservative . On peut vérifier que cette équation exprime au niveau macroscopique et eulérien le transport effectif d'une mesure par le flot associé à un champ de vecteurs.

On considère un champ de vecteur régulier $u_t(x)$ dans $\mathbb{R}^d$, et le flot associé $X_t(x,s)$, défini par

$$\left\{ \begin{array}{rcl} \dfrac{\partial X_t}{\partial t}(x,s) & = & u_t(X_t(x,s)) \\ X_s(x,s) & = & x. \end{array} \right. \tag{1.2}$$

Pour t fixé, le flot

$$x \in \mathbb{R}^d \longmapsto X_t(x,0) \in \mathbb{R}^d$$

transporte une mesure donnée ρ_0 vers une nouvelle mesure notée ρ_t. De façon plus générale ρ_t est la mesure image de ρ_s par l'application $X_t(\,\cdot\,,s)$.

Pour toute fonction régulière $\varphi \in \mathcal{D}(\mathbb{R}^d) = C_c^\infty(\mathbb{R}^d)$, on a en particulier

$$\int_{\mathbb{R}^d} \varphi(y)\rho_t(y)\,dy = \int_{\mathbb{R}^d} \varphi(X_t(x,s))\rho_s(x)\,dx.$$

En dérivant cette identité par rapport au temps t, puis en prenant $s = t$, on obtient

$$\int_{\mathbb{R}^d} \varphi(y)\partial_t\rho_t(y)\,dy = \int_{\mathbb{R}^d} \nabla\varphi(X_t(x,s)) \cdot u_t(X_t(x,s))\,\rho_s(x)\,dx$$

$$= \int_{\mathbb{R}^d} \nabla\varphi(x) \cdot u_t(x)\,\rho_s(x)\,dx = -\int_{\mathbb{R}^d} \varphi(x)\nabla \cdot (u_t(x)\,\rho_t(x))\,dx,$$

d'où

$$\int_{\mathbb{R}^d} \varphi(x)\,(\partial_t\rho_t(x)) + \nabla \cdot (u_t(x)\,\rho_t(x)))\,dx = 0.$$

Cette identité étant valable pour tout instant t, pour tout fonction test φ, on en déduit formellement l'équation de transport conservatif (ou équation de continuité)

$$\partial_t \rho_t + \nabla \cdot (u_t \, \rho_t) = 0.$$

Dans le cas régulier, toute solution de l'équation de transport non conservative (ou conservative avec un champ à divergence nulle) est constante le long des caractéristiques :

Proposition 1.6. *Soit ρ_t une solution régulière de l'équation*

$$\partial_t \rho_t + u_t \cdot \nabla \rho_t = 0,$$

avec u_t régulier (continu, et Lipschitzien par rapport à la variable d'espace). Alors ρ_t est constant le long des caractéristiques définies par (1.2)*.*

Démonstration. On a

$$\frac{d}{dt}\rho_t(X_t(x,s)) = \partial_t \rho_t + \frac{d}{dt} X_t(x,s) \cdot \nabla \rho_t = \partial_t \rho_t + u_t \cdot \nabla \rho_t = 0.$$

□

On en déduit directement, toujours dans le cas régulier, l'expression de la solution de l'équation de transport conservative :

Proposition 1.7. *Soit ρ_t une solution de l'équation*

$$\partial_t \rho_t + \nabla \cdot (\rho_t u_t) = 0,$$

avec u_t régulier (continu et Lipschitz par rapport à la variable d'espace). Alors ρ_τ vérifie

$$\rho_\tau(X_\tau(x,s)) = \rho_s(x) \exp\left(- \int_s^\tau \nabla \cdot u_t(X_t(x,s)) \, dt \right).$$

Noter que l'on peut ainsi exprimer ρ_τ à partir d'une donnée initiale en renversant le flot :

$$\rho_\tau(y) = \rho_0(X_0(y,\tau)) \exp\left(- \int_0^\tau \nabla \cdot u_t(X_t(y,0)) \, dt \right).$$

Remarque 1.8. *En termes de modélisation, on peut voir l'équation de transport de différentes manières, qui conditionnent le sens que l'on peut souhaiter donner aux solutions. La première consiste à se donner un champ de vitesse, et une densité initiale, et à étudier le transport de la densité par le champ. C'est sous cette forme-là que le problème est classiquement étudié d'un point de vue théorique (voir plus loin). Cette situation correspondrait par exemple à l'écoulement d'un fluide qui remplit un certain domaine. On injecte alors dans ce fluide un* traceur passif, *c'est à dire une substance dont on peut suivre le mouvement, mais qui n'a pas d'incidence sur ce dernier. La densité considérée est alors celle du traceur passif. Dans ce premier cas le champ est bien défini indépendamment de la matière (traceur) qu'il transporte. On a toujours une solution particulière, d'un intérêt limité, qui exprime le transport d'une quantité nulle de traceur par le champ de vitesse sous-jacent.*

Une deuxième vision correspondrait à des particules qui évoluent dans le vide (ou dans l'air, dont on pourra négliger les effets dans certains régimes), qui éventuellement intéragissent entre elles, sont soumises à l'action de forces extérieures, etc.... Si l'on connait le champ de vitesse, on souhaite écrire le transport de la matière par le champ de vitesse. Mais ce dernier n'est de façon évidente défini que là où il y a de la matière, il n'est pas donné a priori en tout point de l'espace. D'un point de vue mathématique, le problème est très différent. Les questions typiques que l'on peut se poser sont les suivantes : étant donnée une famille de mesures (ρ_t), existe-t-il un champ de vitesse qui transporte ρ_t ? Est-il ρ_t-presque partout unique ? C'est la version mathématique du problème de l'expérimentateur qui cherche à estimer des vitesses à partir d'observations en termes de positions (de particules, cellules, individus dans une foules, voitures, voire planètes). Dans ce contexte, les champs de vitesses n'ont en général aucune raison de présenter la moindre régularité d'un point de vue Eulérien. *C'est précisément en prenant en compte des interactions entre particules que l'on peut espérer obtenir une certaine régularité, et obtenir des équations aux dérivées partielles (eulériennes, donc) sur lesquelles on pourra espérer dire des choses.*

C'est la version mathématique du problème de l'expérimentateur qui cherche à estimer des vitesses à partir d'observations en termes de position. Cette vision joue un rôle très important dans le cadre du transport optimal, nous proposons ci-dessous une définition de solutions adaptée à ce type de situation.

Solutions faibles de l'équation de transport. Il est important de pouvoir définir des solutions de cette équation pour des densités et des champs moins réguliers.

Definition 1.9. *Soit $t \mapsto \rho_t$ une famille de mesures bornées, et u_t un champ de vecteurs ρ_t-mesurables tel que*

$$\int_0^T \int_{\mathbb{R}^d} |u_t| \, d\rho_t \, dt < +\infty.$$

On dit que le couple (ρ_t, u_t) est solution faible sur $]0,T[$ de l'équation de transport si

$$\int_0^T \int_{\mathbb{R}^d} (\partial_t \varphi + u_t \cdot \nabla \varphi) \, d\rho_t \, dt = 0$$

pour tout $\varphi \in C_c^\infty(\mathbb{R}^d \times]0,T[)$.

Exemple 1.1. L'équation ci-dessus exprime de façon Eulerienne et macroscopique le transport de particules. Considérons une particule de masse m dont la trajectoire est $t \mapsto x(t)$, de vitesse $u(t) = \dot{x}(t)$. On peut représenter ce mouvement de façon Eulérienne en considérans la mesure $\rho_t = m\delta_{x(t)}$, et le "champ" de vitesse $u_t = u(t)$ (cette vitesse n'est définie qu'en $x(t)$, elle n'a pas de sens ailleurs puisque la mesure est supportée en ce point). On a

$$\int_0^T \int_{\mathbb{R}^d} (\partial_t \varphi + u_t \cdot \nabla \varphi) \, d\rho_t \, dt = \int_0^T (\partial_t \varphi(x(t),t) + u(t) \cdot \nabla \varphi(x(t),t)) \, dt$$

$$= \int_0^T \frac{d}{dt} \varphi(x(t),t) \, dt = \varphi(x(T),T) - \varphi(x(0),0) = 0.$$

Remarque 1.10. *On prendra garde au fait suivant : la formulation faible suggère qu'il suffit de se donner un champ de vitesse* presque partout *pour que la notion de solution soit définie*

sans ambiguïté. Mais cette impression n'est justifiée que pour des mesures qui sont absolument continues par rapport à la mesure de Lebesgue, car l'intégrale impliquée dans la formulation faible demande que u_t soit définie ρ_t- presque partout. Prenons par exemple le champ u_t sur $\mathbb{R}$ identiquement égal à un, sauf en 0 où le champ prend la valeur 0. Cette dernière précision peut sembler incongrue car $\{0\}$ est de mesure nulle (relativement à la mesure de Lebesgue), mais la difficulté est que rien dans l'équation n'interdit l'apparition de mesures singulières, qui chargeraient le point 0 en question. On pourra ainsi vérifier que, pour la condition initiale $\rho_0 = \mathbb{1}_{]-1,0[}$, l'équation admet une infinité de solutions, parmi lesquelles on retrouve bien le transport à vitesse constante de la densité initiale

$$\rho_t = \mathbb{1}_{]-1+t,t[},$$

mais aussi

$$\rho_t = \mathbb{1}_{]-1+t,0[} + t\delta_0 \quad \forall t \in [0,1[\,,\ \rho_t = \delta_0 \quad \forall t \geq 1,$$

et, en fait, une infinité de solutions intermédiaires : lors du passage en 0, on peut choisir de laisser passer une fraction arbitraire de masse vers les x positifs, et d'en conserver en 0 le reste (qui va s'accumer pour former une mesure singulière).

Exercice 1.1. Dans l'esprit de la remarque précédente, montrer que la mesure

$$\rho_t = \left| \begin{array}{lll} \delta_0 & \text{sur} &]-\infty, 0[\\ \theta\delta_{-Vt} + (1-\theta)\delta_{Vt} & \text{sur} &]0, +\infty[\end{array} \right.$$

est solution de l'équation de transport pour le champ de vitesse $-V$ sur $]-\infty, 0[$, V sur $]0, +\infty[$, et 0 en 0, avec $V > 0$, quelle que soit la valeur de $\theta \in [0,1]$. Peut-on construire un tel exemple d'indétermination avec le champ de vitesse opposé ? (on pourra se reporter aux notions introduites dans la section 20.5, page 209).

Modèles structurés en âge

L'équation de transport prend une forme particulière lorsque la variable d'espace elle-même correspond en fait à un temps. Ce cadre est naturel lorsque l'on suit une densité de population par tranche d'âge. La forme discrète de cette description correspond à la *pyramide des âges*, utilisée par les démographes. La version continue est basée sur la définition d'une densité $\rho(a,t)$, qui quantifie le nombre de personne à l'âge a. Plus précisément, $\rho(a,t)\,da$ correspond au nombre de personnes entre les âge a et $a + da$.

On obtient typiquement des systèmes de la forme suivante (comme dans la remarque 1.3, la vitesse correspond à un vieilissement d'une unité de temps par unité de temps) :

$$\left| \begin{array}{rcl} \partial_t\rho + \partial_a\rho & = & -\mu(a,t)a, \\ \rho(0,t) & = & \displaystyle\int_0^{+\infty} \beta(a,t)\rho(a,t)\,da, \end{array} \right.$$

où $\mu(a,t)$ correspond au taux de disparition à l'âge a, et $\beta(a,t)$ un taux de fécondité à l'âge a. La dépendance en temps de ces valeurs permet de prendre en compte des facteurs exogènes, du type épidémie momentanée, ou guerre (augmentation de $\mu(a,t)$), ou par exemple la mise en place d'une politique nataliste (augmentation de $\beta(a,t)$). La seconde équation

donne l'impression que l'on fixe le nombre de personnes d'âge 0. Ce terme doit plutôt être interprété comme un terme de flux : de nouvelles personnes (les nouveaux-nés) rentrent dans le circuit, et la valeur $\rho(0,t)$ doit être lue comme un flux $\rho(0,t) \times 1$ (où 1 est une "vitesse" en secondes par seconde), que l'on exprime comme résultant du processus de reproduction.

Aspects théoriques

Malgré sa simplicité apparente, et la trivialité du phénomène qu'elle formalise, l'équation de transport pose des problèmes théoriques extrêmement délicats dès que le champ de vitesse n'est pas régulier. On pourra se reporter à l'article historique de Di Perna & Lions [4], qui établit le caractère bien posé de l'équation de transport (existence et unicité d'une solution pour une condition initiale donnée) dans le cas d'un champ de vitesse $W^{1,1}$, et de divergence uniformément bornée. .

Voir aussi Ambrosio [5] pour une présentation détaillées des différentes approches.

.

1.3 Diffusion

Modèle 1.11. *(Loi de Fick)*
On dit qu'un phénomène de propagation suit la loi de Fick s'il existe un paramètre positif D tel que

$$J = -D\nabla\rho.$$

Remarque 1.12. *D'un point de vue qualitatif, cette loi exprime le fait que la substance a tendance à aller des zones à forte densité vers les zones à faible densité. On peut donc s'attendre à ce qu'un tel phénomène tende à uniformiser les densités. On se reportera à la section 9 pour des exemples de phénomènes de nature (au moins partiellement) diffusive, qui conduisent néanmoins à des répartitions non homogènes de matière dans l'espace.*

Équation de la chaleur. On considère une substance qui diffuse dans un milieu selon la loi de Fick (modèle 1.11). L'équation de conservation (1.1) s'écrit ici

$$\frac{\partial\rho}{\partial t} - \nabla \cdot D\nabla\rho = 0,$$

ou, dans le cas où D est uniforme,

$$\frac{\partial\rho}{\partial t} - D\triangle\rho = 0. \tag{1.3}$$

4. R.J. Di Perna & P.L. Lions,
Ordinary differential equations, transport theory and Sobolev spaces, Invent. math. 98, 511-547 (1989),
`http://perso.crans.org/moussa/dipernalions.pdf`

5. L. Ambrosio, transparents d'un cours donné à Benasque en 2005
`http://benasque.org/benasque/2005pde/2005pde-talks/292Cetraro.pdf`

Noyau de la chaleur. On se place sur l'espace $\mathbb{R}^d$ tout entier. Pour tout $y \in \mathbb{R}^d$, la fonction

$$K(x,t) = \frac{1}{(4\pi Dt)^{d/2}} e^{-\frac{|x-y|^2}{4Dt}}, \tag{1.4}$$

est solution de l'équation de la chaleur (1.3), de telle sorte que, pour toute fonction u_0 suffisamment régulière,

$$u(x,t) = \frac{1}{(4\pi Dt)^{d/2}} \int_{\mathbb{R}^d} e^{-\frac{|x-y|^2}{4Dt}} u_0(y)\,\mathrm{d}y,$$

est la solution de l'équation de la chaleur pour la donnée initiale $u(x,0) = u_0(x)$.

Diffusion non isotrope. Dans le cas où le milieu n'est pas isotrope (i.e. la diffusion est plus importante dans certaines direction), on peut introduire une matrice de diffusion définie positive $\mathbf{D}$ qui conduit a une équation formellement analogue. Ce phénomène traduit la non-isotropie du milieu considéré : lorsque la diffusion se fait plus aisément dans certaines directions, la matrice $\mathbf{D}$ ne sera pas scalaire. Cette situation est courante dans le cas de milieux *fibreux*, comme le sont par exemple les muscles dans le corps humain.

Conditions aux limites. On suppose que le phénomène de diffusion prend place dans une zone délimitée de l'espace. On note Ω cette zone, et l'on suppose que Ω est un ouvert borné. Il est alors licite de prescrire deux types de condition sur la frontière de Ω.

(i) Conditions de Dirichlet : la valeur de la densité est imposée au bord du domaine.

(ii) Conditions de Neumann : on prescrit le flux $J \cdot n$ à travers la frontière du domaine Ω, c'est-à-dire, sous l'hypothèse de flux régi par la loi de Fick, la dérivée normale de la densité, ou plus précisément $-D\partial\rho/\partial n$.

Il est possible de panacher ces deux conditions, c'est-à-dire d'imposer la valeur de ρ sur une partie de la frontière, et la valeur de la dérivée normale sur son complémentaire.

Notons qu'un troisième type de conditions aux limites peut être envisagé, qui implique à la fois la valeur de la fonction et sa dérivée normale, il s'agit des

(iii) Conditions de Robin (ou Fourier) : on prescrit une combinaison linéaire (à coefficient positifs) de la valeur et de la dérivée normale.

Précisons d'où peuvent venir ces dernières conditions en prenant l'exemple de la diffusion de l'oxygène dans le sang au travers de la paroi alvéolaire. On assimile un alvéole à une sphère remplie d'air, au sein duquel l'oxygène diffuse selon la loi de Fick avec un certain paramètre de diffusivité D. La paroi alvéolaire sépare l'alvéole des capillaires dans lesquels circulent le sang, dont les globules rouges vont capter l'oxygène. Au sein de cette paroi, l'oxygène diffuse également et comme elle est très fine, il est licite de négliger au premier ordre la diffusion dans la direction transverse. Si l'on note u_{ext} la concentration en oxygène dans le sang, on peut écrire que le flux d'oxygène au travers de la paroi est proportionnel à la différence de valeurs de part et d'autre, ce qui conduit à écrire

$$\text{Flux alvéole vers sang} = \beta(u - u_{\text{ext}}),$$

où u est la valeur de la concentration dans l'alvéole au voisinage de la paroi alvéolaire, d'où la condition en tout point de la frontière

$$-D\frac{\partial u}{\partial n} = \beta(u - u_{\text{ext}})\,, \text{ i.e. } \beta u + D\frac{\partial u}{\partial n} = \beta u_{\text{ext}}.$$

Noter que cette condition présente l'avantage de contenir d'une certaine manière toutes les autres, puisque l'on retrouve des conditions de Neumann en faisant tendre β vers 0, et des conditions de Dirichlet [6] en faisant tendre β vers $+\infty$.

1.4 Transport - diffusion

Lorsque les deux phénomènes évoqués précédemment coexistent, on parle de transport-diffusion, ou convection-diffusion.

On peut décomposer le vecteur flux en ses deux composantes

$$J = J_u + J_D = u\rho - D\nabla\rho,$$

ce qui conduit à 'équation

$$\frac{\partial \rho}{\partial t} + \nabla\cdot(u\rho) - \nabla\cdot D\nabla\rho = 0.$$

Remarque 1.13. *Dans le cas où le champ de vitesse dérive d'un potentiel, on obtient une équation d'un type particulier (équation dite de* Fokker-Planck*), qui peut s'interpréter comme exprimant un processus de pure diffusion (voir section 9, ou section 1.5 ci-après).*

Definition 1.14. *(Nombre de Péclet)*
Le nombre de Péclet est défini par

$$Pe = \frac{UL}{D},$$

où L représente la taille caractérisque du domaine considéré, U l'ordre de grandeur du module de u, et D le coefficient de diffusion.

Lorsque le nombre de Péclet est petit devant 1, cela signifie que les phénomènes de diffusion sont prépondérants devant les phénomènes de convection. Concrètement, cela signifie que le terme de convection dans l'équation peut être supprimé sans que le champ solution soit modifié de façon significative. Pour Pe $>>$ 1, c'est au contraire la convection qui domine. Dans cette dernière situation, on prendra garde au fait que la suppression du terme de diffusion change profondément la nature de l'équation. Plus précisément, si l'on considère l'équation de convection-diffusion avec des conditions de Dirichlet (valeur de ρ imposée au bord), on peut voir apparaître lorsque a tend vers 0 le phénomène dit de *couche limite*. Dans le cas limite $D = 0$, sur une partie de la frontière où la vitesse est sortante, l'équation ne "voit" pas la condition limite, puisque qu'il n'est pas licite de prescrire la valeur de ρ en un tel point. On aura en général pour des nombres de Péclet grands apparition de très forts gradients de ρ au voisinage de ces zones.

6. Cette technique est couramment utilisée numériquement pour imposer, dans le cadre des méthodes d'éléments finis, des conditions de Dirichlet sans changer la structure de la matrice : il s'agit de la méthode de pénalisation frontière.

Adimensionnement des équations de transport diffusion.
Le nombre de Péclet peut être introduit de la façon suivante : on considère une substance qui se propage par advection et diffusion (champ u et paramètre a), dans un domaine de taille caractéristique L. On note note U l'ordre de grandeur du champ advectant, et $T = L/U$ un temps caractéristique (temps mis par une particule pour être déplacée par advection d'une longueur caractéristique). Écrire l'équation en variables adimensionnées consiste à introduire les variables de temps et d'espaces (sans dimension) $t^\star = t/T$ et $x^\star = x/L$. On note par ailleurs $u^\star = u/U$. Dans ces nouvelles variables, l'équation s'écrit

$$\frac{\partial \rho}{\partial t^\star} + \nabla^\star \cdot (u^\star \rho) - \frac{1}{\mathrm{Pe}} \Delta^\star \rho = 0,$$

Exemple 1.2. (Couche limite)
On considère l'équation de convection-diffusion stationnaire (la dérivée partielle par rapport au temps est égale à 0) sur l'intervalle $]0, L[$, avec une vitesse constante égale à 1, et des conditions aux limites $\rho(0,t) = 1$, $\rho(L,t) = 0$:

$$\partial_x \rho - a \partial_{xx} \rho = 0.$$

La fonction ρ ne dépendant plus du temps, on note ρ' et ρ'' les dérivées en x. On déduit de l'équation de convection diffusion stationnaire que $\ln |\rho'|$ est affine de pente $1/a$, d'où, après prise en compte des conditions aux limites,

$$\rho(x) = \frac{1 - e^{\frac{x-L}{a}}}{1 - e^{-\frac{L}{a}}}.$$

On vérifie que cette fonction, qui prend la valeur 0 en $x = L$, tend uniformément vers 1 sur tout intervalle du type $[0, L - \eta]$, avec $\eta > 0$.

1.5 Advection diffusive

L'équation d'advection diffusion peut aussi s'obtenir dans certains contextes à partir d'un processus de diffusion pure.

Considérons par exmple un processus de diffusion dans $\mathbb{R}^d$ sous une hypothèse de symétrie radiale : $\tilde{\rho}(x,t) = \tilde{\rho}(r,t)$. L'équation sur $\tilde{\rho}$ s'écrit

$$\partial_t \tilde{\rho} - D \Delta \tilde{\rho} = 0\,, \text{ soit } \partial_t \tilde{\rho} - \frac{D}{r^{d-1}} \partial_r \left(r^{d-1} \partial_r \tilde{\rho} \right) = 0.$$

Plus généralement, on peut considérer un processus de diffusion radial dans un espace où la mesure de la sphère de rayon r est donné. Notant $\omega(r)$ cette mesure, on obtient la forme générale

$$\partial_t \tilde{\rho} - \frac{D}{\omega(r)} \partial_r \left(\omega(r) \partial_r \tilde{\rho} \right) = 0.$$

Notons [7] maintenant ρ la densité linéique de masse à distance r de l'origine, $\rho = \omega(r)\tilde{\rho}$ (de telle sorte que $\rho(r)\,dr$ est la quantité totale de matière contenue entre les sphères de rayons r et $r + dr$). Cette nouvelle quantité vérifie l'équation

$$\partial_t \rho - D \partial_{rr} \rho + D \partial_r \left(\frac{\omega'}{\omega} \rho \right) = 0.$$

7. Approche suggérée par B. Merlet.

On obtient ainsi une équation d'advection diffusion avec une vitesse centrifuge $V(r)$ égale à $D\omega'/\omega$. Cette advection centrifuge de nature purement diffusive est liée au fait que, dans cette formulation sur la variable ρ de densité par unité de distance à l'origine, la diffusion est décalée vers les r croissants, puisqu'il y a plus de place lorsque l'on s'éloigne de l'origine (le volume de la couronne entre r et $r + dr$ est plus grand que celui entre $r - dr$ et r. Dans le cas de $\mathbb{R}^d$, pour $d \geq 2$, on a $V = D(d-1)/r$.

2 Fluides

2.1 Tenseur des contraintes, équations générales du mouvement d'un fluide

Definition 2.1. *(Tenseur des contraintes)*
On considère ici un fluide occupant un certain domaine de l'espace, x un point de ce domaine, n un vecteur unité, et $D_\varepsilon(n)$ un disque (ou un segment en dimension 2 d'espace, voire un point[8] *en dimension 1) centré en x, d'aire ε (longueur ε en dimension 2), orthogonal à n.*
On note $F_\varepsilon(n)$ la force exercée sur $D_\varepsilon(n)$ par le fluide situé du côté de n. Si $F_\varepsilon(n)/\varepsilon$ tend vers $F(n)$ quand ε tend vers 0, et si la correspondance $n \longmapsto F(n)$ est linéaire, on appelle tenseur[9] *des contraintes en x le tenseur σ qui représente cette correspondance linéaire.*

$$F(n) = \sigma \cdot n.$$

Le mouvement d'un fluide qui admet partout un tel tenseur peut être formalisé par une équation très générale. On note $\rho = \rho(x,t)$ la densité locale (masse par unité de volume), par u la vitesse[10], et par f une force en volume agissant sur le fluide (typiquement la gravité $f = \rho\, g$). On considère un système matériel $\omega(t)$, c'est à dire à ensemble de particules que l'on suit dans leur mouvement[11]. Le principe fondamental de la dynamique (ou loi de Newton) exprime que la dérivée en temps de la quantité de mouvement pour ce système est égal à la somme des forces extérieures :

$$\frac{d}{dt}\int_{\omega(t)} \rho u = \text{somme des forces extérieures.} \tag{2.1}$$

Le membre de droite est la somme de la contribution des forces en volume $\int_\omega f$, et le bilan des forces exercées sur ω par le fluide à l'extérieur de ω, qui s'écrit, d'après la définition 2.1,

$$\int_{\partial\omega} \sigma \cdot n = \int_\omega \nabla \cdot \sigma.$$

8. Dans ce cas extrême, mais très utile en pratique (la dimension 1, très pauvre pour les fluides incompressibles, permet d'étudier de façon fine les modèles de fluides compressibles), il n'y a évidemment pas lieu de faire tendre la mesure vers 0.

9. On pourra remplacer ici le terme de tenseur par matrice, et considérer que $\sigma \cdot n$, qui représente la contraction de deux tenseurs, correspond à un simple produit matrice vecteur, que l'on verra noté σn dans certains documents.

10. Précisons que le fait de considérer qu'une telle vitesse puisse être définie en tout point est une hypothèse très forte. Par ailleurs, comme dans le cas de la définition du vecteur flux (voir définition 1.1, page 11), parler de vitesse véritablement ponctuelle n'a pas de sens autre qu'abstrait puisque, pour les fluides réels (en particulier pour les gaz) à une échelle inférieure à la taille intermoléculaire, la matière ne peut être vue comme un continuum : la plupart des "points" sont en fait dans le vide, et cela n'a pas de sens de définit une vitesse, dans ce contexte, en l'absence de matière. L'hypothèse sous-jacente est qu'il existe une échelle *mésoscopique* telle que l'on puisse définir à chaque instant une vitesse moyenne sur des volumes élémentaires représentatifs à cette échelle.

11. Si on se donne un sous-domaine $\omega(0)$ comme position initiale du système matériel, on a

$$\omega(t) = \{X_t(x)\,,\ x \in \omega(0)\},$$

où $t \mapsto X_t(x)$ est la trajectoire de la particule située en x à $t = 0$, i.e.

$$\frac{\partial X_t}{\partial t}(x) = u(X_t(x), t)\,,\ X_0(x) = x.$$

Le membre de gauche de 2.1 s'écrit donc (voir équation (A.12), page 254)

$$\frac{d}{dt}\int_{\omega(t)} \rho u = \int_{\omega(t)} \frac{\partial(\rho u)}{\partial t} + \int_{\partial\omega(t)} \rho u (u\cdot n),$$

et le dernier terme peut s'écrire comme une intégrale en volume

$$\int_{\partial\omega(t)} \rho u (u\cdot n) = \int_{\omega(t)} \nabla\cdot\left(\rho u\otimes u\right),$$

où $u\otimes u$ représente la matrice symétrique $(u_i\, u_j)_{i,j}$. Comme le système matériel est arbitraire (en particulier aussi petit qu'on veut), on en déduit l'équation générique suivante :

Modèle 2.2. *(Équation d'évolution générale pour un fluide inertiel)*
On considère un fluide en mouvement de densité $\rho(x,t)$, de vitesse $u(x,t)$, soumis à une force en volume f. On suppose l'existence, en tout point (x,t) du domaine de l'espace-temps occupé par le fluide, d'un tenseur des contraintes $\sigma(x,t)$. La conservation locale de la quantité de mouvement s'écrit

$$\frac{\partial}{\partial t}\left(\rho u\right) + \nabla\cdot\left(\rho u\otimes u\right) - \nabla\cdot\sigma = f. \tag{2.2}$$

La conservation de la masse s'écrit par ailleurs

$$\frac{\partial \rho}{\partial t} + \nabla\cdot\left(\rho u\right) = 0.$$

Modèle 2.3. *(Équilibre des forces pour un fluide non inertiel)*
Quand l'inertie est négligeable, la loi de Newton est remplacée par une relation d'équilibre instantané des forces, qui s'écrit

$$-\nabla\cdot\sigma = f.$$

Remarque 2.4. *On peut légitimement se demander s'il est acceptable d'écrire des dérivées en espace et en temps de quantités scalaires ou vectorielles dont on n'a pas précisé les régularités. Le notion de solution faible de telles équation permet de donner un sens à ce qui précède, même dans le cas de champs peu régulier. Montrons en particulier que l'équation générale écrite ci-dessus (nous ne garderons ici que la partie inertielle) peut être interprétée comme généralisant la loi fondamentale de la dynamique pour des points matériels, si on lui donne un sens pour des distributions de matière ρ singulières. On se place en dimension 1 pour simplifier, on considère $t\mapsto\rho_t$ une courbe de mesures positives de même masse (par exemple des mesures de probabilité), on note u_t le champ de vitesse au temps t, donné comme fonction ρ_t-mesurable, et g un champ de force par unité de masse. On dira que (ρ_t, u_t) est solution faible de*

$$\partial_t(\rho_t u_t) + \partial_x(\rho_t u_t^2) = \rho_t g$$

sur $]0,T[$ si

$$-\int_0^T\int_{\mathbb{R}} \partial_t\varphi u_t\, d\rho_t - \int_0^T\int_{\mathbb{R}} \partial_x\varphi\,(u_t)^2 d\rho_t = \int_0^T\int_{\mathbb{R}} g d\rho_t,$$

pour toute fonction φ régulière à support compact sur $]0,T[\times\mathbb{R}$. Prenons maintenant le cas d'une particule de masse m, soumise à l'action d'une force mg, et dont la trajectoire est $x(t)$. L'expression du principe fondamental de la dynamique pour cette particule est $m\ddot{x} = f$. On représente cette particule de façon Eulerienne par une mesure $\rho_t = m\delta_{x(t)}$, et l'on note $u(t)$ sa vitesse. La masse étant concentrée, il est en effet naturel de voir le "champ" de vitesse

(qui est une fonction ρ_t-mesurable) comme un simple scalaire fonction du temps. Écrivons la formulation faible ci-dessus appliquée à ρ_t, $u(t)$. On obtient

$$-\int_0^T m\,\partial_t\varphi(x(t),t)\,u_t - \int_0^T \partial_x\varphi(x(t),t)\,u(t)^2$$

$$= -\int_0^T m\,u(t)\left(\underbrace{\partial_t\varphi(x(t),t)\,u_t + \partial_x\varphi(x(t),t)\,u_t}_{d\varphi(x(t),t)/dt}\right) = \int_0^T mg\varphi(x(t),t).$$

En intégrant par parties l'intégrale contenant le $d\varphi(x(t),t)/dt$, on obtient

$$\int_0^T \left(\frac{d(mu(t))}{dt} - mg\right)\varphi(x(t),t)\,dt,$$

valable pour toute fonction test, d'où $m\ddot{x} = mg$. On généralise immédiatement cette démarche au cas de plusieurs particules sans croisement de trajectoire. On peut aller au-delà en vérifiant par exemple que la collision de deux particules peut-être représentée de façon Eulérienne par une solution faible de l'équation (dite d'Euler sans pression) ci-dessus. En prenant par exemple un forçage extérieur nul, et

$$\rho_t = \frac{1}{2}\delta_{x_1(t)} + \frac{1}{2}\delta_{x_2(t)}\,,\ x_1(t) = (-1+t)_-\,,\ x_2(t) = (1-t)_+,$$

avec le champ de vitesse correspondant (vitesses opposées jusqu'au temps 1*, nulle ensuite). Mais l'équation elle même ne fait qu'exprimer la quantité de mouvement, sans considération énérgétique. On peut en particulier vérifier que toute loi de collision qui préserve la quantité de mouvement (les particules repartent avec des vitesses opposées) est solution de l'équation ci-dessus.*

L'essentiel de la démarche de modélisation des milieux continus fluides consiste à exprimer le tenseur des contraintes. On distingue deux grandes classes de fluides, les fluides dits *parfaits*, pour lesquels le tenseur des contraintes est diagonal, et les autres fluides, dits *réels*, qui présentent une tendance à résister aux déformations. On s'intéressera en particulier ici aux fluides réels newtoniens incompressibles.

2.2 Fluides parfaits

Un fluide parfait est caractérisé par le fait que, si l'on reprend la définition du tenseur des contraintes, la force exercée sur le disque infinitésimal $D_\varepsilon(n)$ est dirigée suivant n, et son intensité ne dépend pas de l'orientation.

Definition 2.5. *(Fluide parfait)*
Un fluide est dit parfait s'il admet un tenseur des contraintes diagonal, i.e. il existe un champ scalaire p*, appelé champ de* pression *tel que*

$$\sigma(x) = -\mathrm{p}\,\mathrm{Id},$$

où Id *est le tenseur identité.*

Pour un tel fluide, on a

$$-\nabla \cdot \sigma = \nabla \cdot (\mathrm{p}\,\mathrm{Id}) = \nabla \mathrm{p},$$

ce qui conduit à l'équation d'Euler

$$\frac{\partial}{\partial t}(\rho u) + \nabla \cdot (\rho u \otimes u) + \nabla \mathrm{p} = f.$$

Fluide parfait incompressible

Dans le cas d'un fluide homogène (ρ est uniforme) et incompressible (le champ de vitesse est à divergence nulle), on a

$$\nabla \cdot (\rho u \otimes u) = \rho\,(u \cdot \nabla)\,u,$$

où $(u \cdot \nabla)\,u$ est tel que

$$((u \cdot \nabla)\,u)_i = \sum_{j=1}^{d} u_j \frac{\partial u_i}{\partial x_j}.$$

Modèle 2.6. *(Équation d'Euler incompressible)*
On considère un fluide en mouvement de densité $\rho(x,t)$, de vitesse $u(x,t)$, soumis à une force en volume f. On suppose le fluide parfait (on note p *la presison), homogène, et incompressible. Le triplet (ρ, u, p) vérifie alors les Équation d'Euler incompressibles*

$$\left| \begin{array}{rcl} \rho \dfrac{\partial u}{\partial t} + \rho\,(u \cdot \nabla)\,u + \nabla \mathrm{p} & = & f \\ \nabla \cdot u & = & 0 \end{array} \right. \tag{2.3}$$

L'apparente simplicité de cette équation, obtenue en faisant des hypothèses très fortes sur le fluide, est trompeuse. Un fait particulièrement troublant la concernant est lié au *paradoxe de Scheffer-Schnirelman* [12] : on peut construire une solution du système ci-dessus, sans forçage ($f = 0$), non nulle, à support compact en espace temps.

Dans le cas d'un écoulement incompressible stationnaire, on peut montrer formellement la conservation d'une certaine quantité (appelée pression dynamique) le long des lignes de courant.

Proposition 2.7. *("Théorème" de Bernoulli)*
On considère l'écoulement stationnaire d'un fluide parfait homogène incompressible, soumis à l'action d'une force qui dérive d'un potentiel $f = -\nabla\Phi$. On suppose les champs de vitesse et de pression réguliers (continûment différentiables). La quantité

$$\frac{\rho}{2}|u|^2 + \mathrm{p} + \Phi$$

se conserve le long des lignes de courant.

12. On pourra se reporter à la description de cette construction dans :
C. Villani, Paradoxe de Scheffer-Schnirelman revu sous l'angle de l'intégration convexe [d'après C. De Lellis et L. Székelyhidi], Séminaire Bourbaki, Novembre 2008, 61ème année, 2008-2009, no 1001.
`http://cedricvillani.org/wp-content/uploads/2012/08/B10.Bourbaki2.pdf`

Démonstration. On a

$$((u\cdot\nabla)\,u)\cdot u=\sum_{i=1}^{d}u_i\sum_{j=1}^{d}u_j\partial_j u_i=\frac{1}{2}\sum_j u_j\partial_j\left(\sum_i|u_i|^2\right)=u\cdot\nabla\left(\frac{|u|^2}{2}\right).$$

On a donc, en prenant le produire scalaire avec u de la première ligne de (2.3), sans le terme de dérivée en temps (supposé nul),

$$u\cdot\nabla\left(\frac{\rho}{2}|u|^2+\mathrm{p}+\Phi\right)=0,$$

d'où la propriété annoncée. □

Fluide parfait barotrope

Une autre manière de fermer [13] les équations d'Euler est de supposer un lien univoque entre la densité et la pression. On obtient alors le

Modèle 2.8. *(Équations d'Euler barotropes)*
On considère un fluide en mouvement de densité $\rho(x,t)$, de vitesse $u(x,t)$, soumis à une force en volume f. On suppose le fluide parfait (on note p *la pression). Le système d'Euler barotrope s'écrit comme suit*

$$\left|\begin{array}{rcl}\dfrac{\partial\rho}{\partial t}+\nabla\cdot(\rho u) &=& 0,\\ \dfrac{\partial}{\partial t}(\rho u)+\nabla\cdot(\rho u\otimes u)+\nabla\mathrm{p} &=& f\\ \mathrm{p} &=& \mathrm{p}(\rho).\end{array}\right. \tag{2.4}$$

Équations de l'acoustique

Le modèle précedent permet d'obtenir formellement l'équation des ondes, ce qui permet de modéliser la propagation du son dans un fluide compressible.

On se propose ici de montrer formellement comment l'on peut passer des équations d'Euler pour un gaz compressible à l'équation des ondes qui va modéliser la propagation d'ondes au sein de ce milieu. Le point de départ est donc le système d'Euler

$$\partial_t\rho+\nabla\cdot(\rho u)=0,$$

$$\frac{\partial}{\partial t}(\rho u)+\nabla\cdot(\rho u\otimes u)+\nabla\mathrm{p}=0,$$

13. Il peut être très délicat de montrer rigoureusement existence et unicité d'une solution aux équations obtenues, mais cette approche permet d'avoir autant d'équations ($d+2$) que d'inconnues (d pour la vitesse, 1 pour la densité, 1 pour la pression), de telle sorte que le modèle obtenu puisse être considéré comme un *problème*, c'est à dire un système d'équations pour lequel on peut espérer obtenir, sous certaines hypothèses, des résultats théoriques. On peut qualifier ce problème de *posé*, en attente d'être *bien posé* (expression que l'on réserve aux problèmes pour lesquels on a au moins un résultat d'existence et d'unicité, conditionné à d'éventuelles conditions sur l'état initial et le forçage).

avec $\mathrm{p} = \mathrm{p}(\rho)$. On considère que les différentes variables restent au voisinage de valeurs de références ρ_0, p_0, et $u_0 = 0$ pour la vitesse, et l'on garde les notations ρ, p et u pour désigner les (petites) variations au voisinage de ces valeurs. On suppose en outre (on peut montrer que cette hypothèses est réaliste dans un grand nombre de situations) le régime barotrope, c'est à dire que la pression est supposée ne dépendre que de la densité : $p = p(\rho)$. On notera $\beta = p'(\rho_0)$. On réécrit les équations ci-dessus en ne conservant que les termes d'ordre 1 dans les petites variations :

$$\partial_t \rho + \rho_0 \nabla \cdot u = 0,$$

$$\rho_0 \partial_t u + \nabla \mathrm{p} = 0.$$

On a

$$\nabla p = p'(\rho)\nabla\rho \approx p'(\rho_0)\nabla\rho = \beta\nabla\rho,$$

ce qui permet d'éliminer la pression dans la seconde équation. Si l'on prend maintenant la divergence de la seconde équation, la dérivée partielle par rapport au temps de la première , et que l'on fait la différence, on obtient

$$\partial_{tt}\rho - \beta\Delta\rho = 0,$$

avec $\beta = \mathrm{p}'(\rho_0)$, c'est-à-dire une équation des ondes sur la (petite variation de la) densité. On aura donc propagation d'ondes au sein du fluide, à la célérité c, avec $c^2 = \beta$. Dans le cas d'un gaz comme l'air, supposé parfait, de coefficient isentropique $\gamma = 1.4$, on a

$$\frac{p}{p_0} = \left(\frac{\rho}{\rho_0}\right)^\gamma \text{ et donc } \beta = p'(\rho_0) = \gamma\frac{p_0}{\rho_0}.$$

On obtient dans des conditions normales ($p_0 = 10^5$ Pa, $\rho_0 = 1.2$ kg m^{-3}),

$$c = \sqrt{\frac{\gamma p_0}{\rho_0}} \approx 341\, ms^{-1}.$$

2.3 Fluides newtoniens

Les fluides dits *réels* présentent une certaine résistance à la déformation. Pour quantifier cette déformation, on considère une particule de fluide évoluant au voisinage d'une trajectoire $t \mapsto x(t)$. La vitesse au voisinage de x s'écrit

$$u(y,t) \approx u(x,t) + \nabla u(x,t)\cdot(y-x)$$

$$= \underbrace{u(x,t)}_{\text{Translation}} + \left(\underbrace{\frac{\nabla u - {}^t\nabla u}{2}}_{\text{Rotation}} + \underbrace{\frac{\nabla u + {}^t\nabla u}{2}}_{\text{Déformation}}\right)\cdot(y-x).$$

Le mouvement d'un segment matériel $\overline{xy}$ peut ainsi être décomposé en 3 contributions : un mouvement de translation à la vitesse locale, un mouvement de rotation (partie antisymétrique du gradient du champ de vitesse), et une dernière contribution qui correspond aux déformations locales (partie symétrique du gradient du champ de vitesse) On se reportera à la figure 2.1 pour une illustration (en dimension 2 d'espace) de ces trois contributions.

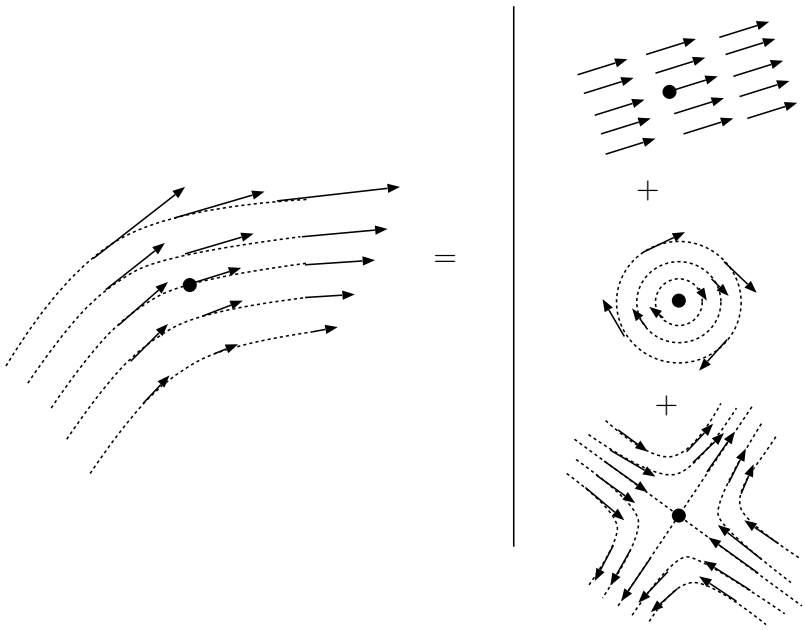


FIGURE 2.1 – Décomposition locale d'un champ de vitesse

Definition 2.9. *(Tenseur des taux de déformation)*
On considère un fluide évoluant selon le champ de vitesse u. Le tenseur des taux de déformations est défini par

$$D = \frac{\nabla u + {}^t\nabla u}{2}.$$

Le modèle le plus simple de fluide réel (nous nous limiterons ici au cas incompressible) est obtenu en considérant que le tenseur des contraintes est, à la contribution diagonale associée à la pression près, proportionnel au tenseur des taux de déformation :

Definition 2.10. *(Fluide (incompressible) newtonien)*
Un fluide incompressible est dit newtonien s'il existe un paramètre positif μ, appelé viscosité, *tel que le tenseur des contraintes s'écrive*

$$\sigma = 2\mu D - \mathrm{p}\,\mathrm{Id} = \mu\left(\nabla u + {}^t\nabla u\right) - \mathrm{p}\,\mathrm{Id},$$

où $\mathrm{p} = \mathrm{p}(x,t)$ est un champ scalaire (pression).

On considère maintenant un fluide incompressible newtonien et homogène (ρ est uniforme). Comme ρ est constant, il peut être sorti de la dérivée en temps. Par ailleurs, comme

$$\nabla \cdot u = \sum_{i=1}^{d} \frac{\partial u_i}{\partial x_i} = 0,$$

on a

$$\nabla \cdot (u \otimes u) = \nabla \cdot (u_i\, u_j)_{i,j} = \left(\sum_{i=1}^{d} u_i \frac{\partial u_j}{\partial x_i}\right)_{1\le j\le d}.$$

Cette quantité exprime la dérivée de la vitesse dans sa propre direction, on la note $(u \cdot \nabla)\, u$ (on peut comprendre cette notation en considérant le bloc $u \cdot \nabla$ comme un opérateur différentiel scalaire $u_1\partial_1 + \cdots + u_d\partial_d$ qui s'applique composante par composante au vecteur u lui même).

Modèle 2.11. *(Équations de Navier-Stokes incompressible)*
L'écoulement d'un fluide newtonien, incompressible et homogène, soumis à l'action d'une force en volume f, suit les équations de Navier-Stokes

$$\left\{ \begin{array}{rcl} \rho\left(\dfrac{\partial u}{\partial t} + (u \cdot \nabla)\, u\right) - \mu\Delta u + \nabla \mathrm{p} & = & f \\ \nabla \cdot u & = & 0. \end{array} \right.$$

Forme adimensionnelle des équations de Navier-stokes

Soit U l'ordre de grandeur de la vitesse pour l'écoulement considéré, L la dimension caractéristique du phénomène étudié, et $T = L/U$ le temps caractéristique associé. On introduit les variables adimensionnées

$$u^\star = \frac{u}{U}\,,\ x^\star = \frac{x}{L}\,,\ t^\star = \frac{t}{T}.$$

En notant $\nabla^\star$ (resp. $\Delta^\star$) le gradient (resp. le Laplacien) relativement à la variable d'espace adimensionnée, on obtient

$$\frac{\partial u^\star}{\partial t^\star} + (u^\star \cdot \nabla^\star)\, u^\star - \frac{\mu}{\rho U L}\Delta^\star u^\star + \nabla^\star \mathrm{p}^\star = f^\star,$$

où $\mathrm{p}^\star = \mathrm{p}/(\rho U^2)$ est la pression adimensionnée, et $f^\star = fL/(\rho U^2)$ le terme de forçage adimensionné.

Definition 2.12. *Le nombre $Re = \rho UL/\mu$ est appelé nombre de Reynolds. Il quantifie l'importance relative des effets inertiels par rapport aux effets visqueux.*

Quand ce nombre (sans dimension) est petit devant 1, on peut considérer que les effets inertiels sont négligeables, de telle sorte que la loi de Newton est remplacée par un équilibre des forces instantané

Modèle 2.13. *(Équations de Stokes incompressibles)*
Un fluide newtonien et incompressible, soumis à une force en volume f, dans un régime d'écoulement où les effets visqueux peuvent être négligés , suit les équations de Stokes incompressibles

$$\left\{ \begin{array}{rcl} -\mu\Delta u + \nabla \mathrm{p} & = & f \\ \nabla \cdot u & = & 0 \end{array} \right. \tag{2.5}$$

Remarque 2.14. *L'absence de dérivée en temps dans ce système s'explique simplement par la disparition des termes d'inertie, mais on évitera de parler d'équation statique, elle exprime plutôt un équilibre instantané des forces à chaque instant, en tout point du fluide. Ce fluide est bien en mouvement, et dans le cas d'un fluide à surface libre, le domaine lui même sera déformé par ce mouvement, malgré l'absence de dérivée en temps.*

Si l'on considère la situation où le fluide remplit un domain délimité par des murs physiques imperméables, on considère en général[14] que le fluide accroche à la paroi, ce qui s'exprime sous la forme de *conditions de Dirichlet homogènes* $u = 0$ sur la frontière $\partial\Omega$.

Écoulements en milieu poreux

Les écoulements en milieu poreux tiennent une place un peu particulière dans les modèles fluides, du fait qu'il mettent en jeu deux phases : l'une est constituée par un fluide visqueux incompressible, et l'autre est une *matrice*[15] rigide et fixe (typiquement un amas tridimensionnels de grains rigides), au travers de laquelle le fluide est susceptible de s'écouler. Même si le fluide est peu visqueux, le fait que l'écoulement du fluide se fasse à une échelle très petite (au travers des *pores* du milieu) permet dans un grand nombre de situations de négliger les effets inertiels : le nombre de Reynolds local est très petit (voir définition 2.12). On a alors une relation de proportionalité entre flux de fluide et gradient de pression. Plus précisément, Darcy a mis en évidence (voir figure 2.2) que le flux d'eau s'écoulant au travers d'un milieu poreux (grains de sable) dépendait linéairement de la différence de pression entre l'entrée et la sortie du domaine. L'écriture locale de cette relation conduit à

Modèle 2.15. *(Loi de Darcy en milieu isotrope)*
On considère l'écoulement d'un fluide visqueux dans un milieu poreux saturé[16].

On dit que cet écoulement suit la Loi de Darcy s'il existe k, appelé perméabilité du milieu, tel que

$$u = -k\nabla \mathrm{p},$$

où μ est la viscosité du fluide, p *la pression au sein du fluide, et u est la vitesse moyenne locale.*

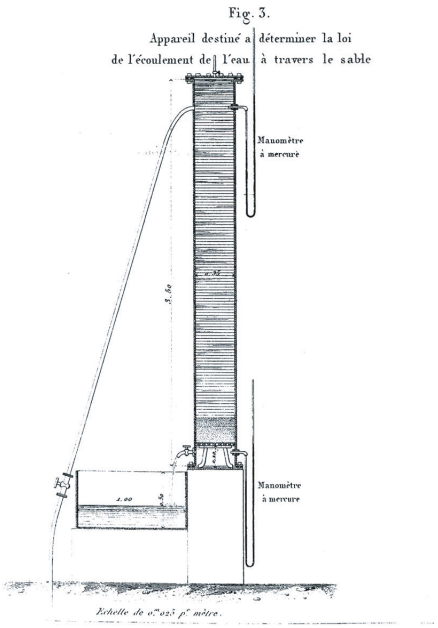
Remarque 2.16. *La notion de vitesse moyenne évoquée ci-dessus correspond en fait à un flux (volumique) par unité de surface. Cette quantité, en* $\mathrm{m}^3\ \mathrm{s}^{-1}$ *par* m^2*, est effectivement homogène à une vitesse, mais on prendra garde au fait que son module peut être très différent de la vitesse effective des particules fluides en mouvement. En particulier, dans le cas d'une porosité (fraction de vide au sein du milieu) très faible, les vitesses effectives des particules seront très supérieures à cette vitesse, appelée vitesse de Darcy.*

14. Cette hypothèse peut être invalidée dans certaines circonstances. Il est parfois plus pertinent d'utiliser les conditions dites de *Navier*, qui préservent la condition de non pénértration du fluide dans la paroi, mais autorisent une vitesse tangentielle non nulle.

15. Au sens bien sûr bassement matériel du terme : il s'agit de décrire une phase solide et immobile quels que soient les efforts exercés sur elle par le fluide.

16. On dit que le mieu est saturé si l'espace libre est entièrement occupé par le fluide visqueux. La proportion d'espace libre est appelée porosité, notée Φ en général. Une valeur typique de Φ est 0.64, qui correspond au *Maximal Random Packing* pour des sphères de même taille (cas *monodisperse*), distribuée "aléatoirement". Le sens de *aléatoirement* ci-dessus est loin d'être trivial, on pourra pour plus de détails se reporter à :
S. Torquato, T. M. Truskett, P. G. Debenedetti, Is Random Close Packing of Spheres Well Defined ?, PRL Vol. 84, No 10, `http://cherrypit.princeton.edu/papers/paper-176.pdf`

16. L'étude des milieux non saturé n'est pas abordée ici. Précisons simplement que l'abandon de l'hypothèse de saturation conduit à des problèmes extrêmement complexes du fait que, l'écoulement fluide au niveau des pores se faisant à petite échelle, les effets de tension surfacique (conditionnés par la nature du fluide, des surfaces solides, et potentiellement du gaz environnant) ne sont en général pas négligeables.



La chambre supérieure de la colonne reçoit l'eau par un tuyau embranché sur la conduite de l'hôpital, et dont un robinet permet de modérer à volonté le débit; la chambre inférieure s'ouvre par un robinet sur un bassin de jaugeage de 1 mètre de côté.

La pression aux deux extrémités de la colonne est indiquée par des manomètres à mercure en U; enfin chacune des chambres est munie d'un robinet à air, essentiel pour la mise en charge de l'appareil.

Les expériences ont été faites avec du sable siliceux de Saône, composé ainsi qu'il suit :

0m 58	de sable passant au crible de	0 mil.	77
0m 13	—	1	10
0m 12	—	2	00
0m 17	de menu gravier, débris de coquilles, etc.		

Il présente environ $\frac{38}{100}$ de vide.

Le sable était versé et tassé dans la colonne préalablement remplie d'eau, afin que les vides de la masse filtrante ne continssent plus d'air, et la hauteur du sable n'était mesurée qu'à la fin de chaque série d'expériences, après que le passage de l'eau l'avait convenablement tassé.

Chaque expérience consistait à établir dans la chambre supérieure de la colonne, par la manœuvre du robinet d'amenée, une pression déterminée ; puis, lorsque par deux observations l'on s'était assuré que l'écoulement était devenu sensiblement uniforme, on notait le débit du filtre pendant un certain temps et on en concluait le débit moyen par minute.

FIGURE 2.2 – Description de l'expérience de Darcy (1856)

On obtient une équation pour le mouvement en écrivant simplement la conservation du volume. Noter que, comme pour le modèle de Stokes, cette équation traduit un équilibre instantané des forces.

Modèle 2.17. *(Écoulement en milieu poreux)*
L'écoulement en milieu poreux saturé d'un fluide visqueux incompressible est régi par

$$\begin{cases} u + k\nabla \mathrm{p} &= U \\ \nabla \cdot u &= 0 \end{cases} \tag{2.6}$$

où p *est la pression au sein du fluide, u la vitesse de Darcy (voir remarque 2.16) , $k = K/\mu$ la perméabilité, et μ la viscosité du fluide. Nous avons noté U la force en volume exercée sur le fluide (c'est plus précisément U/k qui est homogène à une force par unité de volume).*

2.4 Cadre mathématique pour le problème de Darcy

Nous considérons un milieu poreux dont les bords sont "ouverts" (le fluide peut sortir du domaine ou y rentrer), et la pression au niveau du bord est imposée. On cherche un champ de vitesse u et un champ de pression p définis sur Ω tels que

$$\begin{cases} u + \nabla \mathrm{p} &= U \quad \text{dans } \Omega, \\ \nabla \cdot u &= 0 \quad \text{dans } \Omega, \\ \mathrm{p} &= 0 \quad \text{sur } \Gamma, \end{cases} \tag{2.7}$$

où U est un champ de force donné. On se place sur l'espace en vitesses $V = L^2(\Omega)^2$. On pose $\Lambda = H_0^1(\Omega)$, et l'on introduit l'application B de V dans $\Lambda' = H^{-1}$ qui à $v \in V$ associe la forme linéaire Bv définie par

$$\langle Bv\,,\,\mathrm{q}\rangle = \int_\Omega v\cdot\nabla\mathrm{q}.$$

On définit alors $K = \ker B$, et le problème de minimisation sous contrainte s'écrit

$$\left\{\begin{array}{l} u\in K = \left\{v\in L^2(\Omega)^2\,,\ \displaystyle\int_\Omega v\cdot\nabla\mathrm{q} = 0 \quad \forall \mathrm{q}\in H_0^1(\Omega)\right\},\\ J(u) = \displaystyle\inf_{v\in K} J(v), \quad \text{avec } J(v) = \frac{1}{2}\int_\Omega |v|^2 - \int_\Omega v\cdot f. \end{array}\right. \tag{2.8}$$

Proposition 2.18. *Soit Ω un domaine borné de frontière Lipschitz, et $U \in L^2(\Omega)^d$. Le problème de minimisation (2.8) ci-dessus admet une solution unique $u \in K$, et il existe un unique* $\mathrm{p} \in V = H_0^1(\Omega)$ *tel que*

$$u + \nabla\mathrm{p} = U \quad p.p.$$

Démonstration. Le problème (2.8) consiste à minimiser une fonctionnelle quadratique sur un sous-espace K fermé (K s'exprime comme le noyau d'une application linéaire continue). Il admet donc une solution unique $u \in K$.

Il reste à vérifier que le problème de point-selle associé est bien posé. En effet, l'application B est surjective, car son adjoint $B^\star : \mathrm{q} \longmapsto \nabla\mathrm{q}$ est tel que

$$|B^\star\mathrm{q}| = |\nabla\mathrm{q}|_{L^2(\Omega)} \geq \alpha\,|\mathrm{q}|_{H_0^1(\Omega)}\,,$$

d'après l'inégalité de Poincaré 22.43, page 228, ce qui assure bien la surjectivité de B selon la proposition 19.23, page 195. D'après la proposition 23.7, page 239, on a donc existence d'un multiplicateur de Lagrange p tel que $u + \nabla\mathrm{p} = U$, qui est unique du fait du caractère injectif du gradient sur $H_0^1(\Omega)$. □

2.5 Cadre mathématique pour les équations de Stokes

On cherche un champ de vitesse u et un champ de pression p définis sur Ω (les régularités de ces champs seront précisées par la suite) tels que

$$\left\{\begin{array}{rcl} -\triangle u + \nabla\mathrm{p} &=& f,\\ \nabla\cdot u &=& 0, \end{array}\right. \tag{2.9}$$

où f est un champ de force donné. On impose des conditions de Dirichlet homogènes sur la vitesses. La première des deux équations ci-dessus exprime l'équilibre des forces en chaque point du fluide, et la seconde exprime l'incompressibilité du fluide.

Nous allons maintenant préciser comment ce problème rentre le cadre de ce qui a été vu précédemment, en repartant du point de départ usuel qui est le problème de minimisation sous contrainte, puis en reconstruisant le problème de Stokes tel qu'énoncé ci-dessus à partir de la formulation point-selle.

On introduit les espaces

$$V = H_0^1(\Omega)^2\,,\ K = \{u\in V\,,\ \nabla\cdot u = 0\ \ p.p.\}\,,$$

On considère le problème de minimisation sous contrainte

$$\begin{cases} u \in K, \\ J(u) = \inf\limits_{v \in K} J(v), \end{cases} \tag{2.10}$$

où J est la fonctionnelle

$$J(v) = \frac{1}{2}\int_\Omega |\nabla v|^2 - \int_\Omega f \cdot v$$

Proposition 2.19. *La fonctionnelle J admet un unique minimiseur sur K.*

Démonstration. L'application $v \mapsto \nabla \cdot v$ étant linéaire continue (de V dans $L^2(\Omega)$), l'ensemble K est un sous-espace vectoriel fermé de V. De plus la fonctionnelle J est du type

$$J(v) = \frac{1}{2}a(v,v) - \langle \varphi \, , v \rangle,$$

où $a(\,\cdot\,,\,\cdot\,)$ est une forme bilinéaire symétrique continue et coercive sur V, et $\varphi \in V'$. Le théorème de Lax-Milgram assure l'existence et l'unicité d'un minimiseur. □

En vue d'écrire ce problème sous la forme d'une recherche de point-selle, nous introduisons maintenant l'espace

$$\Lambda = L^2_0(\Omega) = \left\{ \mathrm{p} \in L^2(\Omega) \, , \ \int_\Omega \mathrm{p} = 0 \right\},$$

et l'opérateur

$$B \, : \, v \in V \longmapsto Bv = -\nabla \cdot v.$$

L'espace K peut s'écrire

$$K = \left\{ v \in V \, , \ -\int_\Omega \mathrm{q} \nabla \cdot v = 0 \quad \forall \mathrm{q} \in \Lambda \right\},$$

ce qui conduit au Lagrangien

$$L(v, \mathrm{q}) = \frac{1}{2}\int_\Omega |\nabla v|^2 - \int_\Omega f \cdot v - \int_\Omega \mathrm{q} \nabla \cdot v.$$

Le caractère bien posé de la formulation point-selle est assuré par la

Proposition 2.20. *Soit Ω un domaine borné de frontière Γ Lipschitz, et $f \in L^2(\Omega)^N$. Le Lagrangien L défini ci-dessus admet un unique point-selle $(u, \mathrm{p}) \in V \times \Lambda$, où u est la solution du problème de minimisation sous contrainte (2.10). De façon équivalente, il existe un unique couple $(u, \mathrm{p}) \in H^1_0(\Omega)^N \times L^2_0(\Omega)$ tel que*

$$\int_\Omega \nabla u : \nabla v - \int_\Omega \mathrm{p} \nabla \cdot v \ = \ \int_\Omega f \cdot v \quad \forall v \in H^1_0(\Omega)^N \tag{2.11}$$

$$\int_\Omega \mathrm{q} \nabla \cdot u \ = \ 0 \quad \forall \mathrm{q} \in L^2_0(\Omega). \tag{2.12}$$

Démonstration. Malgré l'analogie formelle avec le problème de Darcy (l'opérateur B est l'opérateur de divergence dans les deux cas), la démonstration est plus délicate (voir par exemple [4]). L'existence et l'unicité d'un point-selle est une conséquence de la surjectivité de l'opérateur de divergence B, qui est assurée par le lemme 2.21 ci-après. □

Lemme 2.21. *Soit Ω un domaine connexe, borné, de frontière Γ Lipschitzienne, et soit* q *dans $L^2_0(\Omega)$. Il existe $v \in H^1_0(\Omega)$ tel que $\nabla \cdot v = \mathrm{q}$.*

Démonstration. On se reportera à [4, lemme 3.2] pour la démonstration de ce résultat. Noter que le théorème de l'application ouverte assure l'existence d'une constante C telle que l'antécédent v peut être choisi tel que $\|v\|_{H^1} \leq C \|\mathrm{q}\|_{L^2}$. □

Remarque 2.22. *Comme il a été précisé, établir l'existence et l'unicité d'une solution pour le problème de Stokes en formulation vitesse-pression est plus délicat que pour le problème de Darcy. Cette différence peut se préciser ainsi : dans le cas de Darcy, la démonstration repose sur une inégalité qui assure l'injectivité de $B^\star$ et le caractère fermé de son image. L'opérateur $B^\star$ va de $H^1_0(\Omega)$ dans $L^2(\Omega)^2$, et l'inégalité est conséquence directe de l'inégalité de Poincaré*

$$\|\mathrm{q}\|_{L^2(\Omega)} \leq C \|\nabla \mathrm{q}\|_{L^2(\Omega)^N} \quad \forall \mathrm{q} \in H^1_0(\Omega),$$

qui est vérifiée dès que Ω est borné dans une direction (voir proposition 22.43, page 228). Dans le cas de Stokes, la surjectivité de l'opérateur B peut être établie comme conséquence directe d'une inégalité à première vue très similaire, l'opérateur $B^\star$ étant toujours dans un certain sens l'opérateur de gradient, mais vu cette fois comme un opérateur de $L^2(\Omega)$ dans $H^{-1}(\Omega) = (H^1_0(\Omega)^N)'$. Cette inégalité peut s'écrire

$$\|\mathrm{q}\|_{L^2(\Omega)} \leq C \|\nabla \mathrm{q}\|_{H^{-1}(\Omega)} \quad \forall \mathrm{q} \in L^2_0(\Omega),$$

où $\nabla \mathrm{q}$ représente la forme linéaire sur $H^1_0(\Omega)^N$ définie par

$$v \longmapsto \int_\Omega \mathrm{q} \nabla \cdot v\,, \ \|\nabla \mathrm{q}\|_{H^{-1}(\Omega)} = \sup_{v \in H^1_0(\Omega)} \frac{\displaystyle\int_\Omega \mathrm{q} \nabla \cdot v}{\|v\|_{H^1_0(\Omega)^N}}.$$

2.6 Ecoulement de Poiseuille, notion de résistance

On s'intéresse ici à l'écoulement d'un fluide visqueux incompressible dans un conduit cylindrique à section circulaire.

$$\left\{ \begin{array}{rcl} -\mu \Delta u + \nabla \mathrm{p} & = & 0 \\ \nabla \cdot u & = & 0, \end{array} \right.$$

Le domaine est défini par

$$\Omega = \left\{ (x, y) \in \mathbb{R}^2\,, \ r^2 := x^2 + y^2 < a^2 \right\} \times (0, L).$$

On considère que le fluide adhère ($u = 0$) aux parois latérales. Le problème admet une solution exacte qui peut s'écrire en coordonnées cylindriques :

$$u(x, y, z) = U \left(1 - \frac{r^2}{a^2} \right) \vec{e}_z\,, \ \mathrm{p}(x, y, z) = -4 \frac{\mu U}{a^2} (z - z_0), \tag{2.13}$$

où U est la vitesse maximale (au centre). La pression est uniforme sur chaque section droite du tuyau,. Cela conduit à une relation linéaire entre le flux Q est le saut de pression :

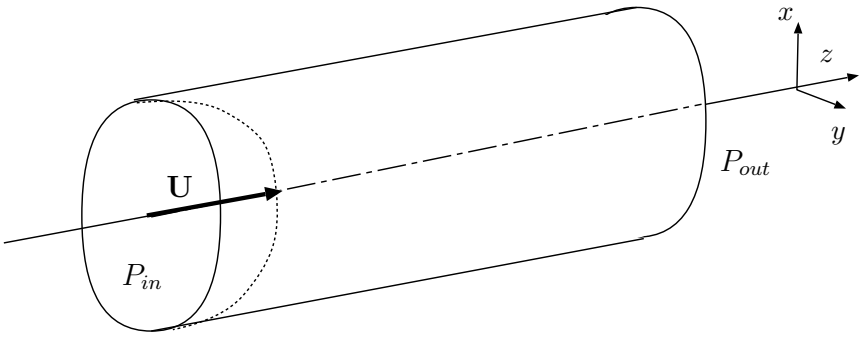


FIGURE 2.3 – Écoulement de Poiseuille

$$Q = U\pi\frac{a^2}{2} = \frac{\pi}{8}\frac{a^4}{\mu L}(P_{in} - P_{out}). \tag{2.14}$$

Cette relation s'appelle la *Loi de Poiseuille*, et s'écrit en général [17]

$$P_{in} - P_{out} = RQ, \tag{2.15}$$

avec

$$R = \frac{8\mu}{\pi}\frac{L}{a^4}. \tag{2.16}$$

La résistance visqueuse s'exprime en Pa s m^{-3}, Les forces de viscosité dissipent l'énergie au taux [18]

$$\mathcal{P} = \mu \int_\Omega |\nabla u|^2 .$$

Un calcul direct permet d'établir que $\mathcal{P} = RQ^2$ (on reconnaitrait un équivalent fluide de la loi de Joule), où Q est le flux défini précédemment.

On peut définir de façon générale la résistance d'un domaine $\Omega \in \mathbb{R}^d$, dont la frontière Γ se décompose en trois composantes

$$\Gamma = \Gamma_{in} \cup \Gamma_{out} \cup \Gamma_w,$$

Le *Pressure Drop Problem* s'écrit de la façon suivante

$$\begin{cases} -\mu\Delta u + \nabla \mathrm{p} &= 0 & \text{in } \Omega, \\ \nabla \cdot u &= 0 & \text{in } \Omega, \\ u &= 0 & \text{on } \Gamma_w, \\ \mu\nabla u \cdot n - \mathrm{p}\, n &= -P_{in}\, n & \text{on } \Gamma_{in}, \\ \mu\nabla u \cdot n - \mathrm{p}\, n &= -P_{out}\, n & \text{on } \Gamma_{out}. \end{cases} \tag{2.17}$$

17. Noter l'analogie entre cette loi et la loi d'Ohm

$$U = RI,$$

où I est le courant électrique au travers d'un conducteur, U la différence de potentiel , et R la résistance (électrique) du conducteur.

18. L'expression devrait être

$$\frac{\mu}{2}\int_\Omega \left|\nabla u + {}^t\nabla u\right|^2 ,$$

mais on peut montrer dans ce contexte, de fait que la vitesse s'annule au bord du domaine et est constante selon sa propre direction (bords libres), que les deux expressions sont équivalentes.

Les conditions en Γ_{out} et Γ_{in} sont appelées conditions de *sortie libre*, bien qu'elles concernent également l'entrée de fluide (dans le cadre linéaire, il n'y a pas lieu de distinguer l'entrée de la sortie). Elles expriment l'hypothèse que les deux composantes (amont Γ_{in} et aval Γ_{out}) sont placées toutes deux en contact avec un milieu pression fixée, qui équilibre la contrainte normale.

On peut définir la résistance du domaine :

Definition 2.23. *(Résistance d'un domaine (Stokes))*
Soit u le champ de vitesse solution de (2.17)*. Le flux Q est défini comme*

$$Q = -\int_{\Gamma_{in}} u \cdot n = \int_{\Gamma_{out}} u \cdot n. \tag{2.18}$$

Par linéarité des équations de Stokes, ce flux dépend linéairement su saut de pression $P_{in} - P_{out}$, et la résistance $R = R(\Omega)$ entre Γ_{in} et Γ_{out} est définie par

$$P_{in} - P_{out} = RQ. \tag{2.19}$$

On peut définir cette résistance de façon varationnelle, comme le minimum de l'énergie dissipée parmi les vitesses qui réalisent un flux unitaire au travers du domaine :

Proposition 2.24. *On définit*

$$K = \left\{ v \in H^1(\Omega)^d \,,\ v_{|\Gamma_w} = 0 \,,\ \nabla \cdot v = 0 \,,\ \int_{\Gamma_{in}} v \cdot n = -1 \right\}.$$

La résistance (définition 2.23) s'exprime alors

$$R = \inf_{v \in K} \mu \int_\Omega |\nabla u|^2 \,.$$

3 Piétons - Micro - ordre 1 en temps - approche granulaire

On s'intéresse ici à la modélisation microscopique (les agents sont individualisés) de mouvements de foules d'un type particulier : on considère que chaque personne tend à suivre sa vitesse *souhaitée* (vitesse qu'elle souhaiterait avoir si elle était seule), et que la vitesse effective de la collection d'individus est la vitesse globale la plus proche (au sens des moindres carrés) de la vitesse souhaitée globale.

3.1 Modèle monodimensionnel

On considère N individus assujettis à se déplacer en ligne droite (comme dans un couloir étroit). Les positions sont notées q_1, ..., q_N, initialement ordonnées conformément à l'indexation, et l'on considérera que les personnes sont identifiées à des disques rigides de rayon r (ou ici à des segment de longueur $2r$). On considèrera comme admissibles les configurations de

$$K = \left\{ q = (q_1, q_2, \dots, q_N) \in \mathbb{R}^N \,,\ q_{i+1} - q_i \geq 2r \,,\ i = 1, \dots, N-1 \right\}.$$

On suppose qu'une vitesse souhaitée U_i est attachée à chaque individu, et que la vitesse effective de la population est la plus proche (pour la norme euclidienne) de la vitesse globale souhaitée, parmi les vitesses admissibles. L'ensemble des vitesses admissibles est défini par [19]

$$C_q = \left\{ v = (v_1, \dots, v_N) \in \mathbb{R}^N \,,\ q_{i+1} - q_i - 2r = 0 \Longrightarrow v_{i+1} - v_i \geq 0 \right\}.$$

Le problème s'écrit donc

$$\frac{dq}{dt} = u \,,\ u = P_{C_q} U.$$

Formulation point-selle

Le problème de projection qui définit la vitesse instantanée consiste à minimiser la fonctionnelle

$$J(v) = \frac{1}{2} \left| v - U \right|^2, \tag{3.1}$$

sur l'ensemble C_q des configurations admissibles. Cet ensemble est une intersection de demi-espaces affines, il s'agit donc bien d'un convexe fermé, l'existence et l'unicité d'un minimiseur est alors immédiate.

Le critère d'admissibilité consiste en la vérification d'une série de contraintes affines. On peut rassembler ces contraintes sous forme matricelle, en introduisant la matrice B donc une ligne est du type

$$(0, \dots, 0, 1, -1, 0, \dots, 0),$$

où les éléments non nuls correspondent à deux indices successifs i et $i+1$, où i est tel que $q_{i+1} - q_i - 2r = 0$ (contact entre i et $i+1$). On peut ainsi écrire

$$C_q = \left\{ v \in \mathbb{R}^N \,,\ Bv \leq 0 \right\}. \tag{3.2}$$

19. On écrit simplement que, lorsque 2 individus sont en contact, la distance ne peut pas diminuer.

Proposition 3.1. *Le problème consistant à minimiser la fonctionnelle J (définie par* (3.5)*) sur C_q (défini par* (3.2)*) est équivalent à la formulation point-selle suivante*

$$\left|\begin{array}{rcl} u + B^\star p & = & U, \\ Bu & \leq & 0, \\ p & \geq & 0, \\ Bu \cdot p & = & 0. \end{array}\right. \tag{3.3}$$

Plus précisément, u étant la solution du problème de minimisationsous contrainte, il existe un unique p tel que le système ci-dessus soit vérifié. Réciproquement, si le couple (u, p) vérifie ce système, alors u est bien la solution du problème de minimisation sous contrainte.

Démonstration. Les contraintes étant affines, elles sont automatiquement qualifiées (définition 23.21, page 244). La proposition 23.22 assure donc l'existence d'un vecteur p de multiplicateurs de Lagrange tel que le système (3.6) ci-dessus soit vérifié. Réciproquement, si (u, p) est solution du système, le théorème 23.29, page 248 assure que ce couple est point-selle du Lagrangien

$$L(v, q) = \frac{1}{2}\left|v - U\right|^2 + q \cdot Bv,$$

et donc que u minimise la fonctionnelle quadratique sous la contrainte $Bu \leq 0$ (d'après la proposition 23.28, page 247). □

Si l'on considère une rangée de personnes 1, ..., N saturée, i.e. chaque individu est en contact avec ses voisins la matrice des contraintes s'écrit

$$B = \begin{pmatrix} 1 & -1 & 0 & \dots & \dots \\ 0 & 1 & -1 & \dots & \dots \\ 0 & 0 & \ddots & \ddots & \dots \\ 0 & 0 & \dots & 1 & -1 \end{pmatrix}$$

Cette matrice exprime une version discrète de $-\partial_x$ (opposé de la divergence en dimension 1), et $B^\star$ correspond à ∂_x (gradient). Dans le cas où toutes les contraintes sont saturées (par exemple si l'on suppose que les vitesses souhaitées sont décroissantes : les personnes devant ont tendance à aller moins vite que les personnes derrière), on aura $Bu = 0$, ce qui implique

$$BB^\star p = BU.$$

La matrice $BB^\star$, d'ordre $N-1$, est exactement la matrice du Laplacien discret en dimension 1 avec conditions de Dirichlet aux extrémités (matrice donnée par (A.13), page 257). Le champ des pressions entre individus apparaît donc comme solution d'un problème de *Poisson* discret, avec un terme source qui quantifie, à partir de l'information sur les vitesses souhaitées, la tendance à violer la contrainte de non chevauchement. On retrouve bien, conformément à l'intuition, que si BU est positif (vitesse souhaitée décroissante), toutes les pressions seront non nulles.

Remarque 3.2. *Les remarques précedentes (sur le fait que B encode l'opposé d'une divergence discrète) renforcent l'analogie formelle entre le problème* (3.6) *et le problème de*

Darcy, telle qu'elle apparaît pour modéliser les écoulements en milieux poreux (équation (2.6)*, page 31, ou sous forme plus abstraite dans le cadre des réseaux résistifs (équation* (4.1)*, page 45).*

Remarque 3.3. *Cette formulation permet de comprendre, dans un contexte très simplifié, les phénomènes d'accumulation de pression au sein d'une foule présentant des tendances concentrantes (ce qui se traduit ici par une divergence de la vitesse discrète négative, i.e. BU localement positif). Si l'on considère par exemple le cas de $N/2$ personnes souhaitant aller vers la droite, et $N/2$ personnes, sur leur droite, souhaitant aller vers la gauche, BU est la version discrète d'une masse de Dirac au point de contact entre les deux populations, et le champ de pression est de type affine par morceaux (fonction chapeau), avec une pression maximale au point de jonction. Toute choses égales par ailleurs, la pression maximale tend vers $+\infty$ quand le nombre d'individu tend vers $+\infty$, dans ce contexte de "mêlée" monodimensionnelle. Notons aussi que le le caractère* sphère dure *du modèle considéré conduit à des effets non locaux, avec propagation de l'information à vitesse infinie au sein du réseau de personnes. Dans l'exemple ci-dessus, le chagement de vitesse souhaitée d'un individu particulier va changer instantanément les vitesses réelles de tous les individus.*

3.2 Modèle en dimension 2 (disques rigides)

On représente comme précédemment les individus par des disques de rayon r, on introduit le vecteurs des positions :

$$q = (q_1, q_2, \dots, q_N) \in \mathbb{R}^{2N}.$$

L'ensemble des configurations admissibles est défini par

$$K = \left\{ q \in \mathbb{R}^{2N} \,,\ D_{ij} = |q_j - q_i| - 2r \geq 0 \quad \forall i \neq j \right\}.$$

On se donne comme une collection de vitesses souhaitées

$$U = (U_1, \dots, U_N).$$

L'hypothèse la plus simple consiste à supposer que chaque U_i ne dépend que de la position de l'individu i (qui n'adapte donc pas sa stratégie aux positions de ses voisins), dans ce cas on aura $U_i = U_0(q_i)$, où U_0 est un champ de vitesse commun à tous les individus. On peut considérer des modèles plus complexes en écrivant plus généralement $U = U(q)$, qui exprime que la vitesse souhaitée d'un individu dépend de sa propre position, mais aussi potentiellement des positions des autres individus (possibilité de modéliser des stratégies individuelles).

Notons $G_{ij} = \nabla D_{ij}(q)$ le gradient de la fonction distance de i à j. Le cône des vitesses admissibles associé à une configuration q est alors

$$C_q = \{ v \,,\ D_{ij}(q) = |q_j - q_i| - 2r = 0 \Rightarrow G_{ij} \cdot v \geq 0 \}. \tag{3.4}$$

Noter que $G_{ij} \in \mathbb{R}^{2N}$ n'a que 4 composantes non nulles, correspondant aux positions des individus i et j. Le modèle d'évolution exprime simplement le fait que la vitesse effective de la population est la plus proche au sens des moindres carrés de la vitesse souhaitée :

$$\dot{q} = P_{C_q} U(q),$$

où P_{C_q} est la projection pour la norme euclidienne sur le convexe fermé C_q, définie de façon unique (proposition 20.7, page 197) et stable (proposition 20.10).

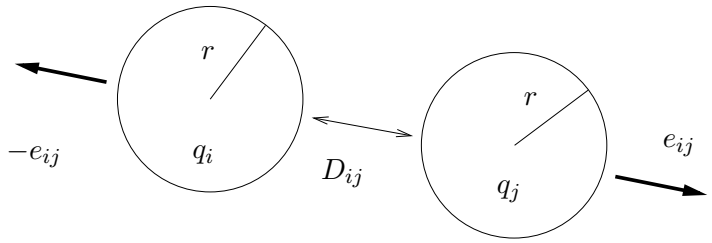


FIGURE 3.1 – Notations.

Formulation point-selle

Comme dans la situation précédente, le problème de projection qui définit la vitesse instantanée consiste à minimiser la fonctionnelle

$$J(v) = \frac{1}{2}\left|v - U\right|^2, \tag{3.5}$$

sur l'ensemble C_q des configurations admissibles, qui peut s'écrire sous forme matricielle

$$C_q = \left\{ v \in \mathbb{R}^N \,,\ Bv \leq 0 \right\},$$

où chaque ligne de la matrice B exprime une contrainte de non chevauchement entre deux disques en contact dans la configuration courante. Plus précisément, pour 2 entités i et j en contact, on définit le vecteur unitaire centre à centre (voir figure 3.1)

$$e_{ij} = \frac{q_j - q_i}{|q_j - q_i|}.$$

Le gradient de la distance entre i et j, vue comme fonction de l'ensemble des degrés de liberté, s'écrit

$$G_{ij} = (0, \ldots, 0, -e_{ij}, 0, \ldots, 0, e_{ij}, 0, \ldots, 0) \in \mathbb{R}^{2N}.$$

Proposition 3.4. *Le problème consistant à minimiser la fonctionnelle J (définie par* (3.5)*) sur C_q (défini par* (3.4)*) est équivalent à la formulation point-selle* (3.6)*, qui peut s'exprimer sous la forme suivante*

$$\left|\begin{array}{rcll} u - \displaystyle\sum_{i\sim j} p_{ij}\, G_{ij} & = & U, & \\ -G_{ij}\cdot u & \leq & 0 & \forall i \sim j, \\ p & \geq & 0, & \\ G_{ij}\cdot u > 0 & \Longrightarrow & p_{ij} = 0. & \end{array}\right. \tag{3.6}$$

Démonstration. La démonstration est parfaitement analogue à celle de la proposition 3.1. □

On s'intéresse maintenant aux propriétés de la matrice $BB^\star$, identifée précédemment à (l'opposé d'un) opérateur de Laplace discret dans le cas de la dimension 1.

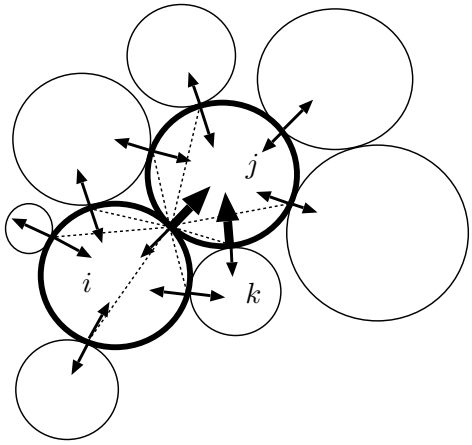


FIGURE 3.2 – Stencil non structuré

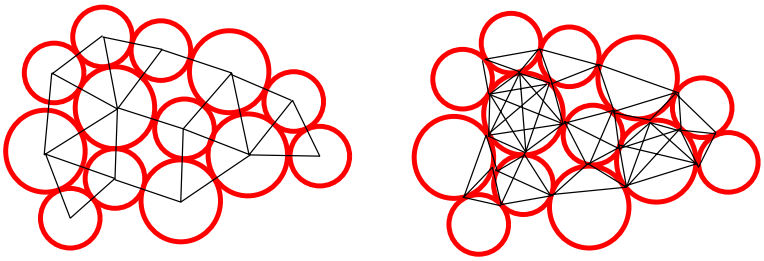


FIGURE 3.3 – Réseaux primal (gauche) et dual (droite)

Considérons une configuration $q \in K$ (voir figure 3.2), et la matrice associée B, dont chaque ligne exprime une contrainte du type

$$-G_{ij} \cdot u \leq 0,$$

où G_{ij} est le gradient de la distance $D_{ij} = |q_j - q_i| - r_i - r_j$ par rapport à $q = (q_1, \ldots, q_N)$. L'opérateur discret $B^\star$ a été identifié dans le cas de la dimension 1 à un gradient discret. Considérons dans le cas présent une collection p de multiplicateurs de Lagrange. L'opération $-B^\star$ réalise l'action de ces forces d'interaction sur le réseau primal de degré de liberté associés aux centres des particules. dans le cas d'une configuration structurée, (par exemple réseau cartésien, ou réseau triangulaire comme représenté sur la figure 3.4) un champ de pression p uniforme est de gradient discret nul sur les points intérieurs au réseau [20]. Cependant, dans le cas général, (quand l'arrangement des disques ne présente pas de symétrie particulière), cette propriété est invalidée. Par exemple dans le cas de la figure 3.2 on vérifiera immédiatement que la somme des vecteurs unitaires pointant vers l'intérieur de chacun des deux grains en gras n'est pas nulle. Le cas bidimensionnel non structuré présente une autre particularité. Considérer le cluster représenté sur la figure 3.4. Le nombre de disques est 14, donc le nombre

20. On retrouve ici la version discrète d'annulation du gradient d'une fonction constante. Plus précisément, pour comprendre la présence d'une résultante non nulle au bord, on peut penser, dans le cas continu, au gradient faible d'une fonction caractéristique d'un domaine borné. Son gradient est effectivement nul à l'intérieur, nul à l'intérieur de l'extérieur, mais il s'identifie globalement à une distribution vectorielle de simple couche supportée par la frontière de l'ensemble.

de degrés de liberté primaux est 28, et le nombre de contacts (nombre de degrés de liberté duaux) est 29. En conséquence, le noyau de $B^\star \in \mathcal{M}_{29,28}(\mathbb{R})$ est non trivial : il existe un champ de pression non identiquement nul (mais nul au bord d'une certaine manière, selon la remarque ci-dessus), induisant une force non nulle sur les grains [21]. Une conséquence de ces comportements pathologiques est que l'opérateur discret $BB^\star$, que l'on pourrait être tenté de considérer comme un Laplacien discret défini sur le graphe dual du réseau de disques (représenté à droite de la figure 3.3) ne vérifie pas le principe du maximum : il peut exister des champs de pression p tels que $BB^\star \geq 0$ (i.e. les pressions contribuent à l'augmentation de toutes les distances entre centre), alors que certaines composantes de p sont strictement négatives.

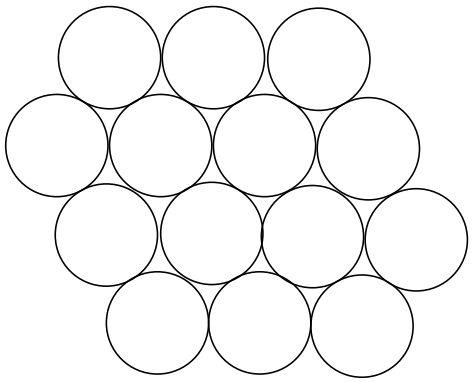


FIGURE 3.4 – Situation hyperstatique (28 degrés de liberté pour 29 contraintes)

L'opérateur discret $BB^\star$ peut se décrire comme suit : considérant un champ de pressions $p = (p_{k\ell})$, où (k, ℓ) parcours l'ensemble des contacts actifs, le vecteur $BB^\star p$ est un vecteur qui vit lui même sur le graphe dual (comme les pressions), et la valeur correspondant aux disques i et j est

$$\sum_{(k,\ell)\sim(i,j)} p_{k\ell} G_{ij} \cdot G_{k\ell}.$$

Par analogie avec la méthode des différences finies, il est tentant de parler de *stencil* associé à cet opérateur. Ce stencil est représenté sur la figure 3.2. La non vérification du principe du maximum est due au fait que, lorsque l'on considère 3 particules i, j, et k, il peut arriver que l'on ait

$$e_{ij} \cdot e_{kj} > 0,$$

où e_{ij} est le vecteur unitaire $(q_j - q_i)/\mid q_j - q_i|$. Des exemples de tels vecteurs sont représentés sur la figure 3.2 en gras. Cette propriété est générique pour des collections de disques congestionnées. Certains éléments extra diagonaux de la matrice $BB^\star$ sont alors *strictement positifs*, et ainsi la matrice $BB^\star$ n'est *pas* une M-matrice [22]. Le réseau résisif associé à cet

21. On peut illustrer cette propriété de la façon suivante : si l'on considère par exemple deux disques rigides, statiques, en contact (éventuellement collés entre eux) posés sur un support parfaitement glissant, on sait que la force d'interaction entre eux est nulle. Ça n'est plus vrai pour la configuration de la figure 3.4 : il est possible que les forces d'interactions soient non nulles. On peut en revanche montrer (grâce au théorème de Hahn Banach) que ces forces ne peuvent pas être toutes positives

22. Une M-matrice est une matrice carrée dont tous les mineurs principaux sont strictement positifs, et dont tous les éléments extra-diagonaux sont négatifs (au sens large). Tous les éléments de l'inverse d'une telle matrice sont positifs, de telle sorte que $Ap = b$, avec $b \geq 0$, implique $p \geq 0$.

opérateur possède donc des résistances *négatives* : on retrouve la situation de certaines matrices résultant de la discrétisation du Laplacien par éléments fini, sur un maillage contenant des triangles *amblygone*[23] (voir section 17.6, page 176).

23. Terme désignant un triangle qui a un angle obtus, peu utilisé depuis quelques siècles, mais quand même plus élégant que *obtusangle*.

4 Réseaux résistifs

On s'intéresse ici à la propagation d'une quantité au travers d'un réseau, en supposant que le flux au travers de chaque arête est proportionnel à la différence de potentiels définis à ses extrémités (sommets, ou points de bifurcation du réseau).

Dans le cas de l'écoulement d'un fluide visqueux, c'est la *pression* aux nœuds qui jouera le rôle du potentiel, dont la différence induit un flux selon la loi de Poiseuille (équation (2.15), page 35). Pour un réseau électrique, c'est le potentiel électrique aux extrémités de chaque arête qui induira le passage d'un courant électrique quantifié par son intensité. On peut aussi imaginer des compartiments séparés par des interfaces faiblement perméables à une certaine substance qui diffuse. Dans l'hypothèse de pressions partielles uniformes dans chaque compartiment, et de flux au travers des interface proportionnels aux sauts de pression partielle, on aura aussi une représentation naturelle du phénomène de diffusion sous forme de réseau résistif, où les pressions partielles jouent le rôle du potentiel électrique.

Dans tous les cas, on écrira le bilan de matière au sein du réseau (loi de Kirchhof, ou loi des *nœuds*). Nous ferons par la suite la distinction entre des points *internes*, en lesquels la loi de Krichhof s'applique, et les autres, au travers desquels le réseau est susceptible d'échanger de la matière avec l'extérieur.

4.1 Cadre formel, problème de Laplace discret

Definition 4.1. *(Réseau résistif)*
*Un réseau résistif fini est défini comme un triplet $N = (V, E, r)$, où V est un ensemble fini de côté (*Vertices*), $E \subset V \times V$ un ensemble d'arrêtes (*Edges*) supposé symétrique*[24] *:*

$$(x, y) \in E \Longrightarrow (y, x) \in E,$$

et $r \in \mathbb{R}_+^E$ est le champ des résistances, défini sur E (avec $r(x, y) = r(y, x)$ pour tout $(x, y) \in E$). On notera $\mathcal{N} = (V, E, r, o, \Gamma)$ un réseau dans lequel on distingue une racine o parmi les sommets, et une frontière Γ, sous-ensemble non vide de $V \setminus \{o\}$. L'ensemble $V \setminus (\{o\} \cup \Gamma)$ des sommets intérieurs est noté $\mathring{V}$, il correspond aux sommets (ou nœuds) en lesquels on imposera la conservation de la matière, alors que de la matière peut entrer ou sortir du domaine par les points de Γ, ou par la racine o.

Un champ de pressions sur le réseau est une collection de réels associés aux sommets ($p \in \mathbb{R}^V$), et les flux sont définis sur les arêtes ($u \in \mathbb{R}^E$). Les flux sont antisymétriques : $u(x, y) = -u(y, x)$.

Pour une arête $e = (x, y)$ du réseau, la loi de Poiseuille s'écrit

$$p(x) - p(y) = r(x, y)u(x, y) = r(e)u(e).$$

Si l'on note maintenant $j(x)$ le flux de matière injectée dans le réseau au travers du nœud x la loi de Kirchhof (ou loi des nœuds) s'écrit

$$\sum_{y \sim x} u(x, y) = j(x),$$

24. On considèrera cependant que, dans les sommes sur l'ensemble des arêtes, on ne compte qu'une fois chaque paire de points connectés.

où $y \sim x$ signifie que y est relié à x (i.e. $(x,y) \in E$).

On note d l'opérateur de divergence discrète (il s'agit en fait de l'*opposé* formel de la divergence)

$$\begin{aligned} d \ : \ u \in \mathbb{R}^E \quad &\longmapsto du \in \mathbb{R}^V \\ du(x) \quad &= -\sum_{y \sim x} u(x,y). \end{aligned}$$

Nous nous intéresserons dans la suite à des flux conservatifs, i.e. tels que $du(x) = 0$ pour tout sommet x dans $\mathring{V} = V \setminus (\{o\} \cup \Gamma)$. On défini l'adjoint formel [25] $d^\star$ (équivalent discret de l'opérateur de gradient) comme

$$\begin{aligned} d \ : \ p \in \mathbb{R}^V \quad &\longmapsto d^\star p \in \mathbb{R}^E \\ d^\star p(e) \quad &= p(y) - p(x). \end{aligned}$$

Remarque 4.2. *On établit immédiatement un équivalent discret du théorème de la divergence*

$$\int_\Omega \nabla \cdot v = \int_{\partial\Omega} v \cdot n.$$

On a en effet, pour tout $e = (x,y) \in E$, $u(x,y) + u(y,x) = 0$, d'où, en sommant sur toutes les arêtes, et en écrivant la somme sur les sommets :

$$\sum_x du(x) = 0,$$

qui exprime simplement le bilan de matière sur l'ensemble du réseau. On peut l'écrire

$$\sum_{x \in \mathring{V}} du(x) + \sum_{x \in \{o\} \cup \Gamma} du(x) = 0.$$

Le premier terme est le pendant discret de (l'opposé de) l'intégrale de la divergence dans le domaine, et le second terme est la somme pour tous les points du bord des flux qui sortent par ces points, i.e. l'équivalent discret de l'intégrale sur la frontière de $u \cdot n$.

L'écriture de la loi de Poiseuille en chaque arête, et de la loi de Kirchhoff's en chaque nœud conduit à un problème de type Darcy

$$\left\{ \begin{array}{rcl} u + cd^\star p &=& 0 \quad \text{sur} \quad E \\ du &=& 0 \quad \text{sur} \quad \mathring{V}. \end{array} \right. \tag{4.1}$$

où c (conductance) est $1/r$, i.e. $c(e) = 1/r(e)$ pour tout $e \in E$. On s'intéresse au problème consistant à calculer les pressions et les flux sur l'ensemble du réseau, quand les pressions sont prescrites en o et sur Γ. Après élimination de la vitesse, on obtient un problème de Poisson discret pour la pression, avec conditions de Dirichlet :

$$\left\{ \begin{array}{rcl} dcd^\star p(x) &=& 0 \quad \forall x \in \mathring{V}, \\ p(o) &=& 0 \\ p(x) &=& P(x) \quad \forall x \in \Gamma, \end{array} \right. \tag{4.2}$$

où P est une collection de pressions prescrites sur la frontière Γ.

25. On a

$$\sum_x q(x) dv(x) = \sum_x \sum_y q(x) v(y,x) = \sum_e v(e) \underbrace{(q(y) - q(x))}_{d^\star q(e)}.$$

Proposition 4.3. *On suppose le réseau $\mathcal{N}$ connexe. Le problème* (4.2) *est alors bien posé.*

Démonstration. On définit H comme l'ensemble des champs de $\mathbb{R}^V$ nuls en o, et H_0 le sous-espace des champs nuls sur Γ. Comme dans le cas continu (début de la section 17), on peut écrire une formulation variationnelle en considérant $q \in H_0$, en multipliant la première équation de (4.2) par $q(x)$, et en sommant sur les x, pour obtenir (en utilisant $q(y) = 0$ pour tout $y \in H_0$)

$$\sum_x q(x) \sum_{y \sim x} c(x,y)(p(x) - p(y)) = \sum_e c(e)(p(y) - p(x))(q(y) - q(x)) = 0$$

On reconnait les conditions d'optimalité pour la fonctionnelle

$$q \longmapsto J(q) = a(q,q) = \frac{1}{2} \sum_e c(e)(q(y) - q(x))^2,$$

minimisée sur l'espace affine $H_P \subset H$ des champs qui valent P sur Γ. Cette fonctionnelle J est une forme quadratique définie positive dès que le réseau est connexe, car les champs qui annulent J sont constants sur le réseau, et nuls en o. □

On remarquera que

$$a(p,p) = \sum_e c(e) \left| p(y) - p(x) \right|^2,$$

est le taux d'énergie effectivement dissipée au sein du réseau : la solution de (4.2) est, parmi les champs de pression qui vérifient les conditions aux limites, celui qui induit une puissance dissipée minimale.

Remarque 4.4. *Noter que, dans le problème d'optimisation intervenant dans la preuve précédente, on n'impose pas la loi des nœuds sur les flux associés à la pression p. La conservation au niveau des points interieurs est* conséquence *du caractère minimisant de p.*

Definition 4.5. *(Résistance équivalent d'un réseau)*
Soit $\mathcal{N} = (V, E, r, o, \Gamma)$ un réseau (selon la Def. 4.1). On impose un champ de pression uniforme $P \equiv 1$ sur Γ. On note p la solution du problème de Dirichlet (4.2), et par $u = -cd^\star p$ le flux associé. Le flux global Q est obtenu en sommant les flux au travers de Γ, ou de façon équivalent en considérant le flux qui sort par la racine o :

$$Q = -\sum_{x \sim o} u(o,x) = du(o). \tag{4.3}$$

La résistance équivalente de $\mathcal{N}$ est définie comme $R(\mathcal{N}) = 1/Q = 1/du(o)$. Par linéarité, le flux associé à une pression uniforme P sur Γ vérifie $P - 0 = RQ$.

Proposition 4.6. *(Loi de Joule pour un réseau)*
Soit $\mathcal{N} = (V, E, r, o, \Gamma)$ un réseau, et p la solution du problème (4.2) *associée à une pression uniforme P. Le taux d'énergie dissipée dans le réseau s'écrit*

$$\mathcal{P} = RQ^2,$$

où $Q = du(o)$ est le flux de Γ à o.

Démonstration. C'est une conséquence de la formule de Green discrète (sommation par parties). L'énergie dissipée s'écrit

$$\begin{aligned}\mathcal{P} &= \sum_E c(x,y)(p(x)-p(y))^2 \\ &= \sum_{x\in\mathring{V}} p(x) \underbrace{\sum_{y\sim x} c(x,y)(p(x)-p(y))}_{=dcd^\star p(x)=0} + \sum_{x\in\{o\}\cup\Gamma} p(x)\sum_{y\sim x} c(x,y)(p(x)-p(y)) \qquad (4.4)\\ &= P\sum_{x\in\Gamma} dcd^\star p(x) = -P\sum_{x\in\Gamma} du(x) = Pdu(o) = Rdu(o)^2,\end{aligned}$$

ce qui termine la preuve. □

Remarque 4.7. *Précisons les similarités et différences entre ce cadre discret et le cadre continu (équations de Darcy* (2.6)*, page 31). La formle de Green utilisée précédemment*

$$\sum_E c(x,y)(p(x)-p(y))(q(x)-q(y)) = \sum_{x\in V} q(x)\sum_{y\sim x} c(x,y)(p(x)-p(y)),$$

est analogue à la même formule dans un domain continu sans bord (par exemple pour l'espace entier, ou un domaine périodique). De fait, la notion de frontière pour un réseau est arbitraire, et nous n'avons d'ailleurs fait aucune hypothèse sur les sommets de Γ*. En particulier, il peuvent être situés au sein même du réseau, avoir un nombre arbitraire de voisins, etc . . .Nous avons obtenu une sorte de terme de bord en décomposant l'ensemble des sommets entre* $\mathring{V}$ *et* $\{o\}\cup\Gamma$*, et la formule obtenue n'a pas véritablement d'équivalent continu. En effet, la transposition du cadre discret conduit à considérer le problème*

$$-\Delta \mathrm{p} = 0 \quad in\ \Omega\setminus X$$

où Ω *est un domaine sans frontière, et* X *une collection finie* (x_i) *de points de* Ω*, avec une valeur de pressions* p_i *prescrite en* x_i*, de telle sorte que*

$$-\Delta \mathrm{p} = \sum_i u_i \delta_{x_i}$$

où u_i *est le flux rentrant e,* x_i*. On a alors formellement*

$$\int_\Omega |\nabla \mathrm{p}|^2 = \sum_i u_i \mathrm{p}_i,$$

qui serait l'équivalent discret de (4.4)*. Le problème est que cette expression n'a pas de sens, car les points ont une capacité nulle en dimension* $d\geq 2$ *(voir exercice 22.1, page 223).*

Pour obtenir une formule de Green avec termes de bords qui contiendraient un équivalent discret de $\int_\Gamma \partial p/\partial n$*), on doit introduire un ensemble d' "arêtes frontières"* E^Γ*, i.e. l'ensemble des* Γ *arêtes qui contiennent un point de* Γ *On a alors*

$$\begin{aligned}\sum_E c(x,y)(p(x)-p(y))(q(x)-q(y)) &= \sum_{x\in\mathring{V}} q(x) \underbrace{\sum_{y\sim x} c(x,y)(p(x)-p(y))}_{=dcd^\star p(x)} \\ &\quad + \sum_{x\in\{o\}\cup\Gamma} q(x)\sum_{y\sim x} c(x,y)(p(x)-p(y)) \\ &= \sum_{x\in\mathring{V}} q(x)dcd^\star p(x) - \sum_{e=(x,y)\in E^\Gamma} c(x,y)q(x)d^\star p(e),\end{aligned}$$

qui est maintenant l'équivalent discret de

$$\int_\Omega k\nabla \mathrm{p}\cdot\nabla \mathrm{q} = -\int_\Omega q\nabla\cdot k\nabla \mathrm{p} + \int_\Gamma k\frac{\partial \mathrm{p}}{\partial n}.$$

4.2 Squelette métrique associé à un réseau résistif

Dans le contexte de circulation de flux étudié dans la section précédente, il est naturel d'associer à un réseau $\mathcal{N} = (V, E, r)$ l'espace métrique défini de la façon suivante. En premier lieu, on métrise V (relativement à E et r) en considérant que la longueur l'une arête $e = (x, y) \in E$ (donc la distance de x à y) est $r(e)$. Pour deux points du réseaux non directement connectés, on définit la distance entre eux comme la longueur du plus court chemin qui les relie. On peut donner un peu de "corps" à cet espace métrique en considérant maintenant chaque arête (x, y) comme un segment plein, ensemble de points définis de façon abstraite [26] comme

$$[e] = [x, y] = \{(1-\theta)x + \theta y\,,\ \theta \in [0, 1]\}\,.$$

On dira que la distance d'un tel point à x (resp. y) est $\theta\, r$ (resp. $(1-\theta)r$). Ce choix définit de façon immédiate une métrique sur la réunion des segments. On notera $\overline{\mathcal{N}}$ le nouvel espace métrique ainsi défini.

Si l'on considère maintenant un champ de pression de $\mathbb{R}^V$, on peut définir de façon canonique un champ de pression $\overline{p}$ sur $\overline{\mathcal{N}}$ affine par morceaux (sur chaque arête), et un champ de flux $\overline{u}$ constant par morceaux. Si $u = -cd^\star p$ (sur $\mathcal{N}$), on a immédiatement, sur chaque arête

$$\overline{u}(s) \equiv u(e) = -\frac{1}{r(e)}(p(y) - p(x)) = -\partial_s \overline{p}.$$

Avec des notations évidentes, on peut écrire le taux d'énergie dissipée sous une forme intégrale

$$\sum_e r(e)u(e)^2 = \sum_e \int_e u(e)^2\, ds = \sum_e \int_e |\partial_s \overline{p}|^2\, ds = \int_{\overline{\mathcal{N}}} |\partial_s \overline{p}|^2\, ds.$$

On retrouve de cette manière l'expression classique de la semie-norme de Sobolev. On prendra garde au fait que l'abscisse curviligne (tout comme la variable d'espace qui intervient dans la dérivée) est homogène ici à une *résistance*.

4.3 Cadre stochastique

Soit un réseau $\mathcal{N} = (V, E, r)$ (voir définition 4.1), on considère la marche aléatoire sur V associée aux probabilités de transitions π_{xy}, définies par

$$\pi_{xy} = \frac{c(x, y)}{C(x)}\,,\ C(x) = \sum_{y\sim x} c(x, y)\,, \tag{4.5}$$

26. Cette démarche peut en effet être menée dans un cadre assez abstrait : chaque segment de notre espace métrique sera de fait isométrique à un segment de longueur $r(e)$ dans $\mathbb{R}^d$, mais il n'est pas nécessaire de plonger le réseau dans l'espace euclidien pour définir le nouvel espace, pour lequel les points de bifurcation restent des points abstraits, indépendamment de toute structure affine. On pourrait d'ailleurs décider de dédoubler certaines arêtes, qui se retrouveraient confondues dans une réprésentation plate et rectiligne du réseau, mais en restant différente pour $\mathcal{N}$ (la distance entre leurs milieux serait par exemple r).

où $c(x,y) = 1/r(x,y)$ est la conductance de l'arête (x,y). La chaîne de Markov associée est irréductible dès que le réseau est connexe, ce que nous supposerons ici. Elle admet donc une unique mesure stationnaire (voir théorème A.8, page 257), que l'on identifie immédiatement comme $C(x)$ (on normalise les résistances de départ de façon à ce que C soit effectivement de masse totale égale à 1).

On considère maintenant un réseau $\mathcal{N} = (V, E, r, o, \Gamma)$ et la donnée d'un champ de pressions $(P(x))_{x\in\Gamma}$ sur la frontière, et $P(o) = 0$. On définit $p \in \mathbb{R}^V$ comme suit : considérant un sommet $x \in V$, on note i la variable aléatoire correspondant à l'instant où la marche aléatoire issue de x atteint Γ ou o :

$$X_0 = x\,,\ X_1, \dots\,,\ X_i \in \Gamma \cup \{o\}\,,$$

avec $X_j \notin \Gamma \cup \{o\}$ pour $0 < j < i$. La valeur de P en X_i (qui est nulle si $X_i = o$) est une variable aléatoire, dont on note $p(x)$ l'espérance. On peut établir le lien suivant avec le problème de Dirichlet (4.2).

Proposition 4.8. *Le champ $p \in \mathbb{R}^V$ défini précédemment est la solution du problème* (4.2).

Démonstration. Remarquons en premier lieu que les conditions de Dirichlet sont automatiquement vérifiées par la probabilité p (quand $x \in \Gamma \cup \{o\}$, l'indice i est 0, et la variable aléatoire considérée est en fait déterministe). Considérons maintenant $x \in \mathring{V}$. On a

$$p(x) = \sum_{y\sim x} \pi_{xy} p(y),$$

qui peut s'écrire (d'après (4.5))

$$C(x)p(x) - \sum_{y\sim x} c(x,y)p(y) = 0,$$

de telle sorte que p est harmonique. Il s'agit donc nécessairement de l'unique solution du problème de Dirichlet (4.2). □

Remarque 4.9. *La matrice de transition P associée à la marche aléatoire définie précédemment est reliée au Laplacien discret de la façon suivante :*

$$P = (p_{xy})_{x,y\in V}\ ,\ p_{xy} = \frac{c(x,y)}{C(x)} \text{ for } \ (x,y) \in E,$$

avec $p_{xy} = 0$ quand x et y ne sont pas connectés (i.e. $(x,y) \notin E$). En notant C la matrice diagonale dont les entrées sont les $C(x)$, on a la relation

$$-\Delta = dcd^\star = C\,(\mathrm{Id} - P)\,.$$

Cette propriété peut être utilisée pour obtenir une expression stochastique de la résistance entre o et Γ. On considère le cas $P \equiv 1$. Le champ p défini précédemment est alors la probabilité de fuite par Γ : pour $x \in V$, $p(x)$ est la probabilité que la marche aléatoire issue de x atteigne Γ avant o.

Proposition 4.10. *On considère une marche aléatoire sur $\mathcal{N} = (V, E, r, o, \Gamma)$ issue de o, avec des probabilités de transition données par* (4.5)*. On a*

$$\frac{1}{R} = C(o)\, p_{esc}, \tag{4.6}$$

où p_{esc} est la probabilité que la marche atteigne Γ avant de revenir en o, et R est la résistance du réseau entre o et Γ (voir Def. 4.5).

Démonstration. Soit p la solution du problème (4.2), avec $P \equiv 1$ sur Γ. Du fait du choix particulier de P, pour tout $x \in V$, $p(x)$ (défini précédemment comme une espérance), est la *probabilité*, partant de x, d'atteindre Γ avant o Par définition 4.5, la résistance R est $1/d(o)$. Par ailleurs on a

$$p_{esc} = \sum_{x\sim o} \pi_{ox} p(x) = \frac{1}{C(o)} \sum_{x\sim o} c(o,x)(p(x) - p(o)) = \frac{1}{C(o)} du(o) = \frac{1}{C(o)} \frac{1}{R},$$

qui donne le résultat. □

On considère la marche aléatoire sur un réseau $\mathcal{N} = (V, E, r)$, dont les probabilités de transition sont définies par (4.5). Partant d'une loi de probabilité p^0 sur la position initiale, on note p^n la loi que suit la position de la particule à l'étape n, définie par

$$p^{n+1}(x) = \sum_{y\sim x} \pi_{yx} p^n(y).$$

Proposition 4.11. *Pour toute fonction φ de $\mathbb{R}^+$ dans $\mathbb{R}$ convexe, la fonctionnelle*

$$S \;:\; p \longmapsto \sum_{x\in V} \varphi\left(\frac{p(x)}{C(x)}\right) C(x)$$

est décroissante le long de la trajectoire discrète, i.e. $S(p^{n+1}) \leq S(p^n)$.

Démonstration. On a

$$S(p^{n+1}) = \sum_{x\in V} \varphi\left(\frac{p^{n+1}(x)}{C(x)}\right) C(x).$$

Chaque terme de la somme s'écrit

$$\varphi\left(\frac{p^{n+1}(x)}{C(x)}\right) C(x) = \varphi\left(\sum_{y\sim x} \frac{c(x,y)}{C(x)} \frac{p^n(y)}{C(y)}\right) C(x) \leq \sum_{y\sim x} \frac{c(x,y)}{C(x)} \varphi\left(\frac{p^n(y)}{C(y)}\right) C(x)$$

car φ est convexe.

On a donc finalement

$$S(p^{n+1}) \leq \sum_{x\in V} \sum_{y\sim x} c(x,y) \varphi\left(\frac{p^n(y)}{C(y)}\right) = \sum_y \left(\frac{p^n(y)}{C(y)}\right) \sum_{x\sim y} c(x,y) = \sum_y \left(\frac{p^n(y)}{C(y)}\right) C(y),$$

ce qui termine la preuve. □

Corollaire 4.12. *En prenant $\varphi(a) = a \log a$, on obtient en particulier la décroissance de l'entropie relative (ou divergence de Kullback-Leibler) de p relativement à la mesure stationnaire C :*

$$S(p) = \sum_{x\in V} \frac{\rho(x)}{C(x)} \log\left(\frac{\rho(x)}{C(x)}\right) C(x) = \sum_{x\in V} \rho(x) \log\left(\frac{\rho(x)}{C(x)}\right).$$

Plan de transport

Etant donnée une distribution de probabilité p^0 définie sur les sommets d'un réseau résistif $\mathcal{N} = (V, E, r)$, ce qui précède revient à définir un plan de transport vers une nouvelle mesure discrète p^1. En effet, avec des notations naturelles, le plan $\gamma \in \mathbb{R}_+^{V\times V}$ défini par

$$\gamma_{yx} = \pi_{yx} p^0(y)\,,\ \pi_{yx} = \frac{c(y,x)}{C(y)}\,,\ C(y) = \sum_x c(y,x)\,,\ c(x,y) = r(x,y)^{-1}$$

transporte p^0 vers p^1 (on a $\gamma = (\gamma_{yx}) \in \Pi_{p^0,p^1}$ avec les notations du début de la section 15, page 118).

Équation de la chaleur sur un réseau

On peut établir une équation d'évolution sur le réseau, en définissant de façon différente la marche aléatoire : on considère que, pour $\tau \in]0,1]$, on reste sur place avec une probabilité $1-\tau$, et l'on se déplace avec probabilité τ, le déplacement se fait alors selon la loi définie par (4.5). On note p_τ^n la loi d'un point évoluant suivant ces principes, on a

$$p_\tau^{n+1}(x) = (1-\tau)p_\tau^n(x) + \tau \sum_{y\sim x} \pi_{yx} p_\tau^n(y),$$

d'où

$$\frac{p_\tau^{n+1}(x) - p_\tau^n(x)}{\tau} = -p_\tau^n(x) + \sum_{y\sim x} \pi_{yx} p_\tau^n(y),$$

soit, en faisant tendre formellement le pas de temps τ vers 0,

$$\frac{dp}{dt}(x) = -p(x) + \sum_{y\sim x} \pi_{yx} p(y) = -(\mathrm{Id} - {}^tK)p.$$

On obtient une structure plus familière en considérant la variable $\rho(x)$ exprimant la densité de p relativement à la mesure stationnaire C (cette mesure stationnaire est de façon évidente la même pour la marche aléatoire initiale, et pour cette nouvelle version alourdie), i.e. $\rho(x) = p(x)/C(x)$. En divisant l'équation précédente par $C(x)$ on obtient

$$\frac{d\rho}{dt}(x) + \rho(x) - \sum_{y\sim x} \pi(x,y)\rho(y),$$

qui peut s'écrire matriciellement

$$\frac{d\rho}{dt} + (\mathrm{Id} - K)\rho = 0. \tag{4.7}$$

Noter que l'on retrouve une matrice symétrique en multipliant l'équation précédente par la matrice diagonale C associée canoniquement à la mesure stationnaire.

4.4 Modèle de flânage

On cherche à modéliser le mouvement d'un individu, ou d'une collection d'individus, dans un lieu d'exposition. On considère le lieu constitué de travées, sur les côtés desquels se trouvent des stands, chaque travée reliant deux nœuds. Chaque nœud corresponds dans l'évolution du promeneur à un point de bifurcation : il va poursuivre son cheminement en emprutant l'une des travées accessibles. On associe à un tel lieu d'exposition un graphe non orienté (V, E), où V est l'ensemble des sommets (nœuds du réseau), et E l'ensemble des côtés (travées), sous ensemble symétrique de $V \times V$.

Évolution pilotée par l'intérêt. On considère chaque travée affectée d'un *score*, qui quantifie l'intérêt du promeneur pour la travée en question. On suppose que le promeneur arrivé au noeud x est capable d'estimer, par vision directe, le score associé aux différentes arêtes issues de x. On définit une marche aléatoire sur le réseau en affectant aux différentes possibilités des probabilités proportionnelles au score, ce qui conduit à la définir la matrice de transition suivante (on écrit $a \sim b$ si $(a, b) \in E$)

$$K(x,y) = \left| \begin{array}{lll} \dfrac{s(x,y)}{\displaystyle\sum_{z \sim x} s(x,z)} & \text{si} & y \sim x \\ 0 & \text{si} & (x,y) \notin E \end{array} \right.$$

On se retrouve donc dans le cadre de la section 4.3, où les conductances sont ici remplacées par des scores, mesurant l'intérêt relatif des différentes travées pour le flâneur. Ce modèle est de façon évidente loin d'être satisfaisant, en particulier le flâneur ainsi modélisé est d'une certaine manière sans mémoire : il est susceptible de revenir sur ces pas, pour revisiter la travée qu'il vient de quitter. Nous décrivons ci-dessous quelques extensions possibles du modèle, de façon à le rendre plus réaliste (au prix d'un éloignement du cadre formel décrit dans la section 4.3).

Extensions.

Le parcours effectif d'une personne dans un tel contexte peut difficilement se concevoir comme un processus purement Markovien, tel que décrit ci-dessus. Il est raisonnable d'intégrer des ingrédients supplémentaires dans le modèle d'évolution, notamment :

1. La probabilité de retourner sur ses pas en arrivant à un point de bifurcation, sauf situation particulière, est très faible.
2. La trajectoire d'un individu a une certain persistence : lorsque l'on arrive à un point de bifurcation, il y a une tendance à continuer tout droit. On peut penser que cette tendance s'amenuise lorsque le nombre de pas dans la même direction devient grand.
3. Les travées qui ont déjà été visitées sont moins attractives.

Une heuristique simple pour gérer ces différents points est la suivante :

On se donne une matrice de scores à l'instant n : $S^n = (s(x,y)) \in \mathbb{R}_+^E$. Partant d'un point x, on récupère les scores de la ligne correspondant à x : $(s(x,y))$. Venant de z, on multiplie le score $s(x,z)$ par un facteur d'inhibition $f_{back} \in [0,1[$. On note n_s le nombre de pas effectués

sans avoir changé de direction. On prend en compte la persistence en multipliant le score de (x, y_s) par un facteur du type

$$f_s = 1 + k \exp\left(-n_s/N_s\right),$$

où N_s est une longueur typique de trajectoire rectiligne avant changement de direction. On calcule ensuite les probabilités de transition en normalisant les scores. Si le sommet suivant est y, on multiplie le score $s(x, y)$ par un facteur d'inhibition $f_m \in [0, 1[$ qui prend en compte la réduction de l'intérêt que l'on accorde à une travée déjà visitée.

4.5 Plongement dans l'espace euclidien

On considère un réseau $\mathcal{N} = (V, E, \Gamma)$ (la racine n'est plus ici distinguée comme un point particulier de la frontière) plongé dans l'espace euclidien $\mathbb{R}^d$, c'est à dire que chaque sommet de V est associé à un point x de $\mathbb{R}^d$, et les côtés sont associés aux sommets entre ces points. On suppose que la correspondance Sommet $\mapsto$ Point est injective, et on suppose que les segments ne se croisent pas [27]. Nous simplifierons les notations en ne faisant pas de distinction entre les sommets du réseau abstrait et les points de $\mathbb{R}^d$ associés. On considère une collection de flux $u \in \mathbb{R}^E$ supposée obéir à la loi de Kirchhof sur les sommets intérieurs. On note $\vec{e}$ la mesure vectorielle associée à l'arête e. Plus précisément, pour tout

$$e = (x, y) \in \mathbb{R}^d \times \mathbb{R}^d\,,\ n_e = \frac{y - x}{|y - x|}$$

on définit la distribution vectorielle (ou mesure vectorielle) $\vec{e}$ comme

$$\varphi \in C_c^\infty(\mathbb{R}^d)^d \longmapsto \langle \vec{e}\,, \varphi \rangle = \int_e \varphi \cdot n.$$

Proposition 4.13. *La mesure vectorielle G définie par*

$$G = \sum_{e \in E} u(e)\, \vec{e} \tag{4.8}$$

vérifie l'équation de conservation (dans $\mathcal{D}'$)

$$\nabla \cdot G = -\sum_{x \in \Gamma} du(x) \delta_x,$$

où la divergence d'une mesure vectorielle est la distribution d'ordre 1 définie par

$$\langle \nabla \cdot G\,, \varphi \rangle = -\langle G\,, \nabla \varphi \rangle \quad \forall \varphi \in \mathcal{D}(\mathbb{R}^d).$$

Démonstration. Pour tout $\varphi \in C_c^\infty$, on a

$$\langle \nabla \cdot G\,, \varphi \rangle = -\langle G\,, \nabla \varphi \rangle = -\sum_{e \in E} u(e) \langle \vec{e}\,, \nabla \varphi \rangle = -\sum_{e \in E} u(e) \int_x^y n_e \cdot \nabla \varphi$$

$$= -\sum_{e \in E} u(e) \int_x^y \partial \varphi / \partial s \, ds = -\sum_{e \in E} u(e) \left(\varphi(y) - \varphi(x)\right) = \sum_{x \in V} \varphi(x) \sum_{y \sim x} u(x, y)$$

$$= -\sum_{x \in V} du(x) \varphi(x) = -\sum_{x \in \Gamma} du(x) \langle \delta_x\,, \varphi \rangle,$$

d'où la propriété annoncée. □

27. Si $d = 2$, le graphe est alors qualifiée de *planaire*.

Remarque 4.14. *Dans le cas où* Γ *se décompose en* Γ_0 *(entrée) et* Γ_1 *(sortie), qui portent respectivement les mesures (positives, de même masse)* μ_0 *et* μ_1*, considérées comme des flux, et auxquelles on associe les mesures atomiques (on garde la même notation)*

$$\mu_0 = \sum_{x\in\Gamma_0} \mu_0(x)\delta_x\,,\ \mu_1 = \sum_{x\in\Gamma_1} \mu_1(x)\delta_x,$$

on peut alors écrire

$$\nabla\cdot G = \mu_0 - \mu_1.$$

4.6 Premier pas vers le transport branché

Le cadre introduit dans la section précédente permet de formaliser une classe très générale de problèmes, qui n'ont été considérés que récemment, et qui suscitent de fait un grand nombre de questions encore ouvertes[28]. On considère deux mesures atomiques μ_0 et μ_1 sur $\mathbb{R}^d$, de supports finis (et disjoints, pour simplifier), de même masse totale (par exemple 1), et l'on note Λ_{μ_0,μ_1} l'ensemble des réseaux (V,E,Γ) plongés dans $\mathbb{R}^d$ (les sommets sont identifiés à des points de $\mathbb{R}^d$, et les arêtes à des segments[29] reliant ces points), tels que $\mathrm{supp}(\mu_0)\cup\mathrm{supp}(\mu_1) = \Gamma$. Pour tout $\mathcal{N}\in\Lambda_{\mu_0,\mu_1}$, tout champ de flux $u\in\mathbb{R}^E$, on note G_u la mesure vectorielle associée à u (on considérera que la notation u encode non seulement le champ des valeurs des flux, mais aussi le réseau $\mathcal{N}$ sur lequel ils sont définis) selon (4.8) (voir section 4.5). On dira que u est admissible, ce qu'on écrira $u\in\Pi_{\mu_0,\mu_1}$, si

$$\nabla\cdot G_u = \mu_0 - \mu_1, \tag{4.9}$$

au sens de la proposition 4.13.

Remarque 4.15. *Il est tentant de dire que* u *transporte* μ_0*, vers* μ_1*. On prendra cependant garde au fait que ce transport est très différent de celui définit dans le cadre du transport optimal (voir section 15). On ne se préoccupe notamment pas ici de savoir "qui va où" : si l'on considère par exemple une bifurcation de mélange (deux arêtes rentrantes* 1 *et* 2 *et une arête sortante), suivie (sur l'arête sortante) par une bifurcation de séparation (deux arêtes sortantes* $1'$ *et* $2'$*), la seule connaissance de* u *ne donne pas d'information sur la proportion dans* $1'$ *de matière venant de* 1*. On verra cependant que, la recherche de réseaux optimaux dans un sens assez général aura tendance à faire disparaître cette ambiguité (le réseau évoqué précédemment comporte un* cycle*, alors que les réseaux optimaux n'en contiendront pas). Par ailleurs,* μ_0 *et* μ_1 *doivent ici être vus comme des flux (quantité de matière par unité de temps) plus que comme des masses statiques. On peut évidemment passer de l'un à l'autre en intégrant l'équation* (4.9) *sur un temps unitaire, mais le problème se pose bien ici nativement en termes de flux.*

Dans le contexte précédemment défini, on définit le coût associé à u de la façon suivante

$$u\in\Pi_{\mu,\nu} \longmapsto C(u) = \sum_e |u(e)|^\alpha\,|e|\,,$$

28. Pour une présentation générale du domaine, voir par exemple :
M. Bernot, V. Caselles, J.-M. Morel, Optimal Transportation Networks, Lecture Notes in Mathematics 1955, Springer Verlag Berlin Heidelberg 2009.

29. En toute généralité, il serait naturel d'identifier les arêtes à des courbes rectifiables, mais on se limitera ici à des segments.

où α est un nombre positif ou nul, et $|e|$ est la longueur de l'arête e.

Le contexte physique d'intensité électrique ou d'écoulement fluide suggère un choix $\alpha = 2$, qui correspondrait à la situation suivante : on considère des sources électriques, et des puits, il s'agit de faire passer une intensité prescrite entre ces puits et ces sources au travers d'un réseau de fils électrique de caractéristique donnée (résistivité prescrite, donc résistance proportionnelle à la longueur), en minimisant la puissance dissipée. Ce problème est dégénéré, comme on peut s'en convaincre en considérant le cas de deux masses de Dirac. En reliant les électrodes ponctuelles par des fils [30] en nombre croissant (en parallèle), on fait diminuer la résistance, et donc la puissance dissipée, l'infimum est ainsi nul, et n'est pas atteint [31]. Le problème pour $\alpha = 2$ (on plus généralement $\alpha > 1$), peut devenir consistant si l'on rajoute des contraintes, par exemple sur la longueur totale du fil, ou si l'on interdit les cycles, mais cette démarche n'a pas été poursuivie à notre connaissance.

Les problèmes de transport branché tels qu'on les conçoit généralement portent sur le cas d'une puissance inférieure à 1, qui exprime une diminution du coût de transport par mutualisation de l'usage des segments (on peut penser à un réseau routier). Le cas $\alpha = 0$ correspond au problème dit de *Steiner*, qui consiste à trouver un réseau reliant tous les points, en minimisant la longueur totale du réseau. Le cas $\alpha = 1$ correspond essentiellement au problème de Monge, pour le coût associé à la distance euclidienne (qui correspond à la distance W_1). Le cas général $\alpha \in]0, 1[$ correspond à un domaine des mathématiques à part entière [32].

4.7 L'arbre bronchique humain comme réseau résistif

Comme modèle simplifié d'arbre bronchique, on considère un arbre régulier à N générations : une première arête (qui correspond à la *trachée*) se sépare en deux branches-filles, et ainsi de suite pour chacune des nouvelles branches, jusqu'à atteindre la génération Nned. la première correspond à la génération 0, de telle sorte que l'arbre comporte en fait $N + 1$ niveaux, et 2^N feuilles. À titre d'ullistration, la figure 4.1 (gauche) représente un arbre à 4 generations. On suppose ici l'arbre *symétrique*, ce qui signifie que la résistance est uniforme sur chaque génération.

La résistance globale de la génération k est $\overline{r}_n = r_n/2^n$, de telle sorte que la résistance globale vaut

$$\overline{R} = \sum_{n=0}^{N} \overline{r}_n = \sum_{n=0}^{N} \frac{r_n}{2^n}. \tag{4.10}$$

Plus précisément, si l'on considère que les bronches d'une même génération n ont la même

30. Le fait que les fils, selon nos hypothèses, doivent être rectilignes, ne pose pas de problème, on peut construire un faisceau de fils distincts, en considérant des trajets affines par morceau.

31. On peut faire un lien avec le fait que la diffusion dans un domaine continu, par exemple d'une source ponctuelle à un puit ponctuel, tend à uniformiser les flux, ce qui correspond d'une certaine manière à une infinité de fils conducteurs en parallèle.

32. Voir : M. Bernot, V. Caselles, J.-M. Morel, Optimal Transportation Networks, Models and Theory, Lecture Notes in Mathematics.

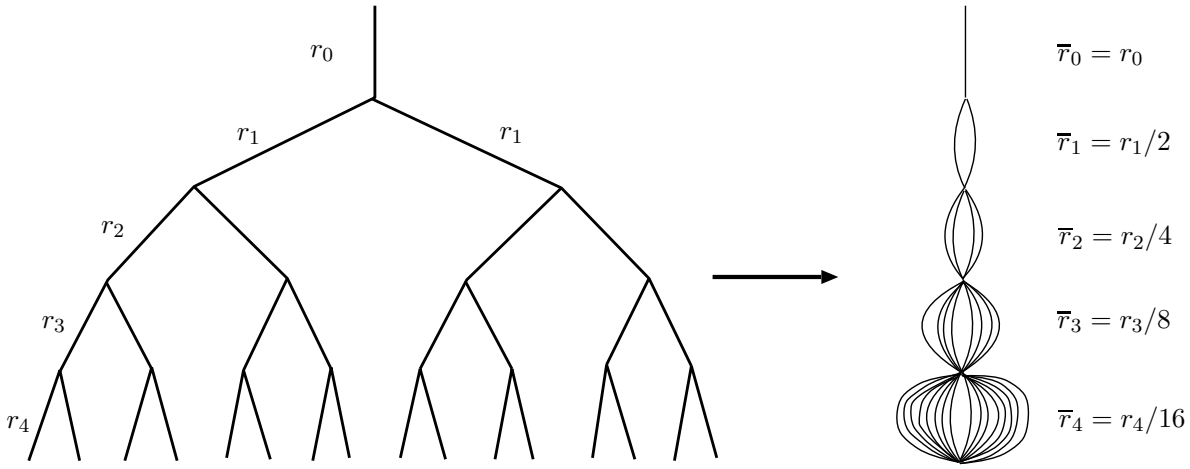


FIGURE 4.1 – Regular dyadic tree

longueur ℓ_n et le même diamètre d_n, la loi de Poiseuille (2.16), précisé l'expression (4.10) :

$$\overline{R} = C \sum_{n=0}^{N} \frac{1}{2^n} \frac{\ell_n}{d_n^4}. \tag{4.11}$$

Si l'on suppose que l'arbre est de plus géométrique, i.e. les dimensions des bronches évoluent géométriquement au fil des générations (paramètre d'homothétie λ d'une génération à la suivante), on a

$$\overline{R} = r_0 \sum_{k=0}^{N} \frac{1}{2^k} \frac{1}{\lambda^{3k}}. \tag{4.12}$$

Remarque 4.16. *Remarquer que cette serie diverge dès que λ est inférieur à $2^{-1/3}$. Selon les données expérimentales, λ est estiùé autour de $0.85 > 2^{-1/3}$ (≈ 0.79), de telle sorte que le poumon "réel" semble se situer dans la zone de convergence.). Mais, pour la même raison, la série des volumes (d'ordre $2^k\lambda^{3k}$ pour la génération k) diverge, de telle telle sorte que le poumon infini extrapolé remplit (très largement, d'une certaine manière, du fait de l'inagalité stricte) l'espace euclidien.*

Exercice 4.1. (Inspiré de Mauroy et al. [33])
On s'intéresse à la résistance équivalente d'un réseau dyadique de tuyaux, du type de celui constitué par l'arbre bronchique humain. On suppose que cet arbre est composé de $N+1$ générations (la première correspondant à la trachée). Si l'on suppose que tous ces tuyaux ont la même forme (identiques à une homothétie près), la résistance à l'écoulement d'un tuyau élémentaire, selon la loi de Poiseuille, est proportionnelle à l'inverse de son volume. On suppose que tous les tuyaux d'une génération p ont le même volume u_p. Sous ces hypothèses, la résistance equivalente R et le volume V ont les expressions qui sont données ci-dessous.

On définit, pour tout $u = (u_0, \dots, u_N)$, $u_p > 0$ pour tout $p \leq N$,

$$R(u) = \sum_{p=0}^{N} \frac{1}{2^p u_p}, \ V(u) = \sum_{p=0}^{N} 2^p u_p.$$

33. B. Mauroy, M. Filoche, E. R. Weibel, B.Sapoval, An optimal bronchial tree may be dangerous, Nature, 427, 633-636, 12 February 2004.

On note $U =]0, +\infty[^{N+1}$, et l'on s'intéresse à la minimisation de la fonction $R(u)$ sur l'ensemble

$$K = \{u = (u_0, u_1, \ldots, u_N) \in U\,,\ V(u) \leq M\}$$

où $M > 0$ est donné (volume maximal : volume de la cage thoracique).

a) Montrer que l'infimum de R sur K est strictement positif, et qu'il est atteint en un point $u \in K$ unique.

b) Écrire la condition d'optimalité associée au problème de minimisation de R sur K, et préciser pourquoi, nécessairement, V(u) = M. Calculer u.

4.8 Réseaux infinis

Nous donnons ici quelques éléments sur l'étude de réseaux infinis, en prolongement direct de ce qui a été vu précédement. On considère un réseau $\mathcal{N} = (V, E, r, o)$, où V est un ensemble dénombrable de sommets, et o un sommet particulier. On supposera que le degré (nombre de voisins) des sommets est uniformément majoré, et que le réseau est connexe. On notera la disparition de Γ dans la définition ci-dessus : l'un des problèmes essentiels dans ce contexte est précisément de déterminer si l'infini (dans un sens à préciser) est susceptible de jouer le rôle de cette frontière Γ. On définit l'espace d'énergie

$$H = \left\{ q \in \mathbb{R}^V\,,\ q(o) = 0\,,\ \sum_e c(x,y)\,|q(y) - q(x)|^2 < +\infty \right\},$$

qui est un espace de Hilbert pour la norme associée canoniquement à la condition d'appartenance, et

$$H_0 = \overline{D},$$

adhérence des champs à support fini dans H.

On peut définir la résistance $R \in]0, +\infty]$ de ce réseau (sous entendu : entre o et l'infini) comme la limite quand N tend vers $+\infty$ de R_N, resistance du sous-réseau des points à distance [34] au plus N de o (avec Γ_N défini comme l'ensemble des sommets à distance exactement N de o).

On énoncera simplement un résultat fondamental [35] établissant un lien entre les espaces fonctionnels ci-dessus, la résistance globale du réseau, et le comportement de la marche aléatoire associée canoniquement au réseau.

Théorème 4.17. *Les trois assertions suivantes sont équivalentes :*

(i) $H/H_0 = \{0\}$ *;*

(ii) $R = +\infty$ *;*

(iii) La marche aléatoire dont les probabilités de transition sont définies par (4.5) *est récurrente.*

34. Il s'agit ici de la distance canonique définie sur le graphe, telle que deux points connectés sont à distance 1.

35. Pour la démonstration, voir par exemple :
P. M. Soardi, Potential Theory on Infinite Networks, Springer-Verlag Berlin and Heidelberg 1994.

On notera que l'équivalence entre (i) et (ii) est une généralisation de la proposition 8.1, page 88, qui se limitait au cas d'un réseau linéaire infini dans une direction.

4.9 Réseaux dynamiques

Des chercheurs japonais [36] ont récemment mis en évidence la capacité de certaines moisissures à constituer des réseaux de transport de nourriture qui présentent à la fois une certaine forme d'optimalité globale et une grande robustesse (vis-à-vis par exemple de la disparition brusque d'une branche). Ils ont proposé un modèle dynamique d'évolution d'un réseau existant basé sur les principes suivants. Le point de départ est un réseau résistif, qui réalise le transport d'un flux entre des points-sources et des points-puits, que l'on définit comme Γ_0 et Γ_1, sous-ensemble de l'ensemble des sommets V. On note $\mu_i \in \mathbb{R}^{\Gamma_i}$, $i = 0, 1$, les flux correspondants (tous deux identifiés à des mesures positives).

La loi des nœuds est vérifiée en tout point intérieur au réseau, et le flux au travers d'un côté est régi par une loi de type Ohm (ou Poiseuille)

$$u(x, y) = \frac{D}{L}(p(x) - p(y)),$$

où L est la longueur de l'arête, et D une mesure de sa conductivité [37]. Pour un réseau donné, avec sa collection de conductivités D_{ij}, et une collection de flux d'entrée et de sortie prescrits, on peut calculer les pressions et flux au travers des arêtes en résolvant un problème de Darcy discret avec condition de flux imposé

$$\left\{ \begin{array}{lcrll} u + cd^\star p & = & 0 & \text{sur} & E, \\ du & = & 0 & \text{sur} & \mathring{V}, \\ du & = & -\mu_0 & \text{sur} & \Gamma_0 \\ du & = & \mu_1 & \text{sur} & \Gamma_1 \end{array} \right. \tag{4.13}$$

Noter que, avec des notations évidentes, on peut regrouper les trois dernières équations en

$$du = -\mu_0 + \mu_1 \quad \text{sur} \quad \Gamma.$$

On peut éliminer les flux pour se ramener à un problème de Poisson sur la pression

$$dcd^\star p(x) = \mu_0 - \mu_1 \quad \text{sur} \quad V.$$

Remarque 4.18. *On notera l'absence de conditions aux limites dans le problème ci-dessus. On peut retrouver une analogie avec un problème aux limites sous forme standard en distinguant les points intérieurs des points sur Γ_0 et Γ_1. On écrira alors que la fonction est harmonique sur les points intérieurs, et vérifient sur les bords des conditions de type Neuman :*

$$du(x) = -dcd^\star p(x) = -\mu_0 \ \textit{sur} \ \Gamma_0,$$

36. A. Tero, S. Takagi, T. Saigusa, K. Ito, D. P. Bebber, M. D. Fricker, K. Yumiki, R. Kobayashi, T. Nakagaki, Rules for Biologically Inspired Adaptive Network Design, SCIENCE, Vol. 327, 2010.
`https://dl.dropboxusercontent.com/u/44213852/BIO.OptNetworkYeast.pdf`

37. Pour un écoulement fluide au travers de tuyaux à section circulaire, D représenterait le diamètre à la puissance 4, voir l'équation (2.16), page 35.

mais comme on le voit, dans le cadre discret, ce choix ne fait que compliquer l'écriture. En fait, dans le contexte discret, la frontière étant un sous ensemble de points de même nature que les points intérieurs, on peut considérer que les conditions aux limites de Neuman n'ont lieu d'être considérées, puisque tout problème à flux imposé sur la "frontière" peut s'écrire comme un problème de Poisson sur le domaine entier (les termes de flux passent dans le second membre du problème de Poisson).

Remarque 4.19. *Comme dans le cas du problème de Neuman dans un domain euclidien, la pression est définie à une constante additive près.*

On choisit alors de faire évoluer les conductivités en favorisant les arêtes les plus actives :

$$\frac{dD_{xy}}{dt} = G(|u(x,y)|) - D_{xy},$$

où $G(\,\cdot\,)$ est une fonction croissante, nulle en 0. Les auteurs considèrent par exemple des fonctions du type

$$G(q) = \frac{q^\gamma}{1 + aq^\gamma}.$$

5 Trafic routier ou piéton – macro – 1d – ordre 1 en temps

Cette section donne, sous une forme très préliminaire, quelques éléments de modélisation du trafic routier ou piétons selon une description macroscopique (densité linéique diffuse).

5.1 Modèle d'évolution

On considère l'évolution d'une population de piétons ou de véhicules sur une voie rectiligne, population représentée par une densité linéique $\rho(x,t)$. On considère que la vitesse des entités est fonction de la densité : $v = v(\rho)$. La manière la plus simple de prendre en compte le fait que la vitesse est d'autant plus faible que la densité est importante est $v(\rho) = U(1 - \rho/\rho_{\max})$. La conservation de la masse s'écrit alors (voir section 1)

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}\left(\rho v(\rho)\right) = 0,$$

qui a la forme d'une équation de conservation que l'on peut écrire sous forme générale

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x} f(\rho) = 0, \tag{5.1}$$

où f est le *flux*.

Propagations des perturbations

Si l'on considère une solution stationnaire ρ_e de l'équation, et une solution perturbée $\rho_e + \tilde{\rho}$, on obtient formellement une équation de transport sur la perturbation :

$$\partial_t \tilde{\rho} + f'(\rho_e)\partial_x \tilde{\rho} = 0 \tag{5.2}$$

qui exprime que les perturbations sont transportées à la vitesse $f'(\rho_e)$.

Supposons que $\rho(x,t)$ est une solution régulière de cette équation. On appelle courbes caractéristiques une courbe $t \longmapsto x(t)$ telle que

$$\dot{x}(t) = f'(\rho(x(t),t).$$

On vérifie immédiatement que ρ est constant le long de telles courbes :

$$\frac{d}{dt}\rho(x(t),t) = \partial_t \rho(x(t),t) + \dot{x}(t)\partial_x \rho(x(t),t) = \partial_t \rho(x(t),t) + f'(\rho(x(t),t)\partial_x \rho(x(t),t) = 0.$$

Comme ρ est constant le long de la trajectoire, la vitesse elle-même est constante : les trajectoires sont des droites

$$t \longmapsto x + t f'(\rho_0(x)).$$

Si l'on se donne une densité initiale ρ_0, on peut ainsi construire la solution associée en reportant la valeur de densité initiale le long des caractéristiques. Cette démarche n'est évidemment possible que tant que les caractéristiques ne se croisent pas.

Pour une densité initiale donnée, supposée lisse (continûment différentiable), on peut considérer le flot associé aux caractéristiques

$$\Phi_t \ : \ x \longmapsto x + f'(\rho(x_0, 0))t.$$

Si l'on suppose que la fonction f est C^2, on peut calculer le jabobien de la transformation

$$J(t, x) = 1 + t\, f''(\rho_0(x))\, \rho_0'(x).$$

Ce Jacobien reste > 0 (la transformation est un difféomorphisme, i.e. les trajectoires ne se croisent pas) pour tout t si $f''(\rho_0(x))\, \rho_0'(x) \geq 0$. Si en revanche cette dernière quantité est négative, alors l'application ne sera régulière que pour

$$t < -\frac{1}{f''(\rho_0(x))\, \rho_0'(x)}.$$

Le temps de vie de la solution lisse sera donc

$$T = \frac{1}{\max |(f''(\rho_0(x))\, \rho_0'(x))_-|}$$

(inverse du max de la partie négative de $f''(\rho_0(x))\, \rho_0'(x)$).

Si l'on considère le flux indiqué précédemment $f(\rho) = U\rho(1 - \rho/\rho_{\max})$, on a $f''(\rho) = -2U/\rho_{\max} < 0$. On aura donc existence de solution lisse si ρ_0 est décroissante, et croisement de caractéristique en temps fini si en revanche ρ_0 est croissante.

Remarque 5.1. *On prendra garde au fait que, bien que l'on ait considéré le Jacobien de l'application Φ_t, ce qui suggère un transport de mesure, n'est aucunement associée à un quelconque transport conservatif de masse.*

Lien avec le modèle microscopique. On peut faire un lien formel avec le modèle microscopique présenté dans la section 6, en notant que la densité linéique (nombre de véhicules ou de piétons par mètre) est l'inverse de la distance entre les personnes : $\rho = 1/d$. Si l'on reprend la fonction φ qui définit la vitesse comme fonction de la distance, on a

$$f(\rho) = \rho v(\rho) = \rho\varphi\left(\frac{1}{\rho}\right) \ , \ f'(\rho) = \varphi\left(\frac{1}{\rho}\right) - \frac{1}{\rho}\varphi'\left(\frac{1}{\rho}\right).$$

qui, exprimée en distance locale $u_e = 1/\rho_e$, donne

$$f'(\rho) = \varphi(u_e) - u_e\varphi'(u_e).$$

Si l'on s'intéresse à l'évolution d'une perturbation autour d'une densité uniforme ρ_e, l'équation (5.2), exprime un transport à la vitesse $f'(\rho_e)$. On retrouve au niveau macroscopique la vitesse de propagation vers l'amont $-u_e\varphi'(u_e)$ trouvée dans la section 6. La vitesse macroscopique contient nativement le terme de vitesse des entités $\varphi(u_e)$, puisqu'il s'agit d'une description *Eulérienne* (la variable est exprimée dans le référentiel fixe du laboratoire, selon l'expression consacrée), par opposition à la description macroscopique qui est nativement *Lagrangienne* (les variables sont afférentes aux entités en mouvement).

Remarque 5.2. *Il est immédiat dans le cadre microscopique Lagrangien de prendre en compte des comportements différents selon les entités. C'est beaucoup plus délicat dans le cadre macroscopique Eulérien que nosu considérons ici. Prendre en compte une telle différentiation nécessiterait de faire dépendre dépendre la fonction flux d'un* label *a qui fait référence à une entité particulière. Le système s'écrit alors*

$$\partial_t \rho + \partial_x f_a(\rho) = 0,$$

où $a(x,t)$ permet de suivre *les entités, i.e. obéit à une équation de transport non conservatif (c'est une quantité intensive, du type* information*, qui est propagée) :*

$$\partial_t a + u\,\partial_x\, a = 0.$$

Remarque 5.3. *Dans le même esprit que la remarque précédente, si l'on souhaite prendre en compte un retard (dans l'esprit de la question 6.4, page 75 pour le modèle microscopique), il est important de modéliser le fait que la relaxation d'une distance subjective (sur laquelle l'entité base sa vitesse) vers la distance réelle est un processus essentiellement Lagrangien. Si l'on appelle ($\tilde\rho(x,t)$ la densité subjective de l'entité située en x au temps t, on écrira*

$$\begin{aligned}\partial_t \rho + \partial_x\left(\rho v(\tilde\rho)\right) &= 0\\ \partial_t \tilde\rho + v(\tilde\rho)\,\partial_x \tilde\rho &= \frac{1}{\tau}\left(\rho - \tilde\rho\right).\end{aligned}$$

5.2 Solutions faibles

Les considérations précédentes indiquent qu'il ne peut, en général, exister de solution lisse globale. Pour donner un sens aux solutions non lisses qui semblent naître spontanément, on définit la notion de solution faible :

Definition 5.4. *On dit que $\rho(x,t)$ est une solution faible de* (5.1) *(sur $\mathbb{R}\times]T_1,T_2[$) si, pour toute fonction φ régulière à support compact dans $\mathbb{R}\times]T_1,T_2[$, on a*

$$\int_{\mathbb{R}}\int_{T_1}^{T_2} \partial_t\varphi\,\rho(x,t)\,dx\,dt + \int_{\mathbb{R}}\int_{T_1}^{T_2} \partial_x\varphi\, f(\rho(x,t))\,dx\,dt = 0.$$

On peut intégrer une condition initiale à cette définition. Dans le cas $T_1 = 0$, $T_2 = T$, on écrira

$$\int_{\mathbb{R}}\int_{0}^{T} \partial_t\varphi\,\rho(x,t)\,dx\,dt + \int_{\mathbb{R}}\int_{0}^{T} \partial_x\varphi\, f(\rho(x,t))\,dx\,dt + \int_{\mathbb{R}} \varphi(x,0)\,\rho^0(x)\,dx\,dt = 0$$

pour toute fonction φ régulière à support compact dans $\mathbb{R}\times[0,T[$

On vérifie immédiatement que toute solution régulière est solution faible. Mais cette définition peut s'appliquer à des solutions qui ne sont pas régulières. Considérons par exemple deux densités qui réalisent le même flux : $F = f(\rho_-) = f(\rho_+)$. La densité

$$\rho = \rho_-\mathbb{1}_{]-\infty,0[} + \rho_+\mathbb{1}_{]0,+\infty[}$$

est solution faible stationnaire de (5.1), de même que la densité obtenue en intervertissant ρ_- et ρ_+. On peut construire des solutions non stationnaires de la façon suivante : on se donne deux densités ρ_L et ρ_R, et l'on cherche une solution ρ constante de part et d'autre d'un point de discontinuité $s(t)$ variable en temps. On vérifie qu'une telle densité est solution faible dès que s vérifie une condition dite de *Rankine-Hugoniot*, comme l'exprime la

Proposition 5.5. *(Relation de Rankine-Hugoniot)*
Soient ρ_L et ρ_R deux valeurs entre 0 et $\rho_{\max}$, et $f(\cdot)$ une fonction flux continue. La densité

$$\rho = \rho_L \mathbb{1}_{]-\infty,s(t)[} + \rho_R \mathbb{1}_{]s(t),+\infty[}$$

est solution faible de (5.1) *si et seulement si la discontinuité s progresse à la vitesse constante*

$$\dot{s} = \frac{f(\rho_L) - f(\rho_R)}{\rho_L - \rho_R}. \tag{5.3}$$

Démonstration. On utilise la définition d'une solution faible, en écrivant la première intégrale double

$$\int_{\mathbb{R}} \int_0^{+\infty} \partial_t \varphi \, \rho = \int_0^{+\infty} \left(\rho_L \int_{-\infty}^{s(t)} \partial_t \varphi + \rho_R \int_{s(t)}^{+\infty} \partial_t \varphi \right),$$

avec

$$\int_{-\infty}^{s(t)} \partial_t \varphi = \frac{d}{dt} \left(\int_{-\infty}^{s(t)} \varphi \right) - \dot{s}(t) \varphi(s(t), t), \quad \int_{s(t)}^{+\infty} \partial_t \varphi = \frac{d}{dt} \left(\int_{s(t)}^{+\infty} \varphi \right) + \dot{s}(t) \varphi(s(t), t).$$

La seconde intégrale double (avec la dérivée en espace sur la fonction test s'écrit

$$\int_{\mathbb{R}} \int_0^{+\infty} \partial_x \varphi \, f(\rho(x,t)) = \int_0^{+\infty} \left(f(\rho_L) \int_{-\infty}^{s(t)} \partial_x \varphi + f(\rho_R) \int_{s(t)}^{+\infty} \partial_x \varphi \right) = \int_0^{+\infty} \varphi(s(t),t))(f(\rho_L) - f(\rho_R)).$$

On obtient donc finalement

$$\int_0^{+\infty} \varphi(s(t),t)) \left(-\dot{s}(t)(\rho_L - \rho_R) + f(\rho_L) - f(\rho_R) \right),$$

qui est identiquement nul pour toute fonction test φ si et seulement si la condition (5.3) est identiquement vérifiée. □

Remarque 5.6. *On peut retrouver la relation* (5.3) *en écrivant simplement un bilan de masse au voisinage de la discontinuité.*

On peut vérifier que, sous sa forme faible, l'équation n'est pas bien posée, au sens où elle admet en général plusieurs solutions. La théorie complète de telles équation dépasse le cadre de ce cours sous sa forme actuelle, disons simplement ici qu'il est possible d'imposer à la solution considérer de vérifier un critère supplémentaire, dit *d'entropie*, qui permet de sélectionner *la* solution physique[38] parmi les nombreuses possibles. Ce critère n'est pertinent que pour discriminer des solutions qui présentent des discontinuités, on peut montrer que ces solutions acceptables sont telles que, lorsque la solution présente une discontinuité, les courbes caractéristiques doivent arriver vers la discontinuité, et non pas en partir.

38. Ce type de critère a été élaboré dans le cadre de la dynamique des gaz. Précisons que, dans le cadre du transport d'entités vivantes, sa légitimité est moins nette

5.3 Résolution numérique

On se place sur l'intervalle $]0, L[$ avec des conditions périodiques. La méthode des volumes finis est basée sur une représentation de la densité par une fonction constante par morceaux sur des *cellules* disjointes qui recouvrent le domaine spatial. Nous considérons ici des cellules associées à à une subdivision uniforme de l'intervalle, de pas Δx. On introduit de la même manière une discrétisation en temps $0 < \Delta t < 2\Delta t < \cdots < N\Delta t = T$. On note ρ_i^n la valeur de la densité approchée sur la cellule i, sur l'intervalle de temps $]n\Delta t, (n+1)\Delta t[$. Le schéma résulte de l'intégration de l'équation de conservation sur la cellule C_i et l'intervalle de temps $[t^n, t^{n+1}]$:

$$\int_{C_i} \rho(x, t^{n+1})\, dx - \int_{C_i} \rho(x, t^n)\, dx + \int_{t^n}^{t^{n+1}} \Big(f(\rho(x_{i+1/2}, t) - f(\rho(x_{i+1/2}, t) \Big) = 0,$$

qui conduit à une classe générale de schémas que l'on note

$$\rho_i^{n+1} - \rho_i^n + \frac{\Delta t}{\Delta x} \Big(f_{i+1/2} - f_{i-1/2} \Big) = 0.$$

La stratégie numérique repose sur la définition des flux discrets $f_{i+1/2}$ et $f_{i-1/2}$. Nous nous limiterons ici à des schéma explicites, basé sur la définition du flux discret comme fonctions des densités de part et d'autre de l'interface :

$$f_{i+1/2} = F(\rho_i^n, \rho_{i+1}^n).$$

6 Trafic routier ou piéton – micro – 1d – ordre 1 en temps

6.1 Le modèle

Le modèle dit *Follow the Leader*[39] est basé sur les principes suivants : on considère $n+1$ véhicules se déplaçant sur une route rectiligne (ou piétons se déplaçant sur une même file), et l'on repère leurs positions respectives au temps t par

$$x_1(t) < x_2(t) < \cdots < x_{n+1}(t). \tag{6.1}$$

On considère dans un premier temps que la vitesse du véhicule i ne dépend que de la distance au véhicule précédent, c'est-a-dire $x_{i+1} - x_i$ (on ne prend pas en compte la taille de l'entité). Le système s'écrit alors

$$\dot{x}_i = \varphi(x_{i+1} - x_i) \quad 1 \leq i \leq n. \tag{6.2}$$

Il est naturel de prendre pour φ une fonction qui s'annule en 0, qui prend la valeur U de la vitesse maximale autorisée quand la distance tend vers l'infini. On pourra considérer par exemple la fonction

$$\varphi(u) = U(1 - \exp(-u/u_s)),$$

où u_s est une distance caractéristique de sécurité (distance observée pour des véhicules roulant approximativement aux 2/3 de la vitesse autorisée, pour le cas de voitures sur l'autoroute). Cette quantité conditionne la raideur (*stiffness* en anglais) du modèle.

Remarque 6.1. *La taille des entités peut être prise en compte en modifiant la fonction :*

$$\varphi(u) = U(1 - \exp(-(u - u_m)/u_s)).$$

Noter que cette modification ne change pas la nature du modèle. En dimension 1*, il est en effet équivalent de travailler sur des entités ponctuelles interagissant en fonction de leurs distances, ou des entités de tailles non nulles (en considérant alors les distances d'objet à objet). Cette prise en compte devient en revanche importante dès que l'on s'intéresse au positionnement des entités sur un voie réelle, par exemple si l'on s'intéresse à la possibilité que l'information remonte une file plus vite qu'elle n'avance, où si l'on souhaite faire le lien avec un modèle macroscopique (pour lequel on aura une densité maximale* $1/u_m$*).*

Proposition 6.2. *On se donne des positions initiales vérifiant la relation d'ordre* (6.1)*. On suppose que la vitesse* $V(t)$ *(et donc la trajectoire) de l'entité de tête* $(n+1)$ *est une fonction continue du temps, donnée, à valeur dans* $[0, U]$*. On se donne une fonction de comportement* φ *Lipshitzienne nulle en* 0 *(prolongée par* 0 *en deça), et prenant ses valeurs dans l'intervalle* $[0, U]$*. Le système* (6.2) *admet une unique solution maximale, qui est globale.*

39. C'est sous cette dénomination qu'il est présenté dans :
B. Argall, E. Cheleshkin, J. M. Greenberg, C. Hinde and P.-J. Lin, A rigorous treatment of a follow-the-leader traffic model with traffic lights present, SIAM J. Appl. Math., 63(1), pp. 149–168 , 2002, `http://www.cs.cmu.edu/~bargall/docs/02siam-argall.pdf`.
Cette dénomination est cependant partiellement impropre dans le cas qui nous intéresse : chaque entité suit de fait l'entité qui la précéde, mais la présence de cette dernière est plus une gêne (qui conduit à une diminution de la vitesse) qu'une incitation positive.

Démonstration. Il s'agit d'une application du théorème de Cauchy-Lipschitz 21.9. Cette solution est globale car la vitesse est bornée (proposition 21.12). □

Il est essentiel de vérifier la viabilité de la solution de l'équation différentielle ci-dessus (nous n'avons pas exclu les cas de distances nulles, voire négatives, entre entités. On peut vérifier que les distances restent strictement positives.

Proposition 6.3. *On se place dans les hypothèses de la proposition précédente. Les distances restent strictement positives.*

Démonstration. On note $L = \|\varphi'\|_\infty$. Tant que $x_{n+1} - x_n > 0$, on a

$$\dot{x}_n = \varphi(x_{n+1} - x_n) \leq L(x_{n+1} - x_n),$$

d'où, si l'on note $u_n = x_{n+1} - x_n$,

$$\dot{u}_n \geq -\eta u_n + V(t) \geq -\eta u_n,$$

d'où $u_n \geq u_n(0)e^{-Lt}$. □

Remarque 6.4. *Le caractère Lipschtiz de φ est essentiel pour éviter les accidents. Si l'on prend par exemple une fonction φ qui se comporte comme u^α au voisinage de 0, avec $\alpha \in]0,1[$, considérant deux véhicules successifs, le premier étant arrêté, on obtient l'équation $\dot{u} = -u^\alpha$, qui conduit à*

$$u(t) = \left(u(0)^{1-\alpha} - (1-\alpha)t\right)^{1/1-\alpha}.$$

On a alors "accident", c'est à dire annulation des distances en temps fini. Noter que le théorème de Cauchy Lipschitz ne s'applique ici que sur l'ouvert $]0,+\infty[$, la solution maximale n'est alors pas globale.

6.2 Stabilité, propagation des perturbations

Supposons que le véhicule de tête en x_{n+1} se maintient à une vitesse constante $V < U$. On vérifie immédiatement que si tous les véhicules sont à distance u_e du précédent, avec $V = \varphi(u_e)$, autrement dit

$$u_e = -u_s \ln\left(1 - \frac{V}{U}\right),$$

ils vont tous à la vitesse V du véhicule de tête. On peut se demander ce qui va se passer en cas de perturbation, par exemple si le véhicule de tête freine brusquement, puis reprend sa vitesse de croisière V.

Remarque 6.5. *Si l'on note $V = \{1, 2, \ldots, n\}$, on peut définir un ensemble A d'arêtes :*

$$(1,2)\,,\ \ldots, (n-1,n),$$

tel que $(i,j) \in A$ si et seulement si le comportement de i est directement influencé par le comportement de j. Pour le modèle considéré, le graphe est de façon évidente acyclique *(voir def. 12.3).*

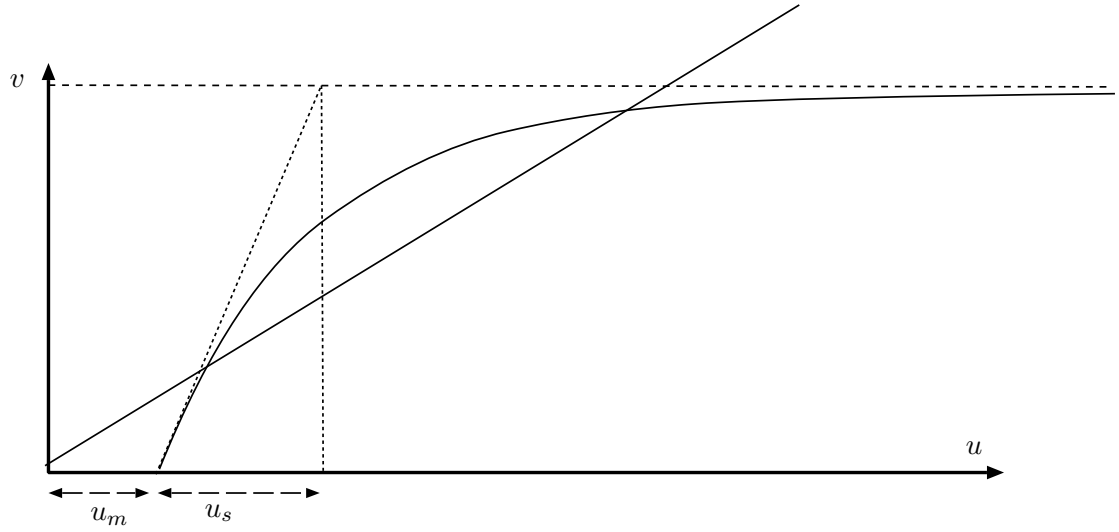


FIGURE 6.1 – Vitesse fonction de la distance

On introduit les variables de distances entre véhicules :

$$u_i = x_{i+1} - x_i\,,\ i = 1, \dots n.$$

Le système s'écrit, pour ces nouvelles variables

$$\dot{u}_i = \varphi(u_{i+1}) - \varphi(u_i)\,,\ i = 1, \dots n\,,\ \text{ ou } \dot{u} = F(u).$$

et $u = (u_e, \dots, u_e)$ est point d'équilibre du système.

Proposition 6.6. *Le point d'équilibre défini ci-dessus est asymptotiquement stable.*

Démonstration. Le linéarisé au point d'équilibre s'écrit

$$\nabla F = \varphi'(u_e) \begin{pmatrix} -1 & 1 & 0 & \cdot & 0 \\ 0 & -1 & 1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & -1 & 1 \\ 0 & \cdot & \cdot & 0 & -1 \end{pmatrix}.$$

On a donc une unique valeur propre $-\varphi'(u_e) < 0$, donc stabilité asymptotique avec un temps caractéristique de retour à l'équilibre[40] égal à $1/\varphi'(u_e)$. □

40. Nous verrons que dans le cas présent d'un gradient non diagonalisable, le temps effectif caractéristique de retour à l'équilibre peut être significativement plus grand que $1/\varphi'(u_e)$, ou plus précisément que le temps de retour effectif à l'équilibre n'est pas uniforme vis-à-vis du nombre n de véhicules, alors que $1/\varphi'(u_e)$ n'en dépend pas.

Propagation des perturbations vers l'amont

Équation de transport. On peut établir un lien informel entre le comportement du système au voisinage de l'équilibre et une équation de transport. Cette approche va nous permettre d'estimer la vitesse de propagation de l'information le long du train de véhicule, une approche plus rigoureuse pour estimer cette vitesse est décrite plus loin.

Considérons une perturbation de l'état d'équilibre correspondant à des entités équidistance de u_e, qui avancent à la vitesse $v_e = \varphi(u_e)$. En se plaçant dans le référentiel qui suit le train, à la vitesse u_e, on peut décrire les petites évolutions du modèle en considérant que les distances sont du type $u_e + w_i$, où w_i est une petite variation de la distance entre x_i et x_{i+1}, que l'on considère comme une variable attachée au milieu du segment (qui est fixe dans le référentiel mobile). On a

$$\dot{w_i} = \varphi(u_e + w_{i+1}) - \varphi(u_e + w_i) \approx \varphi'(u_e)(w_{i+1} - w_i) = u_e\varphi'(u_e)\frac{w_{i+1} - w_i}{u_e}.$$

Les w_i étant définis en des points distants de u_e, on peut interpréter le dernier quotient comme une dérivée en espace d'une fonction $w(x)$, pour laquelle obtient ainsi formellement l'équation

$$\frac{\partial w}{\partial t} - u_e\varphi'(u_e)\frac{\partial w}{\partial x} = 0.$$

il s'agit d'une équation de transport à la célérité $c = -u_e\varphi'(u_e)$. On a donc une remontée à vitesse constante vers l'arrière du train. Cette vitesse est estimée dans le référentiel qui avance à la vitesse $\varphi_(u_e)$. On aura effectivement propagation vers l'arrière[41] (pour l'observateur extérieur) si

$$u_e\varphi'(u_e) > \varphi(u_e) \Longleftrightarrow \varphi'(u_e) > \frac{\varphi(u_e)}{u_e}.$$

Sous cette dernière forme, il apparaît que le critère se ramène à une comparaison entre les pentes de la corde et de la tangente au point considéré $(u_e, \varphi(u_e))$. On note que, pour un même flux, c'est à dire pour une même corde (le flux d'entités par unité de temps est $\varphi(u_e)/u_e$), on a deux régimes de fonctionnement possibles (voir figure 6.1), l'un dense à faible vitesse (régime fluvial), et l'autre dilué à grande vitesse (régime torrentiel). On a de façon évidente propagation de l'information vers l'arrière pour le cas dense. Dans le cas dilué, pour un même flux, la vitesse de propagation est inférieure à la vitesse des véhicules, de sorte qu'une perturbation suit le sens du mouvement pour un observateur extérieur.

Analyse spectrale. Cette propagation vers l'amont décrite informellement ci-dessus peut-être étayée par une étude plus approfondie du système tangent au voisinage du point d'équilibre :

$$\dot{u} = Mu,$$

où M est la matrice du gradient de F au point d'équilibre

On garde la notation u pour désigner le vecteur inconnu, mais les u_i correspondent maintenant à des variations autour du point d'équilibre, qui évoluent au voisinage de 0 (et non pas de u_e).

41. Dans le cas du trafic routier, si l'on est dans cette situation, toute perturbation est susceptible de se propager vers l'arrière et de créer potentiellement un bouchon.

La solution du problème ci-dessus s'écrit

$$u(t) = e^{tM} u_0,$$

où u_0 est une perturbation initiale. La matrice M s'écrit

$$M = \beta\,(-\,\mathrm{Id} + N)$$

avec $\beta = \varphi'(u_e)$, et N une matrice nilpotente

$$N = \begin{pmatrix} 0 & 1 & 0 & \cdot & 0 \\ 0 & 0 & 1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & 0 & 1 \\ 0 & \cdot & \cdot & 0 & 0 \end{pmatrix},\ N^2 = \begin{pmatrix} 0 & 0 & 1 & \cdot & 0 \\ 0 & 0 & 0 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 1 \\ \cdot & \cdot & \cdot & 0 & 0 \\ 0 & \cdot & \cdot & 0 & 0 \end{pmatrix},\ \ldots, N^n = 0.$$

L'exponentielle s'écrit donc

$$e^{tM} = e^{-\beta t}\left(\mathrm{Id} + \beta t N + \frac{(\beta t)^2}{2!}N^2 + \frac{(\beta t)^3}{3!}N^3 + \cdots + \frac{(\beta t)^{n-1}}{(n-1)!}N^{n-1}\right).$$

Montrons que la forme particulière de cette matrice rend compte d'une propagation des perturbations vers les index de vehicules décroissants. On considère pour cela une perturbation du véhicule de tête, qui induit une perturbation du véhicule immédiatement derrière celui-ci. Cette perturbation est donc colinéaire à $u_0 = e_n$, où e_i est le $i-$ème vecteur de la base canonique de $\mathbb{R}^n$. On a

$$Ne_n = e_{n-1}, N^2 e_n = e_{n-2}, \ldots, N^{n-1} e_n = e_1.$$

Le comportement général de la solution du système linéarisé peut donc se traduire en termes de perturbations pour chacun des véhicules de la file, avec, pour le véhicule k, un facteur

$$\frac{(\beta t)^{n-k}}{(n-k)!}e^{-\beta t},\ k = 1, \ldots, n.$$

Dans les premiers instants, cette fonction va avoir un maximum glissant qui correspond au véhicule couramment affecté par la perturbation. On peut par exemple calculer pour quel temps deux véhicules successifs sont affectés de la même manière (ce qui correspond au passage de l'information entre $k+1$ et k) :

$$\frac{(\beta t)^{n-k-1}}{(n-k-1)!}e^{-\beta t} = \frac{(\beta t)^{n-k}}{(n-k)!}e^{-\beta t} \Leftrightarrow t = (n-k)/\beta.$$

La distance entre les véhicules étant de l'ordre de u_e, cela traduit une propagation de l'information vers l'amont du train de véhicule à la célérité

$$c = -\beta u_e = -u_e \varphi'(u_e).$$

On peut retrouver ce résultat en recherchant à quel moment la perturbation ressentie par l'entité $n-k$ est maximale. On a

$$p_k(t) = e^{-\beta t}\frac{(\beta t)^k}{k!},\ p_k'(t) = e^{-\beta k}\frac{\beta^k t^{k-1}}{k!}\,(-\beta t + k)$$

qui s'annule pour $t = k/\beta$.

Question 6.1. Montrer que (le maximum de) l'intensité de la perturbation ressentie par l'entité $n-k$ varie pour k grand comme $1/\sqrt{2\pi k}$.

Exercice 6.2. Montrer que la prise en compte de la taille des véhicules (en considérant que la fonction φ est nulle en dessous d'une longueur minimale u_s, et concave sur $[u_s, +\infty[$) permet de mettre en évidence la possibilité que des ondes d'information remontent le courant vers l'amont plus vite que la vitesse des véhicules-mêmes.

Remarque 6.7. *Pour appréhender ce qui se passe lorsque le nombre de véhicules est important, on considère une file de véhicule infinie dans une direction : une infinité de véhicule suit un véhicule de tête dont la vitesse est fixée. La perturbation au temps t correspond à la loi de Poisson de paramètre βt :*

$$p(t) = (p_k(t))_{k\in\mathbb{N}}\,,\ p_k = e^{-\beta t}\frac{(\beta t)^k}{k!}$$

On a donc $\|p(t)\|_1 = 1$: la "masse" totale de la perturbation reste constante, on n'a donc pas, pour cette norme, stabilité asymptotique.

On a en revanche décroissance vers 0 des normes p, avec $p > 1$, jusqu'à $p = \infty$. On a convergence vers 0 dans ℓ^∞ faible-$\star$ (contre toute suite de ℓ^1), on n'a en revanche pas convergence faible-$\star$ vers 0 dans ℓ^1 vu comme sous espace de $(\ell^\infty)'$ (qui correspondrait pour des mesures sur un espace euclidien à la convergence étroite*). La non-convergence de la suite (comme de toute suite extraite) n'est pas en contradiction avec la compacité de la boule unité de $(\ell^\infty)'$ pour la topologie faible-$\star$, du fait de la non séparabilité de ℓ^∞ (on pourra se reporter à la section 13, page 112, pour plus de détail). Cette convergence est une version discrète de la convergence étroite pour les mesures, on retrouve ici la situation typique d'une famille de mesures de probabilité qui part vers l'infini (ou se concentre sur le bord d'un ouvert), ce qui assure la convergence vers 0 au sens des mesures (i.e. contre les fonctions continues qui s'annulent au bord), sans que l'on ait convergence étroite.*

6.3 Cas périodique

On se place dans un cadre périodique : route de type périphérique sans entrée ni sortie, ou couloir circulaire, représenté par un domaine périodique de longueur L. Le véhicule n voit le véhicule 1, et les équations s'écrivent simplement

$$\dot{x}_i = \varphi(x_{i+1} - x_i)\,,\ i = 1, \dots, n \quad (n+1 \equiv 1),$$

ou, exprimé sur les variables de distance $u_i = x_{i+1} - x_i$ (avec la convention $u_n = x_1 - x_n$)

$$\dot{u}_i = \varphi(u_{i+1}) - \varphi(u_i)\,,\ i = 1, \dots n \quad (n+1 \equiv 1), \tag{6.3}$$

que l'on peut écrire globalement $\dot{u} = F(u)$.

Remarque 6.8. *Comme dans le cas linéaire, on peut définir un graphe orienté (V, A) (voir définition 12.1, page 110), avec $V = \{1, 2, \dots, n\}$, et la règle $(i, j) \in A$ si et seulement si le comportement de i est directement influencé par le comportement de j : $A = \{(1,2)\,,\ \dots\,,\ (n-1, n)\,,\ (n, 1)\}$. Ce graphe contient de façon évidente un cycle*[42].

Si la fonction φ est strictement croissante, le système en distance admet un unique point d'équilibre $u_{eq} = (u_e, \dots, u_e)$, avec $u_e = L/n$.

Proposition 6.9. *On suppose que φ est une fonction C^1 strictement croissante sur $[0, +\infty[$. Le point d'équilibre $u_{eq} = (u_e, \dots, u_e)$, $u_e = L/n$, solution stationnaire de (6.3) est alors asymptotiquement stable.*

Démonstration. On écrit le gradient de F au point d'équilibre u_{eq} :

$$\nabla F(u_{eq}) = \varphi'(u_e) \begin{pmatrix} -1 & 1 & 0 & \cdot & 0 \\ 0 & -1 & 1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & -1 & 1 \\ 1 & \cdot & \cdot & 0 & -1 \end{pmatrix} = \varphi'(u_e) A_{per} = \varphi'(u_e) \left(-\operatorname{Id} + C\right).$$

où C est une matrice circulante, matrice de permutation particulière qui réalise le shift à droite périodique. Cette dernière vérifie $C^n = \operatorname{Id}$ et la famille $(C^k)_{0 \le k \le n-1}$ est libre, son polynôme caractéristique est donc $X^n - 1$, et ses valeurs propres sont ainsi les racines n-ièmes de l'unité. Les valeurs propres de A_{per} sont donc

$$\mu_k = -1 + \exp\left(\frac{2ik\pi}{n}\right), \ k = 1, \dots, n.$$

Toutes les valeurs propres sont donc de partie réelle ≤ 0, ce qui suggère une certaine stabilité du système. Mais pour $k = 0$, on trouve $\mu_0 = 0$, de telle sorte qu'il est a priori impossible de trancher quant à la stabilité de la solution. On peut néanmoins établir cette stabilité en remarquant que l'espace propre associé est $\mathbb{R}e$, où e est le vecteur dont tous les éléments sont égaux à 1. Or, du fait que, par construction, la somme des u_i est constante (égale à la longueur L), les perturbations admissibles sont de moyenne nulle, et donc orthogonale à e. On vérifie immédiatement que $e^\perp$ est stable par A_{per}, on peut donc se ramener à une étude spectrale sur $e^\perp$, dans lequel toutes les valeurs propres ont une parties réelle strictement négative [43]. □

Temps caractéristique de relaxation. La partie réelle de plus petit module est $\varphi'(u_e)(1 - \cos(2\pi/n))$, qui est proche de $\varphi'(u_e)2\pi^2/n^2$, ce qui donne un temps caractéristique de

$$\tau = \frac{1}{2\pi^2} \frac{n^2}{\varphi'(u_e)}.$$

42. Ce cyle est le plus petit, et il est unique au sens suivant : les autres cycles ne sont que des duplications de ce cycle simple (on peut "tourner" un nombre quelconque de fois).

43. On peut se ramener à une démarche plus habituelle en éliminant une variable redondante, dans les u_i, par exemple en écrivant que $u_n = L - \sum_{i=1}^{n-1} u_i$. La dernière équation s'écrit alors $u_{n-1} = \varphi(L - \sum_{i=1}^{n-1} u_i) - \varphi(u_{n-1})$, et le gradient s'écrit

$$\nabla F(u_{eq}) = \varphi'(u_e) \begin{pmatrix} -1 & 1 & 0 & \cdot & 0 \\ 0 & -1 & 1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & -1 & 1 \\ -1 & -1 & \cdot & -1 & -1 \end{pmatrix}$$

Le polynôme caractéristique P_{n-1} de cette matrice vérifie (en développant par rapport à la première colonne) $P_{n-1} = -\lambda P_{n-2} + (-1)^n$, d'où

$$P_{n-1} = (-1)^{n+1} \left(1 + \lambda + \dots + \lambda^{n-1}\right).$$

Les valeurs propres sont donc bien les racines n-èmes non triviales de l'unité.

Cette relaxation se produit selon un vecteur propre de *basse fréquence* en espace.

Corollaire 6.10. *Dans le cas où la fonction φ est nulle sur $[0,\ell]$, puis strictement croissante, sur $[\ell,+\infty[$, on a de même unicité d'un point d'équilibre, qui correspond à un mouvement effectif des véhicules si L est suffisamment grand (plus précisément si $L > n\ell$), sinon à un paquet d'entités immobilisées. Si φ n'est pas strictement croissante, on n'a pas forcément unicité du point d'équilibre. En particulier, si l'on suppose (ce qui est raisonnable) que φ est plate au delà d'une certaine valeur u_+ de la distance (correspondant à la visibilité), on peut avoir de multiples points d'équilibre dès que $L > nu_+$.*

Proposition 6.11. *On considère n entités avançant sur un chemin circulaire et fermé, on suppose l'évolution régie par*

$$\dot{x}_i = \varphi(x_{i+1} - x_i)\,,\ i = 1,\dots,n \quad (n+1 \equiv 1),$$

où φ est une fonction croissante. On note $u_i = x_{i+1} - x_i$, et l'on considère une solution du système (6.3). Pour toute fonction g convexe, la quantité

$$S(u(t)) = \sum_i g(u_i)$$

est décroissante.

Démonstration. Les distances vérifient

$$\dot{u}_i = \varphi(u_{i+1}) - \varphi(u_i)\,,\ i = 1,\dots,N.$$

On a donc

$$\begin{aligned}\frac{d}{dt}\left(\sum_i g(u_i)\right) &= \sum_i g'(u_i)\dot{u}_i = \sum_i g'(u_i)\left(\varphi(u_{i+1}) - \varphi(u_i)\right)\\ &= \sum_i \varphi(u_i)\left(g'(u_{i-1}) - g'(u_i)\right).\end{aligned}$$

Supposons g strictement convexe. La fonction g' étant alors strictement croissante, on peut effectuer le changement de variable $\beta_i = g'(u_i)$. La quantité ci-dessus s'exprime donc

$$\sum_i \varphi \circ (g')^{-1}(\beta_i)\,(\beta_{i-1} - \beta_i)\,,$$

où $\varphi \circ (g')^{-1}$ est une fonction croissante, qui s'écrit donc comme la dérivée d'une fonction convexe : $\varphi \circ (g')^{-1}(\beta) = \psi'(\beta)$. Comme ψ est convexe, on a

$$\psi(\beta_i) + \psi'(\beta_i)(\beta_{i-1} - \beta_i) \le \psi(\beta_{i-1}),$$

de telle sorte que

$$\frac{d}{dt}\left(\sum_i g(u_i)\right) \le \sum_i \left(\psi(\beta_{i-1}) - \psi(\beta_i)\right) = 0.$$

Si g n'est pas strictement convexe, on applique la démarche à $g(u) + \varepsilon u^2$, et on fait tendre ε vers 0. □

Remarque 6.12. *Dans le cas d'une route de longueur* 1*, on peut interpréter* $u = (u_i)$ *comme une mesure de probabilité sur un ensemble à* N *éléments. Prenant* $g(x) = x \log x$ *dans ce qui précède, on a alors décroissance de l'*entropie *(selon la définition 10.1, page 101)*

$$S(u) = \sum_i u_i \log u_i.$$

Remarque 6.13. *Considérons le cas d'un* g *strictement convexe (par exemple* $g(u) = u \log u$*). Si la fonction* φ *est strictement croissante sur l'intervalle de valeurs couvert par les* u_i*, alors la décroissance de l'entropie est stricte, tant que l'on n'a pas l'état stationnaire* $u_1 = u_2 = \cdots = u_N = L/N$*. On converge alors nécéssairement vers l'unique état stationnaire. Si en revanche* φ *n'est pas strictement croissante, la propriété de convergence peut être invalidée (l'état équi-réparti n'est pas asymptotiquement stable). C'est le cas par exemple si, au delà d'une certaine distance, l'entité va à la vitesse maximale, de telle sorte que la fonction* φ *est constante au delà d'une certaine valeur. Si la route circulaire est assez grande, on peut avoir une distribution non régulière d'entités progressant toutes à la vitesse maximale. D'un point de vue macroscopique, cette situation correspond à une onde progressive que l'on observe en effet lorsque la fonction flux (ici la densité multipliée par la vitesse) est affine sur certaines plages de densité.*

Corollaire 6.14. *Dans le cas où la fonction* φ *est nulle sur* $[0, \ell]$*, puis strictement croissante, sur* $[\ell, +\infty[$*, on a la propriété suivante : si les valeurs initiales des distances sont* $> \ell$*, alors la solution est telle que les* u_i *sont minorés par* $\ell + \eta$*, avec* $\eta > 0$*.*

Démonstration. On peut choisir $g(u) = 1/(u-\ell)$, qui est convexe pour $u > \ell$. La décroissance de l'entropie exclut que l'un des u puisse tendre vers ℓ. Plus précisément, on a

$$\sum g(u_i) \leq S_0 = \sum g(u_i^0),$$

d'où, pour tout i,

$$u - \ell > 1/S_0,$$

ce qui conclut la démonstration. □

Propagation des perturbations. L'étude de l'exponentielle de la matrice du système linéarisé, dans le cas non périodique, avait mis en évidence une propagation des perturbations vers l'amont à la célérité $-u_e\varphi'(u_e)$. Plus précisément, nous nous étions intéressés à la propagation d'une perturbation ponctuelle (affectant seulement le véhicule de tête). On se propose ici de quantifier ce phénomène de propagation dans le cas périodique. Le système linéarisé s'écrit

$$\frac{du}{dt} = \varphi'(u_e)\,(-\operatorname{Id} + C)\,u.$$

La matrice est diagonalisable, d'éléments propres

$$\mu_k = \varphi'(u_e)\left(-1 + \exp\left(\frac{2ik\pi}{n}\right)\right),\ w_k = \left(\exp\left(\frac{2ik\pi m}{n}\right)\right)_m.$$

Les parties réelles des valeurs propres,

$$\operatorname{Re}(\mu_k) = -\varphi'(u_e)\left(1 - \cos\left(\frac{2k\pi}{n}\right)\right) \leq 0,$$

quantifient l'amortissement exponentiel selon les différents modes. La propagation en espace est encodée par la partie imaginaire. La partie correspondante de la solution s'écrit

$$\exp\left(\varphi'(u_e)\sin\left(\frac{2k\pi}{n}\right)t\right)\exp\left(\frac{2ik\pi m}{n}\right)=\exp\left(\frac{2ik\pi}{n}\left(m+\underbrace{\frac{\varphi'(u_e)n}{2\pi k}\sin\left(\frac{2k\pi}{n}\right)}_{=-c_k}t\right)\right),$$

où m indexe les n entités impliquées. Cette expression correspond donc à une propagation (sur la suite des indices) à vitesse constante c_k. On retrouve pour k/n petit une célérité de l'ordre de $-\varphi'(u_e)$ (en s^{-1}, ou entités par seconde), ou, si l'on prend en compte le fait que les entités sont séparées de u_e, d'une vitesse effective de $-u_e\varphi'(u_e)$ (en ms^{-1}).

6.4 Extensions, développements

Individus de profils différents . Il est peu réaliste de considérer que tous les individus ont le même comportement. Si l'on reprend le modèle initial sur route rectiligne, avec un véhicule de tête qui va à vitesse constante $v_e=\varphi_{n+1}(u_e)$, et que l'on se donne des courbes de comportement φ_i toutes strictement croissantes (pour $u\geq u_m$), on aura existence et unicité d'un point d'équilibre en distances dès que la vitesse de tête est atteignable par chacun des suivants, i.e.

$$v_e<\max_u\varphi_i(u)\quad\forall i.$$

On écrit u_e^i la distance qui réalise $v_e=\varphi_i(u_e^i)$. Le vecteur $u_e^1,\ldots,u_e^n$ est alors point d'équilibre. L'étude de stabilité de ce point d'équilibre conduit à une matrice du type

$$\nabla F=\begin{pmatrix}-\beta_1 & \beta_2 & 0 & \cdot & 0\\ 0 & -\beta_2 & \beta_3 & \cdot & \cdot\\ \cdot & \cdot & \cdot & \cdot & 0\\ \cdot & \cdot & \cdot & -\beta_{n-1} & \beta_n\\ 0 & \cdot & \cdot & 0 & -\beta_n\end{pmatrix},\ \beta_i=\varphi_i'(u_e^i)\quad i=1,\ldots,n. \tag{6.4}$$

La situation est assez troublante, car, si l'on peut espérer que le phénomène de propagation de l'information vers l'amont soit préservé pour ce système perturbé, la structure du problème est complètement différente. Les β_i n'ont aucune raison d'être identiques, on peut considérer que, même s'ils peuvent être voisins, ils sont génériquement [44] différents deux à deux. Mais alors la matrice est diagonalisable, et l'étude du comportement de la solution du système linéarisé $e^{tA}\,u_{pert}$, est complètement différente. Cette étude est à mener avec précaution, car les matrices diagonalisables de ce type ne sont pas loin d'une matrice qui ne l'est pas, ce qui peut conduire à un comportement singulier. Pour s'en convaincre, considérons la famille de matrices A^ε associées à

$$\beta^\varepsilon=(\beta_1^\varepsilon,\ldots,\beta_n^\varepsilon),$$

où les β_i^ε tendent tous vers le même β limite, que l'on prendra égal à 1 pour simplifier. On vérifie immédiatement que les vecteurs propres u_i^ε normalisés associés convergent (à sous suite

44. Cette notion de *généricité* est très utilisée oralement, elle est à manier avec précaution. Elle signifie ici en substance que, au voisinage d'une situation considérée, l'ensemble des cas pour lesquels la propriété (dite générique) n'est pas vérifiée est de mesure nulle.

extraite près) vers un vecteur propre de la matrice $A = -\operatorname{Id} + N$, qui n'a qu'une droite propre (selon le premier vecteur de base). Tous les vecteurs propres tendent donc à avoir la même direction. La diagonalisation effective d'une telle matrice (pour ε petit mais non nul) risque d'être extrêmement instable, on peut par exemple s'attendre à ce que la plupart des méthodes numériques d'estimation de valeurs propres ne fonctionnent pas. On peut se convaincre de la difficulté du problème, tout en vérifiant que l'on aura bien propagation vers l'amont, en considérant le cas de 2 entités libres (donc de deux distances, i.e. 3 entités, celle de tête ayant une vitesse imposée). On définit

$$A \begin{pmatrix} -1 & 1+\varepsilon \\ 0 & -1-\varepsilon \end{pmatrix}.$$

Cette matrice est évidemment diagonalisable pour $\varepsilon \neq 0$, avec une matrice de passage

$$A \begin{pmatrix} 1 & 1+\varepsilon \\ 0 & -\frac{\varepsilon}{1+\varepsilon} \end{pmatrix}.$$

Si l'on considère maintenant la solution du problème d'évolution linéaire, avec une perturbation sur les distance de tête, on obtient (on n'indique pas la dépendance de P vis à vis de ε pour alléger les notations)

$$e^{tA^\varepsilon} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = PP^{-1} e^{tA^\varepsilon} PP^{-1} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \frac{1+\varepsilon}{\varepsilon} P \begin{pmatrix} e^{-t} \\ e^{-t(1+\varepsilon)} \end{pmatrix} = e^{-t} \begin{pmatrix} \frac{1+\varepsilon}{\varepsilon}(1-e^{-t\varepsilon}) \\ e^{-t\varepsilon} \end{pmatrix},$$

et l'on retrouve bien par développement limité une évolution de la seconde distance (première composante) en te^{-t} (au premier ordre en ε), comme pour la matrice limite non diagonalisable. Noter que l'on est passé par l'intermédiaire de matrices très mal conditionnées [45] : dans une situation où les calculs ne pourraient pas être faits analytiquement, il serait périlleux de suivre cette démarche en cherchant à diagonaliser de façon approchée les matrices de type de celle définie par (6.4), pour des β_i proches les uns des autres.

Question 6.3. Intégrer au modèle le fait que l'on ne réagit pas instantanément à une variation de distance.

Question 6.4. Proposer un modèle *macroscopique* intégrant le fait que l'on définit sa vitesse avec un certain retard. On pourra introduire une distance subjective w_i pour chaque individu, et considérer qu'on a relaxation de cette distance vers la vraie distance instantanée, avec un temps caractéristique τ.

Stratégie dépendant de la vitesse. On se propose ici de baser le modèle sur un principe différent : on considère que chaque entité a une vitesse qu'elle souhaiterait avoir si elle était seule. A chaque instant elle estime la distance à l'entité précédente, ainsi que sa vitesse. A la vitesse estimée elle associe une distance $D(v)$ (qui correspondrait à la distance qui permet d'éviter une collision avec quelqu'un qui avance à la vitesse v, en cas d'arrêt brusque). Si sa distance effective est supérieure à cette distance, elle va à sa vitesse souhaitée, sinon, la

45. Voir section 18.1, page 179 : les matrices sont de norme contrôlée mais, du fait que les vecteurs propres sont quasiment colinéaires, leurs inverses ont une norme qui tend vers $+\infty$ quand les β_i tendent à se confondre.

vitesse souhaitée est significativement réduite (jusqu'à ce que la distance effective redevienne de l'ordre de $D(v)$). Une telle démarche conduit par exemple au modèle suivant :

$$v_i = \dot{x}_i = U_i \left(1 + \exp\left(-\frac{x_{i+1} - x_i - D(v_{i+1})}{u_s}\right)\right)^{-1}.$$

Ce modèle est considérablement plus compliqué que les précédents, car la vitesse de chaque entité dépend de la vitesse des autres de façon non linéaire, ni l'unicité ni même l'existence d'une collections de vitesses réalisant l'ensemble des relations ne sont garanties. Plus précisément, la difficulté du propblème est conditionnée par le type du graphe des dépendances (voir remarques 6.5 et 6.8). Dans le cas d'un graphe acyclique (entités sur une route rectiligne), on fixe la vitesse de l'entité de tête, et les vitesses sont déterminées de façon unique en descendant la hiérarchie. Dans le cas où l'on a des cycles en revanche, comme dans le cas d'une route circulaire, le problème est plus délicat, il peut exister plusieurs collections de vitesses qui vérifient le système.

7 Trafic routier ou piéton – micro – 1d – ordre 2 en temps

7.1 Le modèle

On s'intéresse ici à un modèle de trafic routier (ou piéton) microscopique (les entités sont suivies individuellement) d'ordre 2 en temps. On note $x_i = x_i(t)$ la position de la i-ème entité au temps t, qui évolue sur $\mathbb{R}$ (on considérera par la suite le cas périodique). Le modèle s'écrit

$$\ddot{x}_i = \frac{1}{\tau}(\varphi(x_{i+1} - x_i) - \dot{x}_i), \tag{7.1}$$

où τ est un temps caractéristique d'accession à une vitesse souhaitée. Pour des voitures, τ représente le temps caractéristiques mis par le conducteur pour accéder à la vitesse qu'il souhaite. Ce temps peut dépendre du type de véhicule, du comportement du conducteur, on pourrait même considérer (au prix néanmoins d'un changement profond sur la nature du modèle) qu'il dépend du signe de $\varphi(x_{i+1} - x_i) - \dot{x}_i$ (on peut avoir une voiture au moteur poussif, mais qui possède de bons freins). Nous supposerons que ce temps τ est constant. La fonction $u \mapsto \varphi(u)$ représente la vitesse que souhaite avoir un véhicule à la distance u du véhicule qui le précède. Si l'on ne prend pas en compte la taille des véhicules, on choisira une fonction croissante qui s'annule en 0, qui tend vers une valeur limite U quand u tend vers $+\infty$. Un exemple d'une telle fonction est

$$u \longmapsto U\left(1 - \exp\left(-u/u_s\right)\right), \tag{7.2}$$

où u_s représente l'ordre de grandeur de la distance considérée par le conducteur comme étant de sécurité (pour un vitesse égale à $1-1/e \approx 0.6$ fois la vitesse maximale. Pour un conducteur agressif peu scrupuleux des distances de sécurité, u_s sera donc petit. Nous supposerons pour simplifier les conducteurs tous identiques, ce qui conduit bien au modèle (7.1), avec une fonction φ qui ne dépend pas de i.

Solutions globales et accidents

Si l'on suppose la fonction φ Lipschitzienne, son prolongement par 0 sur $]-\infty, 0]$ reste Lipschitzien, et le théorème de Cauchy-Lipschitz appliqué au système

$$\left|\begin{array}{rcl} \dot{x}_i &=& v_i \\ \dot{v}_i &=& \dfrac{1}{\tau}(\varphi(x_{i+1} - x_i) - v_i), \end{array}\right. \tag{7.3}$$

assure l'existence d'une unique solution maximale, qui est globale d'après la proposition 21.12, page 214. De façon évidente les solutions pour lesquelles les distances sont nulles voire négatives sont à considérer avec une attention particulière. S'il advient que l'une des distances s'annule, cela traduit une collision, et le modèle que nous avons écrit, même s'il est défini mathématiquement, n'a plus de sens. Vérifions que des accidents sont susceptibles de se produire. On considère pour simplifier un véhicule derrière un véhicule à l'arrêt en 0. La position du véhicule en mouvement, notée x, vérifie

$$\ddot{x} = \frac{1}{\tau}\left(\varphi(-x) - \dot{x}\right),$$

avec condition initiales en position et vitesse. On s'intéresse à ce qui se passe au voisinage de l'origine, on a alors $\varphi(-x) \approx -\varphi'(0)\, x$. Notant $\varphi'(0) = 1/\eta$, on obtient

$$\ddot{x} + \frac{1}{\tau}\dot{x} + \frac{1}{\tau\eta}x = 0.$$

Les racines de l'équations caractéristique sont

$$\lambda = \frac{1}{2\tau}\left(-1 \pm \sqrt{1 - \frac{4\tau}{\eta}}\right)$$

On aura donc amortissement non oscillant pour $\tau/\eta < 1/4$. Dans le cas contraire, x va atteindre 0 (à vitesse non nulle), on ne peut donc pas exclure dans ce cas l'occurrence d'accident (et donc la durée de vie finie de la solution en tant que trajectoire viable).

7.2 Stabilité

On peut se demander dans un premier temps si le modèle ci-dessus permet de reproduire des régimes stationnaires stables. Nous nous concentrerons ici sur le cas périodique (route circulaire du type périphérique, circuit de formule 1). Pour cela considérons la situation de N entités sur une route circulaire, équidistants (distance $u_e = L/N$). La configuration où tous les véhiculent roulent à la même vitesse $V = \varphi(u_e)$, correspond au régime stationnaire.

Pour étudier la stabilité de cette situation, on travaille sur les variables de distance $u_i = x_{i+1} - x_i$. Le modèle s'écrit pour cette nouvelle variable

$$\ddot{u}_i = \frac{1}{\tau}(\varphi(u_{i+1}) - \varphi(u_i) - \dot{u}_i), \tag{7.4}$$

pour lequel le vecteur $(u_e, u_e, \ldots, u_e)$ est point fixe. On peut écrire ce modèle $(\dot{u}, \dot{v}) = \Psi(u, v)$, avec $v = \dot{u}$.

La stabilité du point d'équilibre est conditionnée par les propriétés de la matrice

$$\nabla\Psi_{|y=y_f} = \begin{pmatrix} 0 & \mathrm{I_d} \\ \dfrac{1}{\tau}\varphi'(u_e)A_{\mathrm{per}} & -\dfrac{1}{\tau}\,\mathrm{Id} \end{pmatrix}, \quad \text{avec } A_{\mathrm{per}} = \begin{pmatrix} -1 & 1 & 0 & \cdot & 0 \\ 0 & -1 & 1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & \cdot & \cdot & -1 & 1 \\ 1 & 0 & \cdot & 0 & -1 \end{pmatrix}$$

La matrice A_{per} est somme de $-\,\mathrm{Id}$ et d'une matrice circulante C. Cette dernière vérifie $C^n = \mathrm{Id}$, son polynôme caractéristique est donc $X^n - 1$, et ses valeurs propres sont ainsi les racines n-ièmes de l'unité. Les valeurs propres de A_{per} sont donc

$$\mu_k = -1 + \exp\left(\frac{2ik\pi}{n}\right), \; k = 1, \ldots, n.$$

le problème aux valeurs propres pour la matrice globale s'écrit donc

$$v = \lambda u\,, \;\; \frac{\varphi'(u_e)}{\tau}Au - \frac{1}{\tau}v = \lambda v \Longrightarrow \left(\lambda^2 + \frac{\lambda}{\tau} - \frac{\varphi'(u_e)}{\tau}A\right)u = 0$$

Pour tout couple propre u_k, $\mu_k = -1 + \exp\left(\frac{2ik\pi}{n}\right)$ de A_{per}, on aura donc deux valeurs propres pour la matrice globale, qui sont les racines de

$$\lambda^2 + \frac{\lambda}{\tau} - \frac{\varphi'(u_e)}{\tau}\mu_k = 0,$$

c'est à dire

$$\lambda_k^{\pm} = \frac{1}{2\tau}\left(-1 \pm \sqrt{1 - 4\varphi'(u_e)\tau\left(1 - \exp\left(\frac{2ik\pi}{N}\right)\right)}\right)$$

Notons $\alpha = 4\varphi'(u_e)\tau$. Le lieu des $\lambda_k^{\pm}$ est donc l'ensemble image du cercle unité par la transformation (bivaluée) dans le plan complexe

$$z \longmapsto \left(-1 \pm \sqrt{1 - \alpha(1-z)}\right)/2\tau.$$

Le point essentiel est de déterminer si les valeurs propres sont toutes de parties réelles positives. On se ramène donc à la question suivante : la racine carrée du cercle centré (sur l'axe réel) en $1-\alpha$ et de rayon α appartient-elle au demi-espace $Re(z) \leq 1$?

On peut préciser la réponse à cette question :

Lemme 7.1. *La racine carrée du cercle centré (sur l'axe réel) en* $1-\alpha$ *et de rayon* α *intersecte le demi espace* $Re(z) > 1$ *si et seulement si* $\alpha > 2$.

Démonstration. Une première approche consiste à poser le problème à l'envers, en remarquant [46] qu'il y aura des point de l'ensemble recherché qui sont à droite de la droite $\mathrm{Re}(z) = 1$ dès que le carré de cette droite intersecte le cercle C_α en d'autres points que 1. Le carré de cette droite est une parabole, lieu des $z = (1+iy)^2 = 1 - y^2 + 2iy$ pour y décrivant $\mathbb{R}$. Le rayon de courbure en 1 de cette parabole est 2, il est donc plus petit que le rayon α du cercle dès que $\alpha > 2$.

On peut essayer de se faire une idée plus précise du lieu des valeurs propres : l'ensemble que l'on cherche à décrire est l'ensemble des $\bar{x} + i\bar{y}$ tels que

$$\bar{x}^2 - \bar{y}^2 = x\,,\ 2\bar{x}\bar{y} = y$$

où $x + iy$ décrit le cercle d'équation $(x - 1 + \alpha)^2 + y^2 = \alpha^2$. Il s'agit donc d'une courbe quartique d'équation

$$\left(\bar{x}^2 - \bar{y}^2 - 1 + \alpha\right)^2 + 4\bar{x}^2\bar{y}^2 = \alpha^2,$$

qui contient le point $z = 1$.

On pose $X = \bar{x}^2$, $Y = \bar{y}^2$, pour obtenir

$$(X - Y - 1 + \alpha)^2 + 4XY = \alpha^2\,,\ \text{ soit } \ \Psi(X,Y) = 0.$$

La dérivée de Ψ par à X, qui est $2(X + Y - 1 + \alpha)$ est non nulle en $(1,0)$. On peut donc d'après le théorème des fonctions implicites, exprimer X fonction de Y au voisinage de ce point, et estimer la dérivée de cette courbe

$$\frac{dX}{dY}_{|(1,0)} = \frac{2-\alpha}{\alpha},$$

qui est > 0 (ie. les abscisses dépassent strictement 1) dès que $\alpha > 2$. □

46. Astuce suggérée par S. Di Marino

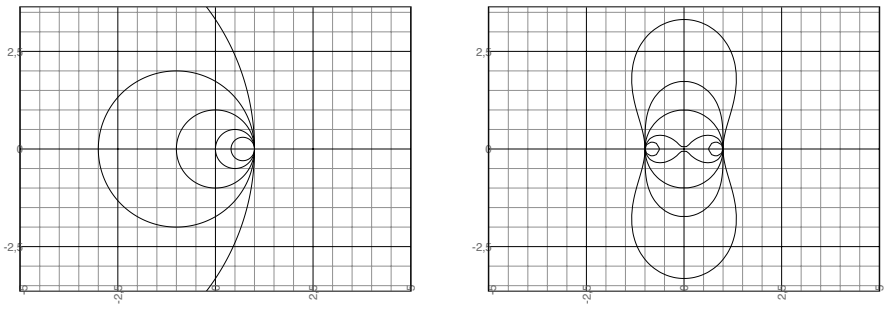


FIGURE 7.1 – Cercles (gauche) et quartiques associées (droite), pour $\alpha = 0.3\,,\ 0.5\,,\ 1\,,\ 2\,,\ 6$.

Remarque 7.2. *Pour α entre 0 et 2, le lieu des valeurs propres est une quartique dans la bande $x \in [-1, 1]$, tangente en 1 à la droite $y = 1$. Noter que, bien que le comportement soit stable, on a des valeurs propres de partie réelle certes négative mais petite en valeur absolue. Ces valeurs propres correspondent à des racines $n-$èmes proches de 1, donc des modes de très basses fréquences (oscillations en espace dont la période est le l'ordre de la longueur totale du chemin).*

Remarque 7.3. *Pour $\alpha = 1/2$, le lieu des valeurs propres est une* lemniscate de Bernoulli *(voir figure 7.1), qui correspond à la transition vers la connexité du lieu des valeurs propres. Pour $\alpha = 1$, la quartique est le cercle unité (en fait deux copies du cercle unité confondues). Pour la valeur critique $\alpha = 2$ on a une forme de stade allongée verticalement, avec une courbure nulle en 1 ; pour $\alpha > 2$, la courbe délimite un ensemble qui n'est plus convexe.*

Mode le plus instable

On peut pousser l'analyse ci-dessus en cherchant à identifier le mode le plus instable. A partir de

$$(X - Y - 1 + \alpha)^2 + 4XY = \alpha^2$$

on obtient

$$\frac{dX}{dY} = -\frac{X + Y + 1 - \alpha}{X + Y - 1 + \alpha}.$$

La variable X est donc maximale pour $Y = -X - 1 + \alpha$. En réinjectant dans l'équation de la courbe, on obtient

$$X = \frac{\alpha^2}{4(\alpha - 1)}.$$

Pour estimer l'angle correspondant au mode le plus instable, on se ramène à la variable x :

$$x - 1 + \alpha = X - Y - 1 + \alpha = 2X.$$

L'angle est donc

$$\theta = \arccos\left(\frac{2X}{\alpha}\right) = \arccos\left(\frac{\alpha}{2(\alpha - 1)}\right).$$

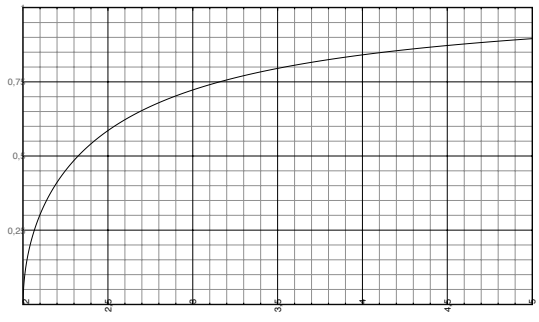


FIGURE 7.2 – Angle θ (mode le plus instable) fonction de α.

Pour α grand, on tend donc vers un angle de $\pi/3$, ce qui correspond à la $N/6$-ième racine N-ième de l'unité (on suppose N divisible par 6, sinon le mode le plus instable est le plus proche de celui-là). Le vecteur propre de la matrice A_{per} associé à la k-ième racine est

$$u_k = \left(e^{2i\pi kn/N}\right)_n,$$

soit, avec $k = N/6$, une oscillation de période 6 en n. Le mode le plus instable est donc un mode de petite période (relativement au nombre total de véhicules, supposé grand), qui affecte typiquement des groupes de 6 entités consécutives, avec alternances de sous paquets de 3 en compression, décompression, etc ....

On peut aussi estimer cet angle au voisinage de l'apparition de l'instabilité ($\alpha = 2^+$), en écrivant $\varepsilon = \alpha - 2$, on a

$$\theta = \arccos\left(\frac{\alpha}{2(\alpha-1)}\right) \arccos\left(\frac{1+\varepsilon/2}{1-\varepsilon}\right) = \arccos\left(1 - \frac{\varepsilon}{2} + o(\varepsilon)\right) \sim \sqrt{\varepsilon} = \sqrt{\alpha - 2}.$$

On aura donc pour $\alpha - 2$ petit un angle θ petit, ce qui correspond à des basses fréquences en espace, mais la croissance de θ vis-à-vis de $\alpha - 2$ est très raide : le mode le plus instable correspond très vite à une mode de haute fréquence (oscillation qui implique localement un nombre faible d'entités). Si l'on prend par exemple $\alpha = 2.3$, on a un angle autour de $\pi/6$, qui correspond à une perturbation qui affecte localement 12 entités (voir figure 7.2). La plage sur laquelle les modes les plus instables sont de basse fréquence est donc extrêmement étroite : il peut être délicat de les observer en pratique [47].

Remarque 7.4. *Le paramètre α qui conditionne la stabilité s'écrit*

$$\alpha = 4\varphi'(u_e)\tau,$$

qui est bien un nombre sans dimension : φ associe à une distance une vitesse, sa dérivée est donc l'inverse d'un temps η. C'est le temps caractéristique associé au modèle d'ordre un en temps (voir proposition 6.6, page 67). La condition d'instabilité s'écrit donc $\tau/\eta > 1/2$. Le temps τ quantifie la réactivité de l'entité. Dans le cas du trafic routier, cette réactivité englobe la réactivité du véhicule. On pourra se faire une idée de ce temps caractéristique en

47. La plage de valeurs sur laquelle on a des basses fréquence, i.e. le voisinage immédiat de 2^+, est d'une amplitude inférieure à la précision que l'on peut espérer avoir sur l'estimation des paramètres τ et $\eta = \varphi'(u_e)$.

imaginant l'expérience suivante : le véhicule nous précédant pile brusquement, quel temps allons nous mettre pour ralentir significativement notre vitesse (i.e. réduction au $2/3$*, pour fixer les idées) ? Ce temps*[48] *est de l'ordre de quelques seconde, disons* 5 *ou* 6*. La condition indique que l'on aura donc un système plus stable dans le cas d'une bonne réactivité (*τ *petit). Le temps* η *qui intervient dans le modèle de comportement est moins directement accessible à l'intuition, puisqu'il apparaît en fait comme l'inverse d'une variation en vitesse relativement à la distance. Dans l'hypothèse raisonnable d'une fonction* φ *concave, défini par exemple par* (7.2)*, on a*

$$\varphi'(u_e) = \frac{U}{u_s} \exp(-u_e/u_s).$$

Dans les cas "dilués" (u_e *grand devant* u_s*),* η *sera très petit, et le système sera stable. La situation intéressante pour un trafic dense, i.e.* $\exp(-u_e/u_s) \approx 1$*. Le temps* η *s'écrit alors* u_s/U*, où* U *est la vitesse maximale autorisée, et* u_s *la distance "typique entre véhicule", plus précisément la distance inter-véhicules correspondant à une vitesse de* $1 - 1/e \approx 0.6$ *fois la vitesse maximale. Sur autoroute, on peut prendre une centaine de mètres comme ordre de grandeur, ce qui donne un* η *de l'ordre de* 2 *ou* 3*. On vérifie ainsi immédiatement que la valeur critique* $1/2$ *correspond à l'ordre de grandeur de* τ/η *: il peut être très délicat en pratique de savoir si l'on est dans une situation stable ou instable.*

Exercice 7.1. On trouve dans les ouvrages de sécurité routière les ordres de grandeur suivant pour la distance totale (temps de réaction + freinage effectif) d'arrêt en fonction de la vitesse :

Vitesse (en km h^{-1})	30	50	70	90	120
Distance (en m)	14	28	46	68	108

En supposant que chaque conducteur adapte sa distance à sa vitesse en considérant qu'il doit pouvoir éviter la collision en cas d'arrêt brusque du véhicule devant lui, estimer le paramètre η en fonction du régime d'écoulement (densité ou distance inter-véhicule), et préciser la condition que doit vérifier le temps τ (qu'on peut considérer encoder le temps de réaction effectif du conducteur et de son véhicule) pour que l'on ait stabilité asymptotique du régime stationnaire.

Exercice 7.2. On considère un modèle de piéton dans un couloir circulaire, avec une fonction φ basée sur les mesures décrites dans la section **??**, plus précisément la figure **??**, page **??**. Ces données permettent de reconstruire la fonction φ qui décrit le comportement des piétons, et donc, à une densité L/n donnée (ou de façon équivalente une distance moyenne entre personnes), on peut estimer le $\varphi'(u_e) = 1/\eta$. Le paramètre τ du modèle décrit dans cette section, qui encode des effets purement instationnaires, n'est pas accessible à partir de ces données. Evaluer la stabilité du système en fonction du régime considéré, et du paramètre τ du modèle.

48. Cette démarche nous met face à un défaut manifeste du modèle. Si l'on renverse l'expérience en considérant une situation où le véhicule devant nous "disparaît" (par exemple un tracteur qui se range sur le bord de la route pour nous laisser passer), le temps mis pour atteindre la vitesse maximale va être beaucoup plus long que ces quelques secondes, à moins d'avoir un véhicule extrêmement puissant. On sent qu'il faudrait un τ pour l'accélération, et un autre pour la décélération. Une telle démarche pourrait s'envisager, mais précisons que la nature du modèle change considérablement, puisque le modèle n'est plus Lipschitz (il est quand même continu car le "switch" se produit précisément quand la différence entre la vitesse souhaitée et la vitesse effective change de signe, et donc s'annule).

7.3 Extensions, développements

Modèle macroscopique associé. Comme dans le cas du modèle d'ordre 1, on peut dériver formellement une équation aux dérivées partielles pour les perturbations de distances au voisinage d'un point d'équilibre. On a

$$\ddot{u}_i = \frac{1}{\tau}\left(\varphi(u_{i+1}) - \varphi(u_i) - \dot{u}_i\right).$$

La situation $u_i \equiv u_e$ est point d'équilibre du système[49]. On considère une perturbation de cette situation, les distances sont de type $u_e + u_i$, où u_i est maintenant une (petite) variation de u_e. On obtient

$$\ddot{u}_i = \frac{1}{\tau}\left(\varphi'(u_e)(u_{i+1} - u_i) - \dot{u}\right) = \frac{1}{\tau}\left(u_e\varphi'(u_e)\frac{u_{i+1} - u_i}{u_e} - \dot{u}\right)$$

Si l'on considère que les u_i sont les valeurs d'une fonction lisse u aux points équidistants de u_e, on obtient formellement

$$\partial_{tt}u + \frac{1}{\tau}\left(\partial_t u - c\partial_x u\right) = 0,$$

avec $c = u_e\varphi'(u_e)$.

Exercice 7.3. Montrer que le modèle macroscopique obtenu précédemment présente un comportement génériquement instable. Préciser ce qui est le plus discutable dans le développement asymptotique formel ayant conduit au modèle, et qui peut expliquer que le régime stable observé pour le modèle microscopique ait complètement disparu au niveau macroscopique.

Exercice 7.4. Proposer un modèle qui prenne en compte à la fois l'inertie de l'entité en mouvement (comme cela a été fait dans cette section), et le fait que les conducteurs ou piétons mettent un certain temps à réagir. On pourra introduire (comme dans l'exercice 6.3) une distance subjective w_i pour chaque individu, et considérer qu'on a relaxation de cette distance vers la vraie distance instantanée, avec un temps caractéristique τ'.

49. On pourra considérer le cas périodique, avec $u_e = L/n$, ou la situation d'entités sur une voie rectigne, derrière une entité de tête à vitesse fixée égale à $v_e = \varphi(u_e)$.

Deuxième partie

Notions, développements transverses

8 Analyse fonctionnelle et modélisation

Nous rassemblons ici quelques interprétations en termes de modélisation de notions théoriques en analyse fonctionnelle.

8.1 Espaces de Sobolev

Système masses-ressort en dimension 1

On considère un ensemble de $N+1$ masses alignées sur l'axe des x, reliées par des ressorts de même raideur k_N et même longueur au repos ℓ_N. On impose $x_0 = 0$ et $x_N = 1$ (la chaîne est accrochée à ses extrémités). On note (x_i) la configuration de référence [50], avec $x_i = i/N$. La position de la masse i est notée $x_i + u_i$. L'énergie potentielle élastique du système est

$$E_N = \frac{1}{2} \sum_{i=0}^{N-1} k_N \mid x_{i+1} - x_i + u_{i+1} - u_i - \ell_N \mid^2.$$

Si l'on choisit ℓ_N de telle sorte que la configuration de référence soit d'énergie nulle, i.e. $\ell_N = 1/N$, on obtient

$$E_N = \frac{1}{2} \sum_{i=0}^{N-1} k_N \mid u_{i+1} - u_i \mid^2,$$

que l'on peut aussi écrire

$$E_N = \frac{1}{2} \sum_{i=0}^{N-1} \ell_N \left(k_N \ell_N\right) \left| \frac{u_{i+1} - u_i}{\ell_N} \right|^2.$$

En choisissant $k_N = K/\ell_N$, on reconnait une somme de Riemann, qui converge donc lorsque N tend vers $+\infty$ (en supposant que u_i est la valeur en x_i d'un champ de déplacement continûement différentiable $x \longmapsto u(x)$), vers

$$\frac{K}{2} \int_0^1 \left|u'(x)\right|^2 \, dx,$$

ce qui permet d'interpréter le carré de la semi-norme H^1 comme l'énergie potentielle mécanique d'un système élatisque obtenu comme limite du système discret de masses reliées par des ressorts, avec une raideur qui tend vers l'infini comme le nombre de masses.

On peut retrouver la norme H^1 complète (avec la partie L^2) en considérant que chacune des masses du système discret est accrochée au point de référence x_i par un ressort de longueur au repos nulle, et de raideur k_N^0. Le surplus d'énergie discrète est alors

$$E_N^0 = \frac{1}{2} \sum_{i=1}^{N-1} k_N^0 \left|u_i\right|^2$$

50. Cette configuration minimise l'énergie potentielle dans le cas où la longueur au repos est inférieure à $1/\ell_N$.

qui tend vers

$$E^0 = \frac{K^0}{2} \int_0^1 u(x)^2 \, dx,$$

si l'on prend $k_N^0 = K^0 \, \ell_N$.

Noter que la raideur des ressorts "externes" tend vers 0, alors que celle des ressorts internes tend vers $+\infty$.

Les fonctions de H^1 sont continues en dimension 1. Si un champ de déplacement u présente une discontinuité, alors pour le système discret associé l'un des $u_{i+1} - u_i$ va tendre vers une valeurs non nulle. Or l'énergie d'un ressort du système discret est $KN \, |u_{i+1} - u_i|^2$, qui tend alors vers l'infini quand N tend vers l'infini.

Système masses-ressort en dimension ≥ 2

En dimension 2, on peut concevoir un ensemble de $(N+1)^2$ masses disposées aux nœuds d'un réseau cartésien. L'extension directe de ce qui précède consiste à considérer des déplacements de masses dans le plan du réseau, donc des déplacements vectoriels (ce qui est possible, et conduirait à une norme du type de celle que l'on tuilise en elasticité pour les déplacements). Pour rester sur un champ scalaire, on considère plutôt ici des déplacements verticaux (dans la direction transverse au plan du réseau), et l'on suppose que les masses sont reliées (entre voisines) par des ressorts de longueur au repos nulle, et de raideur k_N. Les masses sur le bord sont supposés fixées. Si l'on note $u_{i,j}$ le déplacement vertical, l'énergie du ressort entre (i,j) et $(i+1,j)$ s'écrit

$$\frac{k_N}{2} \left(\ell_N^2 + |u_{i+1,j} - u_{i,j}|^2 \right).$$

L'énergie totale du système s'écrit comme

$$\sum_{0 \leq i \leq N-1} \sum_{0 \leq j \leq N-1} \frac{1}{2} k_N \left(2\ell_N^2 + |u_{i+1,j} - u_{i,j}|^2 + |u_{i,j+1} - u_{i,j}|^2 \right)$$

$$= K_N + \sum_{0 \leq i \leq N-1} \sum_{0 \leq j \leq N-1} \frac{1}{2} k_N \ell_N^2 \left(\left| \frac{u_{i+1,j} - u_{i,j}}{\ell_N} \right|^2 + \left| \frac{u_{i,j+1} - u_{i,j}}{\ell_N} \right|^2 \right)$$

qui approche, si l'on prend $k_N = k$ (indépendant de N)

$$k + \frac{k}{2} \int_\Omega |\nabla u|^2 ,$$

où $u_{i,j}$ est la valeur du champ u (supposé continûment différentiable) au point $(i\ell_N, j\ell_N)$. Le k dans l'expression précedente correspond à l'énergie du réseau non déformé (qui est non nulle du fait que les longueurs aus repos ont été prises égales à 0). On trouve donc ici une interprétation mécanique de la semie-norme de Sobolev en dimension 2.

Réseaux résistif

On peut également interpréter la semi-norme de Sobolev comme la version continue d'une énergie dissipée au sein d'un réseau résistif (circuit électrique ou réseau de conduits pour un fluide visqueux). Cette approche est décrite dans la section 4.2, page 48.

On peut (voir section 22.3 ci-après) donner un sens à la partie L^2 de la norme en considérant que les points du réseau sont reliés directement à des points extérieurs portés au potentiel nul (ou pression nulle dans la cas d'un fluide).

8.2 Traces

La démarche de définition d'une *trace* dans un sens assez général peut se formaliser de la façon suivante, pour des fonctions définies sur un domaine de l'espace euclidien (voir plus bas pour une généralisation à d'autres situations).

On considère un domaine Ω de $\mathbb{R}^d$, et un espace vectoriel de (classes de) fonctions sur Ω noté H, muni d'une norme $\|\cdot\|$ qui en fait un espace de Banach. On suppose que H contient l'espace $\mathcal{D}(\Omega)$ des fonctions continues à support compact sur Ω. On note H_0 l'adhérence de $\mathcal{D}(\Omega)$ dans H.

Deux types de questions se posent de façon naturelle :

1. L'espace quotient (voir proposition 19.8, page 192) H/H_0 est-il trivial ou pas ? Question accompagnée d'une question subsidiaire dans le cas où l'espace quotient est trivial : *pourquoi* est-il trivial ? (nous préciserons le sens de cette interrogation plus loin).
2. Si cet espace (défini sans ambiguïté, mais de façon abstraite) n'est pas trivial, peut-on le décrire ? L'identifier à un espace de fonctions définies sur $\partial\Omega$?

Considérons tout de suite une autre situation, sorte de problème-jouet, qui nous permettra de préciser rapidement le sens et l'enjeu des questions précédentes. On considère maintenant que H est un sous-espace vectoriel de $\mathbb{R}^{\mathbb{N}}$, muni d'une norme qui en fait un espace de Banach. On note maintenant D le sous-espace des suites nulles au delà d'un certain rang. Pour $H = \ell^p$, avec $p \in [1, +\infty[$, l'espace quotient est trivial. Pour ℓ^∞, la situation est déjà plus riche, l'espace quotient contient en premier lieu les classes (distinctes) des suites constantes (ces classes s'identifient aux suites qui admettent une limite finie en $+\infty$). On peut en fait vérifier que l'espace quotient n'est pas séparable, alors que H_0 l'est dans ce cas : toute la richesse de l'espace est en fait "au bord" (comportement en $n \longmapsto +\infty$).

Considérons maintenant, pour $(\alpha_n) \in]0, +\infty[^{\mathbb{N}}$ donné, l'espace

$$H = \left\{ u = (u_n) \in \mathbb{R}^{\mathbb{N}}\,,\ u_0 = 0\,,\ \sum \alpha_n \left| u_{n+1} - u_n \right|^2 < +\infty \right\}, \tag{8.1}$$

muni de la norme naturelle associée à sa définition. Il s'agit d'un espace de Banach, et même d'un espace de Hilbert (isométrique à l'espace modèle ℓ^2).

Supposons en premier lieu que $\alpha_n \equiv 1$. On peut alors vérifier (voir proposition 8.1 ci-dessous) que D est dense dans H, donc que l'espace quotient est trivial : il n'y a "rien" en l'infini. Noter que $H = H_0$ ne signifie aucunement que toutes les suites seraient d'une certaine manière nulles en $+\infty$, c'est même plutôt *le contraire* : par exemple la suite $u_n = 1+1/2+\cdots+1/n$, qui tend vers $+\infty$, est dans H. On peut construire aussi très simplement [51] des suites qui tendent vers n'importe quelle valeur réelle en $+\infty$. Symétriquement, dans

51. On peut même avec un peu plus de travail construire des suites dans H dont l'ensemble des valeurs d'adhérences est $\mathbb{R}$ tout entier : c'est vraiment *n'importe quoi*.

ce contexte, il est tentant de dire que par exemple *la suite triviale identiquement nulle ne converge pas vers* 0, c'est à dire que, au vu de la norme définie sur les suites, il n'est pas licite de parler de sa valeur en $+\infty$ comme étant 0, puisqu'elle peut être approchée arbitrairement près par des suites qui ont un comportement très différent en $+\infty$.

Les remarques ci-dessus donnent une première réponse informelle au *pourquoi ?* de la première question au début de cette section : l'espace quotient est trivial parcequ'il est impossible de définir la limite d'une suite de H en $+\infty$.

On peut montrer a contrario que, si la suite des α_n croît suffisamment vite, l'espace quotient est non trivial. On a plus précisément :

Proposition 8.1. *Soit H l'espace défini par* (8.1)*, et H_0 l'adhérence de D (sous espace des suites nulle au delà d'un certain rang). On a*

$$\sum \frac{1}{\alpha_n} < +\infty \Longrightarrow H/H_0 \simeq \mathbb{R}\,,\ \sum \frac{1}{\alpha_n} = +\infty \Longrightarrow H/H_0 \simeq \{0\}.$$

Démonstration. Supposons dans un premier temps que la série des $1/\alpha_n$ converge (vers la valeur $1/\alpha > 0$). Remarquons en premier lieu que, pour tout $u \in H$, tous $p < q$,

$$|u_q - u_p| \ \leq \sum_{k=p}^{q-1} |u_{k+1} - u_k| \ = \sum_{k=p}^{q-1} \frac{1}{\sqrt{\alpha_n}} \sqrt{\alpha_n}\, |u_{k+1} - u_k| \ \leq \left(\sum_{k=p}^{q-1} \frac{1}{\alpha_n} \right)^{1/2} \left(\sum_{k=p}^{q-1} \alpha_n \, |u_{k+1} - u_k| \ ^2 \right)^{1/2},$$

qui tend vers 0 quand p et q tendent vers $+\infty$: la suite est de Cauchy, donc converge vers une valeur réelle. On note φ la forme linéaire qui à une suite de H associe sa limite. On a

$$|u_n| = |\ u_n - u_{n-1} + u_{n-1} - \cdots - u_0 + u_0| \leq \left(\sum \frac{1}{\sqrt{\alpha_n}} \right)^{1/2} \left(\sum \alpha_n \, |u_{n+1} - u_n|^2 \right)^{1/2} \leq \frac{1}{\alpha} \, \|u\|_H \,.$$

Il s'agit donc bien d'une forme linéaire continue, de norme ≤ 1.

Cherchons maintenant à identifier l'orthogonal de H_0. Tout suite h dans cet orthogonal est telle que la quantité $\alpha_n(h_{n+1} - h_n)$ est constante (h est *harmonique* au sens discret). On note q cette constante, on a

$$h_n = \sum_{k=1}^{n} (h_k - h_{k-1}) = q \sum_{k=1}^{n} \frac{1}{\alpha_{k-1}} \longrightarrow \frac{q}{\alpha},$$

de telle sorte que h est entièrement déterminée par sa limite quand n tend vers ∞.

Considérons maintenant la situation où la série des $1/\alpha_n$ diverge, et montrons que toute suite u de H peut être approchée par une suite de D, ce qui assurera la trivialité de H/H_0 (absence de trace). Pour $u \in H$ donné, on construit u^N de la façon suivante : u^N_n est égal à u_n pour $n \leq N$, et u^N_n décroît (ou croît si u_n est négatif) vers 0 entre N et un indice $M > N$ que nous fixerons ultérieurement. La suite u^N ainsi construite est dans D On impose

$$\alpha_n(u^N_{n+1} - u^N_n) = q$$

constant pour n entre N et $M-1$. On a donc

$$u_N = u^N_N = u^N_N - u^N_{N+1} + \cdots - u^N_{M-1} + u^N_{M-1} - u^N_M = q \sum_{n=N}^{M-1} \frac{1}{\alpha_n} = q r_{NM}.$$

On a donc

$$\sum_{n=N}^{M-1} \alpha_n (u_{n+1}^N - u_n^N)^2 = q^2 r_{NM} = (u_N)^2 \frac{1}{r_{NM}}.$$

Par divergence de la série, $1/r_{NM}$ peut être rendu arbitrairement petite, on choisit par exemple $M = M(N)$ tel que $(u_N)^2/r_{NM} < 1/N$. On a ainsi convergence de u^N vers u pour la norme de H.

□

Comme suggéré précédemment, on peut avoir trivialité de l'espace quotient pour des raisons différentes. Considérons par exemple, sous l'hypothèse $\sum 1/\alpha_n < \infty$, l'espace

$$H = \left\{ u = (u_n) \in \mathbb{R}^{\mathbb{N}}\,,\ u_0 = 0\,,\ \sum u_n^2 + \sum \alpha_n \left| u_{n+1} - u_n \right|^2 < +\infty \right\}. \tag{8.2}$$

L'espace D des fonctions nulles au delà d'un certain rang est dense dans H, l'espace quotient H/H_0 est donc trivial. La situation est pourtant très différente du cas d'absence de trace de la proposition précédente : ici, on peut définir d'une certain manière une trace (les suites de H sont de Cauchy d'après la partie différentielle de la norme), mais cette trace est nécessairement nulle.

Interprétation en termes de modélisation

Les espaces de suites définis ci-dessus peuvent s'interpréter de la façon suivante : on considère une infinité de fils électriques, de résistances r_1, ..., r_n, ..., mis bout à bout. On note $\alpha_n = 1/r_n$ la conductivité du fil n. Pour faciler la représentation mentale d'un fil global qui possède bien 2 bouts (en 0 et en $+\infty$), on pourra imaginer que les longueurs des fils forment une série convergente, et que l'on peut ainsi identifier la chaîne à un fil de longueur finie, que l'on peut plonger dans l'espace euclidien.

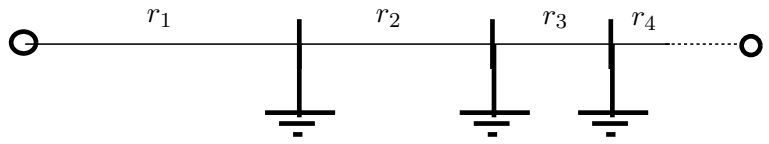


FIGURE 8.1 – Réseau linéaire semi-infini

On note u_n et u_{n+1} les potentiels électriques aux extrémités du n-ième fil, on a par hypothèse un potentiel nul à l'extrémité 0. La question qui se pose est de savoir s'il cela a un sens d'imposer un potentiel non nul U à l'extrémité ∞. Pour le fil tronqué à N bouts, on s'intéresse à la minimisation de

$$\sum_{n=1}^{N} \alpha_n \left| u_n - u_{n-1} \right|^2 = \sum_{n=1}^{N} \frac{1}{r_n} \left| u_n - u_{n-1} \right|^2,$$

avec valeurs imposées 0 et U aux extrémités. Le minimum est atteint en une collection u de potentiels unique, tels que

$$q_n = \alpha_n (u_n - u_{n-1}) = q$$

est constant. Cette quantité q correspond à l'intensité électrique qui traverse le fil, et la somme ci-dessus vaut

$$\sum_{n=1}^{N} \frac{1}{r_n} |u_n - u_{n-1}|^2 = \sum_{n=1}^{N} r_n |q_n|^2 = \underbrace{\sum_{n=1}^{N} r_n}_{=R_N} |q|^2 ,$$

qui exprime la puissance dissipée (effet Joule). L'appartenance à l'espace H exprime le fait que le courant électrique généré par les potentiels (u_n) induit une puissance dissipée finie. On prendra garde au fait que H contient des potentiels *non harmoniques*, i,e. tels que les intensités peuvent varier d'un segment à l'autre : la loi des nœuds n'est pas vérifiée, de l'intensité peut rentrer ou sortir du domaine par les points de jonction, mais sans induire de puissance dissipée supplémentaire (voir ci-après une situation qui pénalise énérgétiquement ces fuites). Le cas correspondant à $\alpha_n \equiv 1$ exploré précédemment correspond ici plus généralement à $R = \sum r_n = \sum 1/\alpha_n = +\infty$: la résistance globale du fil "infini" est infinie, ce qui signifie qu'il est impossible de faire passer une intensité non nulle dans le fil en dissipant une quantité finie d'énergie. Si l'on reprend le fil tronqué précédemment, il apparaît que, quel que soit le potentiel U imposé en sortie, l'intensité tend vers 0 quand N tend vers $+\infty$. on a aussi convergence simple vers 0 de toutes les potentiels ponctuels. Pour le fil infini, la conséquence est que l'on peut imposer n'importe quel potentiel à l'extrémité $+\infty$ sans qu'il se passe quoi que ce soit. L'extrémité ∞ est isolante : le potentiel imposé n'est pas *vu* par le système. Cette situation correspond au cas d'un espace-quotient trivial (pas de trace), avec valeur au bord quelconque.

La situation qui correspondrait au cas alternatif d'un espace quotient trivial par nullité forcée des champs au bord peut être construite comme suit : on considère maintenant un fil infini de résistance globale finie, en supposant $\sum r_n = \sum 1/\alpha_n < +\infty$. On a alors $H/H_0 \neq \{0\}$, cet espace s'identifie à $\mathbb{R}$, ce qui signifie que cela a un sens d'imposer un potentiel non nul en ∞ (il s'agit en fait d'un problème de *Dirichlet discret*). Considérons maintenant que chaque point de jonction soit lui même relié à la terre (potentiel nul) par un fil de résistance unitaire. La puissance dissipée par effet Joule dans l'un de ces fils transverses est $\alpha_n(u_n - 0)^2$. L'espace d'énergie du problème (ensemble des potentiels qui induisent une puissance dissipée finie) est maintenant défini par l'équation (8.2). On retrouve la situation l'un espace quotient nul, mais pour une raison bien différente : le potentiel en ∞ est nécessairement nul. Plus précisément, imposer un potentiel non nul induirait une puissance dissipée infinie.

Remarque 8.2. *Cette construction peut se faire dans un cadre mécanique, en considérant un système mécanique constitué d'une infinité de ressorts. Les potentiels sont alors remplacés par des déplacements, les intensités par des forces, et les conductances α_n par des constantes de raideur. Un tel système mécanique sans trace est alors localement infiniment mou (on peut déplacer le "point" du bord infiniment facilement, ou alors (dans le cas où l'on attache les points de jonction, simplement reliés entre eux dans le premier cas, à un support fixe) infiniment raide (il est impossible de déplacer le point au bord avec une énergie finie).*

Nous avons abordé la première des deux questions initiales, qui portait sur la possibilité de stucturer de façon non triviale le comportement des fonctions (ou des suites) au bord du domaine. Comme le suggère l'exemple des suites, c'est une certaine rigidité de la norme lorsque l'on s'approche du bord qui conduit au fait que l'espace quotient n'est pas trivial. Dans le cadre de la proposition 8.1, c'est dans le cas où les α_n croissent suffisamment (donc rigidifient la suite en pénalisant l'écart entre valeurs successives) que l'on peut identifier un

espace de trace non trivial. La seconde étape consiste à décrire cet espace quotient non trivial, par exemple en l'identifiant à un espace de fonctions qui vivent sur la frontière du domaine. Nous allons voir que c'est maintenant une certaine forme de *rigidité transverse* de la norme qui va conditionner le comportement des objets au bord du domaine.

Dans le cas des suites, la situation est évidemment assez pauvre, puisqu'il n'y a qu'un point à l'infini, on ne peut donc trouver que $\mathbb{R}$ ou $\{0\}$. On peut néanmoins se faire une première idée de cette notion de rigidité transverse en considérant un réseau de fils électrique en forme d'échelle semie-infinie (voir figure 8.2), et en définissant l'espace de potentiels aux nœuds de ce réseaux qui correpondent à une puissance dissipée finie. On note $\alpha_n = 1/r_n$, et l'on définit

$$H = \left\{ u = (u_n^1, u_n^2)\,,\ u_0^1 = u_0^2\,,\ \sum \alpha'_n \left| u_n^2 - u_n^1 \right|^2 < +\infty\,,\ \sum \alpha_n \left| u_{n+1}^i - u_n^i \right|^2 < +\infty\,,\ i = 1\,,\ 2 \right\}$$

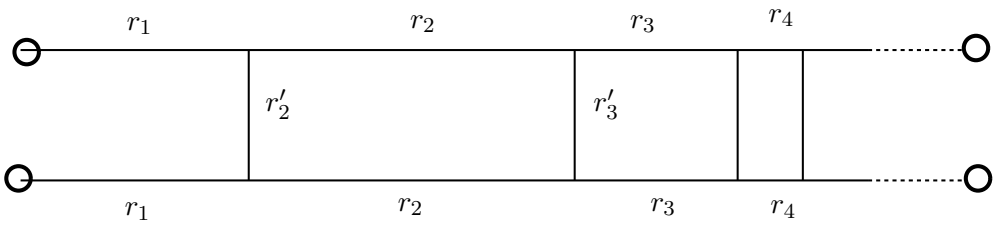


FIGURE 8.2 – Réseau semi-infini

On suppose que la série des inverses des α_n converge (ce qui revient à dire ici que la résistance de chacun des "rails" est finie). Pour tout u dans H, les suites (u_n^1) et (u_n^2) sont de Cauchy, donc convergent vers des valeurs U_1 et U_2. Si les α'_n sont nuls (résistances r'_n infinies), les deux rails sont indépendants, et l'on a un espace de trace H/H_0 qui s'identifie à $\mathbb{R}^2$. Maintenant considérons par exemple que les α'_n sont minorés (les résistances tranverses sont majorées). Alors les deux suites de Cauchy précédentes sont nécessairement adjacentes, et les limites sont donc les mêmes. On peut donc avoir H/H_0 de dimension 1 ou 2, selon la *rigidité transverse* induite par les conductances α'_n. Si l'espace est de dimension finie comme ici, le problème se ramène à déterminer sa dimension, et éventuellement à identifier une norme naturelle sur cet espace.

Dans le cas de fonctions définies sur un domaine euclidien, ce qui joue le rôle des deux "bouts" est une variété (le bord de Ω), ou par exemple les directions vers l'infini si Ω est l'espace entier. Les deux valeurs aux bouts sont remplacées par une fonction qui vit sur cette variété. On pourra alors retrouver le cas H/H_0 trivial sous deux formes : la situation d'une trace indéfinie (on peut avoir essentiellement n'importe quelle fonction au bord), ou la situation de fonction nécessairement nulle. Cette propriété dépendra de la rigidité de la norme quand on s'approche du bord. Pour le cas $H/H_0 \neq \{0\}$, selon l'importance de la rigidité transverse, on pourra retrouver le cas où la fonction est nécessairement constante, ou des cas extrêmes pour lequel la fonction ne présente par de régularité particulière, mais aussi des situations intermédiaires dans lesquels la rigidité transverse impose une certaine régularité aux traces, qui s'exprime par exemple dans le cas où H est l'espace de Sobolev $H^1(\Omega)$, sous la forme d'une régularité Sobolev fractionnaire $H^{1/2}$ en l'occurrence, pour un bord régulier.

9 Diffusion et hétérogénéité

9.1 Considérations générales

Une substance qui diffuse à tendance à se répartir uniformément dans l'espace disponible. Cette tendance est exprimée par exemple par la convergence (uniforme) vers 0 du noyau de la chaleur (équation (1.4), page 18) quand t tend vers $+\infty$. Dans le cas d'un domaine borné, une solution de l'équation de la chaleur, interprétée comme une courbe de densités de probabilité, voit son entropie [52] diminuer au cours du temps (voir (10.4)). Au niveau discret la suite de densités de probabilité associée au processus de diffusion associé à un réseau résistif voit de la même manière son entropie relative par rapport à la mesure stationnaire décroitre (voir section 4.3). Ce type de processus ne conduit ainsi pas spontanément à la création de "formes" non triviales, mais au au contraire à un étalement de la matière vers une situation d'homogénéité maximale. Sans rentrer dans des considérations philosophiques qui dépasseraient le cadre de cet ouvrage et les compétences de son auteur, il est même surprenant que le monde qui nous entoure, dont chaque sous-système fermé obéit au second principe de la thermodynamique, puissent être aussi hétérogène et empli de formes (*patterns*) non triviales, à différentes échelles. Nous nous proposons ici d'explorer quelques mécanismes impliquant un processus de diffusion couplé avec un ou plusieurs autres ingrédients, et qui conduisent de façon transitoire ou asymptotique à des distributions *hétérogènes* de matière, par opposition à l'étalement vers la distribution uniforme associé au processus de diffusion seul.

Le plus simple de ces mécanismes est basé sur l'ajout d'un terme de transport *concentrant* au terme de diffusion dans l'équation de la chaleur, encodant une tendance à se diriger dans une certaine direction. Considérons l'exemple d'un champ de vitesse u qui dérive d'un potentiel Ψ, l'équation de transport diffusion associée s'écrit

$$\frac{\partial \rho}{\partial t} - D\Delta\rho + \nabla \cdot (\rho u) = 0, \quad u = -\nabla\Psi.$$

Si Ψ est par exemple strictement convexe coercive, elle admet un minimum unique sur $\mathbb{R}^d$, et ce point de minimum est un attracteur global du système dynamique associé : toutes les trajectoires $\dot{x} = -\nabla\Psi(x)$ convergent vers ce même point. L'équation ci-dessus, appelée *équation de Fokker-Planck*, ajoute au transport de la diffusion, et ρ, que l'on peut voir comme une densité de probabilité, aura tendance à se concentrer autour de ce point de minimum d'autant plus que le coefficient de diffusion est faible. Dans le cas d'un potentiel quadratique $\Psi = |x|^2$, la densité limite sera une Gaussienne centrée en 0, de variance proportionnelle à D.

Cette équation est parfois utilisée pour modéliser le mouvement de particules ayant tendance à diffuser tout en ayant la faculté de se mouvoir préférentiellement dans la direction du gradient d'une certaine quantité (chimiotaxie). Dans ce contexte on écrira plutôt $u = \nabla\Psi$, où Ψ est par exemple une concentration en oxygène, et ρ une densité de bactéries dans un fluide. On peut penser aussi à une foule décrite de façon macroscopique par une densité, composée d'individus "agités" qui ont tendance à se diriger dans la direction d'un (ou plusieurs) point(s) d'intérêt commun(s).

Noter que cette équation qui semble coupler diffusion et transport peut s'interpréter dans certains contextes comme une équation de diffusion pure, dans un milieu hétérogène. Le

52. Dans le cas plus général d'une équation de diffusion et transport par un gradient, vers l'entropie relative par rapport à la mesure stationnaire (définie par (10.3)).

processus de diffusion associé à un réseau résistif décrit dans la section 4.3 est d'ailleurs l'équivalent discret de cette équation de Fokker Planck, le transport préférentiel selon certaines directions étant encodé par les variations locales de conductances, qui biaisent la marche aléatoire dans un sens ou dans l'autre. Le caractère essentiellement diffusif du phénomène sous-jacent à cette équation est aussi d'une certaine manière attesté (voir section 10.2) par la propriété de décroissance de l'entropie relative par rapport à la mesure stationnaire. Cette mesure stationnaire est la (à constante multiplicative près) solution de l'équation stationnaire, on retrouve son expression en remarquant que

$$-D\Delta\rho - \nabla\cdot\left(\rho\nabla\Psi\right) = -\nabla\cdot\rho\nabla\left(D\log\rho + \Psi\right) = -\nabla\cdot\rho\nabla\left(D\log\left(\frac{\rho}{\eta}\right)\right),$$

avec $\eta = Ce^{-\Psi/D}$ (mesure de Gibbs), qui est donc nul pour $\rho = \eta$.

Remarque 9.1. *Plus précisément, l'évolution peut être interprétée comme un flot de gradient pour cette fonctionnelle d'entropie relative si l'on se passe dans le cadre adapté de la métrique de Wasserstein (voir section 11). Noter que cette analogie peut s'étendre au niveau discret (équation* (4.7)*, page 51) grâce à l'introduction récente d'une métrique de type Wasserstein sur l'espace des mesures portées par les sommets d'un réseau résistif*[53]*.*

Le cadre précédent s'appuie sur une composante exogène (le potentiel Ψ). Nous explorons maintenant la possibilité de modéliser un phénomène d'agrégation *en boucle fermée*, en considérant des entités qui manifestent une tendance à se regrouper. Au niveau macroscopique, une écriture brutale de le principe, par exemple en considérant que le flux J (définition 1.1, page 11) est proportionel au gradient de ρ, ne conduit pas à un modèle pertinent, puisqu'il s'agit de l'équation de la chaleur *rétrograde*, qui est mal posée selon tous les cadres formels utilisables dans un contexte de modélisation. On obtient une situation plus riche et exploitable (et considérée comme représentant assez fidèlement certains phénomènes expérimentaux) en introduisant une quantité intermédiaire S, qui correspond à la concentration d'un *chimio-attractant* émis par les entités elle-mêmes, est dont le mouvement diffusif est complété par un biais dans la direction du gradient de cette nouvelle quantité. On obtient ainsi les équations de *Kelle-Segel*, développées dans la section 9.2 ci-après.

Une autre approche permet de reproduire des distributions non uniforme de matière, elle consiste à prendre en compte des mécanismes de réaction non linéaires afférents à la population considérée, voire à plusieurs populations coexistantes. L'équation de ce type la plus simple (en termes de modélisation tout du moins) est l'équation de Fisher KPP,

$$\frac{\partial\rho}{\partial t} - D\Delta\rho = k\rho(1-\rho/\rho_{max}),$$

qui conduit génériquement à l'apparition d'une zone pleine ($\rho\approx 1$) qui remplit progressivement l'espace, séparée d'une zone vide ($\rho\approx 0$) par une interface plus ou moins diffuse suivant la valeur du coefficient de diffusion. Ce modèle est présenté dans la section 9.3.

Une situation plus riche est obtenue lorsque l'on considère un terme source possédant 2 états d'équilibre stables, séparés par un état intermédiaire instable. Même si cette équation

53. Voir : J. Maas, Gradient flows of the relative entropy for finite Markov chains, Journal of Functional Analysis, 261(8), Pages 2250-2292 (2011).
`http://www.janmaas.org/papers/discrete.pdf`

est en général motivée par la modélisation de phénomènes de séparation de phase, on peut penser à une population ayant tendance à diffuser et à croitre (avec un terme de limitation logistique) lorsque la densité dépasse une certaine valeur critique. En dessous de cette valeur, la population tend à s'éteindre, et au dessus à croitre vers une valeur maximale (comme pour l'équation de Fisher KPP). L'équation avec diffusion s'écrit

$$\frac{\partial \rho}{\partial t} - D\Delta\rho = \rho(1-\rho)(\rho - a),$$

elle modélise une compétition entre les deux états stables (0 et 1), compétition équilibrée si $a = 1/2$, de sorte que, selon la distribution initiale, l'évolution peut conduire à la dispartion d'un des deux états, ou un coexistence entre les deux états. Si a est plus proche de 0 par exemple, le bassin d'attraction de 1 s'en trouve agrandi, et on peut vérifier que l'on a convergence vers l'état uniforme 1. Ce modèle est présenté dans la section 9.4.

Un autre point de vue a été apporté par Turing au début des années 50. Il a mis en évidence (par des arguments de stabilité linéaire) le fait que, si l'on considère deux populations réagissant entre elles de façons adaptée, un état d'équilibre au départ stable si l'on considère simplement le système différentiel représentant les interactions mutuelles, pouvait être déstabilisé, paradoxalement, par la prise en compte de mécanismes de diffusion de chacune des espèces en jeu, sous réserve que les coefficients de diffusion respectifs soient significativement différents. La section 9.5 détaille l'étude de stabilité permettant de mettre en évidence ce phénomène.

Les mécanismes évoqués ci-dessus conduisent à des formes variables, mais pour l'essentiel régulières, la situation la plus riche de ce point de vue correspondant aux instabilités de Turing, qui peuvent conduire à des distributions de motifs (textures) de type tâches ou rayure De tels modèles sont utilisés pour expliquer l'apparition de motis sur le pelage de certains animaux comme des félins (tigres ou léopards) ou des poissons [54].

L'apparition spontanée de motifs plus irréguliers persistants nécessite de faire appel à de nouveaux ingrédients, tout en gardant une place centrale à la diffusion. Précisons tout de suite que l'apparition spontanée de motifs en dendrites, ou en filament, ne peut reposer que sur des modèles moins bien posés que ceux considérés précédemment, et que l'analyse mathématique en est en général plus délicate. On retrouvera sous différentes formes un principe d'évolution très général, une sorte d'inverse de la loi de Fick à la base du processus de diffusion, qui consiste à renforcer l'hétérogénétité en faisant grandir ce qui est déjà grand, et diminuer ce qui est petit.

L'un des mécanismes conduisant à l'apparition de dendrites est connu sous le terme DLA (*Diffusion Limited Aggregation*). On peut décrire ce mécanisme très informellement de la façon suivante : on considère une première particule (on pourra se représenter ces particules commes des entités de taille finie) fixe. On considère une seconde particule qui se déplace de façon aléatoire (mouvement brownien ou marche aléatoire dans le cas discret) à partir d'une position initiale lointaine. Lorsque cette particule rencontre la première, elle se colle à elle. On fait ensuite partir une troisième particule, qui se collera à l'amas déjà formé dès le premier contact. Le mécanisme de croissance associé à ce principe présente la particularité suivante : si l'amas courant est de forme irrégulière, i.e. si son contour présente des creux et des bosses,

54. Voir par exemple : K. J. Painter, P. K. Maini, and H. G. Othmer, Stripe formation in juvenile Pomacanthus explained by a generalized Turing mechanism with chemotaxis, `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC21897/`

la particule suivante a plus de chance de le rencontrer au niveau d'une bosse que d'un creux. Le mécanisme tend donc à renforcer les irrégularités, à toute échelle.

Une déclinaison déterministe de ce principe, sous la forme d'un système d'équations aux dérivées partielles, est présentée dans la section 9.6.

9.2 Chimiotaxie, équations de Keller-Segel

On note $\rho(x,t)$ la densité locale d'une population d'entités mobiles (dans un fluide ou un substrat considéré lui-même comme immobile), animées d'une vitesse dirigée vers les zones les plus riches en une certaine substance (appelée *chemo-attractant*), dont on notera c la densité. Si l'on suppose que la vitesse est proportionnelle au gradient de S, que les entités sont également l'objet d'un processus de diffusion, et enfin que le chimio-attractant diffuse lui même au sein du milieu, on obtient le système

$$\begin{aligned} \partial_t \rho - D\Delta\rho + \nabla\cdot(\beta\rho\nabla S) &= 0 \\ \partial_t S - D_S \Delta S &= 0. \end{aligned}$$

Dans le cas (que l'on rencontre en pratique pour certaines bactéries) où la substance c est émise par l'entité elle-même[55], on obtient, en rajoutant un terme de disparition du chimio-attractant, le système dit de *Keller-Segel*

$$\begin{aligned} \partial_t \rho - D\Delta\rho + \nabla\cdot(\beta\rho\nabla S) &= 0 \\ \partial_t S - D_S \Delta S &= k\rho - \gamma S. \end{aligned}$$

On suppose que le flux (diffusif en l'occurrence) de chimio-attractant au travers de la frontière est nul :

$$D_S \frac{\partial S}{\partial n} = 0,$$

ainsi que celui de ρ. Ce flux s'écrit a priori

$$-D\partial\rho/\partial n + \beta\rho\nabla S\cdot n,$$

mais, du fait de la condition de Neuman sur S, ce flux est purement diffusif, ce qui conduit là aussi à une condition de Neumann homogène

$$D\frac{\partial \rho}{\partial n} = 0.$$

Remarque 9.2. *Noter que, dans le cas extrêmes ou cette attraction entre congénères ne passerait pas par l'intermédiaire d'une substance émise et ressentie, on aurait une vitesse chemotactique du type $u = \beta\nabla\rho$, ce qui conduirait à phénomène d'*anti-diffusion*, et à une équation de la chaleur rétrograde :*

$$\partial_t\rho + \beta D\Delta\rho = 0$$

55. On peut penser à la *phéromone* émise par les fourmis, qui leur permet de choisir préférentiellement les parcours déjà empruntés par leurs congénères.

qui est mal posée. Le système de Keller Segel, en prenant en compte de façon plus souple cette attraction entre entités, traduit une tendance à la concentration qui pose de fait des problèmes en termes de régularité de la solution, tout en permettant du fait du retard, que des solutions régilières puissent exister, au moins localement en temps. On peut en particulier montrer[56] *que, sous certaines conditions, le système conduit à l'apparition (en temps fini) de points de concentration (masses de Dirac).*

Si l'on suppose que la diffusion du chimio-attractant de concentration c est quasi instantanée par rapport aux autres phénomènes (i.e. si D_c est grand), on peut remplacer, ce qui est souvent fait en pratique, la seconde équation par une équation statique

$$\gamma S - D_S \Delta S = k\rho.$$

On obtient alorsle système de Keller -Segel dit *parabolique-elliptique* (par opposition au système de départ, appelé *parabolique-parabolique*.

Conservation. On peut vérifier la conservation de ρ sur l'ensemble du domaine (ce qui n'est pas une surprise, puisque les équations expriment précisément cette conservation sur tous les sous domaines) en intégrant l'équation en ρ :

$$\frac{d}{dt}\int_\Omega \rho + \int_\Gamma \underbrace{(-D\partial\rho/\partial n + \beta\rho\partial S/\partial n)}_{=0} = 0.$$

Pour S, on a l'équation de bilan

$$\frac{d}{dt}\int_\Omega S + \int_\Gamma \underbrace{(-D_S\partial S/\partial n)}_{=0} = k\int_\Omega \rho - \gamma\int_\Omega S,$$

qui exprime la variation de la quantité totale de chimio-attractant comme le bilan entre la création et la disparition naturelle.

Exercice 9.1. Proposer des modifications au modèle de Keller-Segel, fondées sur le prise en compte de phénomènes réalistes, qui pourraient empêcher (ou au moins retarder) le phénomène de concentration.

9.3 Équation de Fisher KPP

Pour désigner les facteurs qui sont de nature à limiter la croissance d'un population, comme la prédation, la limitation des ressources en nourriture, on utilise le terme d'*effets logistiques*. Dans le contexte des équations différentielles ordinaires, lorsque l'on décrit une population par sa seule taille, la manière la plus simple de les prendre en compte est de considérer un terme de croissance du type $\rho(1-\rho)$, qui exprime que le taux de croissance tend vers 0 lorsque ρ tend vers une valeur limite ici fixée à 1.

56. A. Blanchet, J. Dolbeault, B. Perthame, Two-dimensional Keller-Segel model : optimal critical mass and qualitative properties of the solutions,
Electronic Journal of Differential Equations 2006, (2006) 1–32,
`https://hal.archives-ouvertes.fr/hal-00021782`

Si l'on s'intéresse maintenant à une espèce distribuée non uniformément dans l'espace, soumise à un processus de diffusion, on aboutit à l'équation dite de Fisher KPP :

$$\frac{\partial \rho}{\partial t} - D\Delta\rho = k\rho(1-\rho/\rho_{max}),$$

où k correspond à un taux de reproduction à faible densité, et ρ_{max} est la capacité du milieu.

Pour ce modèle, l'équation différentielle associée présente deux états d'équilibre (en 0 et en ρ_{max}). Le premier est instable, le second est stable. La concentration aura donc tendance à tendre partout vers la valeur 1 correspondant à l'état stable.

Fisher KPP en dimension 1.
Il peut être intéressant de s'interroger sur la possible existence de solutions de type onde progressive, en dimension 1, pour l'équation de Fisher KPP

$$\frac{\partial \rho}{\partial t} - D\Delta\rho = k\rho(1-\rho/\rho_m).$$

Le second membre a été modifié par souci d'homogénéité, ρ_m est la densité maximale, dite *capacité* du milieu, et k est un taux de reproduction (homogène à l'inverse d'un temps) sous conditions optimales (à densité faible). On cherche une solution de la forme

$$\rho(x,t) = U(x-ct)\,,\ c>0\,,\quad \text{avec } U \geq 0,$$

de l'équation

$$\frac{\partial \rho}{\partial t} - D\partial_{xx}\rho = k\rho(1-\rho/\rho_m).$$

On a

$$-cU' - DU'' = kU(1-U/U_m),$$

d'où, en écrivant $u = U$, $v = U'$,

$$u' = v \tag{9.1}$$

$$v' = \frac{1}{D}\left(-ku(1-u/u_m) - cv\right). \tag{9.2}$$

Ce système $u = F(u)$ admet deux point d'équilibre, $(0,0)$ et $(u_m,0)$. Le gradient s'écrit

$$\nabla F = \begin{pmatrix} 0 & 1 \\ \dfrac{k}{D}\left(2u/u_m - 1\right) & -\dfrac{c}{D} \end{pmatrix}.$$

Les racines du polynôme caractéristique s'écrivent

$$\lambda = \frac{-c \pm \sqrt{c^2 - 4kD}}{2D} \quad \text{en } (0,0),$$

$$\lambda = \frac{-c \pm \sqrt{c^2 + 4kD}}{2D} \quad \text{en } (u_m,0),$$

Le point $(u_m,0)$ est donc instable (l'une des deux valeurs propres est réelle positive), alors que le point $(0,0)$ est stable. On a des trajectoires issues de $(1,0)$ (pour $x=-\infty$) qui vont converger exponentiellement vers le point stable $(0,0)$. Mais seules les trajectoires telles que u reste positif nous intéressent ici. Or, si $c < 2\sqrt{kD}$, les valeurs propres ont une partie imaginaire non nulle, de telle sorte que les trajectoires vont s'enrouler autour de l'origine, et u prendra des valeurs négatives. On a donc nécessairement $c \geq 2\sqrt{kD}$.

9.4 Équations d'Allen-Cahn

On considère ici une équation de réaction diffusion avec un terme source correspondant à deux états stables, par exemple en 0 et en 1, et un état instable pour une valeur a entre 0 et 1. Il s'agit de l'équation dite d'*Allen Cahn*, qui s'écrit

$$\frac{\partial \rho}{\partial t} - D\Delta\rho = \rho(1-\rho)(\rho - a).$$

Dans ce cas, si $a = 1/2$ (situation équilibrée entre les deux états stables), on peut avoir convergence vers une situation où co-existent les deux valeurs 0 et 1.

9.5 Motifs de Turing

Nous nous intéressons ici à des systèmes d'espèces en interaction, selon un système différentiel

$$\frac{du}{dt} = f(u,v) \tag{9.3}$$
$$\frac{dv}{dt} = g(u,v). \tag{9.4}$$

On suppose que ce système admet un point d'équilibre stable, que l'on fixe en $(0,0)$ (quitte à changer les fonctions f et g). On note F le champ définissant le système, et

$$\nabla F(0,0) = \begin{pmatrix} \partial_u f(0,0) & \partial_v f(0,0) \\ \partial_u g(0,0) & \partial_v g(0,0) \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

La stabilité du point d'équilibre se traduit par

$$a + d < 0\,,\ ad - bc > 0.$$

L'idée de Turing [57] a été de recherche la possibilité (a priori paradoxale) que rajouter de la diffusion en espace à un tel système pouvait *déstabiliser* le processus d'évolution. On s'intéresse donc au problème

$$\frac{\partial u}{\partial t} - \Delta u = f(u,v) \tag{9.5}$$
$$\frac{\partial v}{\partial t} - D\Delta v = g(u,v), \tag{9.6}$$

dans un domaine Ω, avec des conditions de Neuman homogènes.

On s'intéressera en particulier à la version linéarisée de ce problème :

$$\frac{\partial u}{\partial t} - \Delta u = au + bv \tag{9.7}$$
$$\frac{\partial v}{\partial t} - D\Delta v = cu + dv. \tag{9.8}$$

57. A. M. Turing, The Chemical Basis of Morphogenesis, Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, Vol. 237, No. 641. (Aug. 14, 1952), pp. 37-72
`http://www.math.u-psud.fr/~maury/ULM/PAPS/BIO_Turing.Turing52.pdf`

On s'intéressera au cas où la seconde espèce diffuse mieux que la première, i.e. $D > 1$. On introduit la base Hilbertienne (w_k) des fonctions propres du Laplacien avec conditions de Neuman :

$$-\Delta w_k = \lambda_k w_k.$$

En décomposant chacune des fonctions sur cette base :

$$u = \sum u_k(t) w_k(x)\,,\ v = \sum v_k(t) w_k(x),$$

et en prenant le produit scalaire avec l'une des fonction w_k, on obtient, du fait du caractère orthogonal de cette base,

$$\dot{u}_k = a u_k + b v_k - \lambda_k \tag{9.9}$$
$$\dot{v}_k = c u_k + d v_k - D\lambda_k. \tag{9.10}$$

qui peut s'écrire

$$\dot{Y} = \left(A - \lambda_k \begin{pmatrix} 1 & 0 \\ 0 & D \end{pmatrix} \right) Y.$$

L'étude de stabilité du mode w_k passe donc par la recherche des valeurs propres de la matrice ci-dessus, somme de A et de la matrice diagonale multipliée par $-\lambda_k$. On note que la trace de cette matrice,

$$a + d - (1 + D)\lambda_k,$$

reste négative. On aura donc un mode instable lorsque le déterminant est négatif (valeurs propres de signes opposés) :

$$\underbrace{ad - bc}_{>0} - \lambda_k(Da + d) + D\lambda_k^2.$$

Pour les grandes valeurs propres, le déterminant reste positif. mais pour des valeurs propres "petites" (en un sens à préciser), il est possible que ce déterminant devienne négatif si $a > 0$ et D est plus grand que 1. Noter que l'on a forcément $d < 0$ (pour que la trace non perturbée soit négative), et, du fait que le déterminant non perturbé est positif, i.e. $ad - bc > 0$, b et c doivent être de signes opposés. On peut alors avoir une plage de valeurs propres associés à des modes instables, qui peuvent expliquer l'apparition de "motifs".

Exemple. Un des exemples les plus simples est le suivant

$$\frac{\partial u}{\partial t} - \Delta u = u^2 v - u \tag{9.11}$$
$$\frac{\partial v}{\partial t} - D\Delta v = r(1 - u^2 v), \tag{9.12}$$

où r et D sont des paramètres. On a un point fixe $(1, 1)$, avec une matrice du problème linéarisé qui s'écrit

$$\begin{pmatrix} -1 + 2uv & u^2 \\ -2ruv & -ru^2) \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ -2r & -r \end{pmatrix}.$$

Pour $r > 1$, on a donc bien une trace $(1 - r)$ négative, et un déterminant $(= r)$ positif, donc possibilité d'instabilités.

9.6 Croissance Dendritique

On considère[58] une population d'entités (de type bactéries) soumises à un processus de diffusion non linéaire (paramètre $m > 1$ dans l'équation ci-dessous) lorsqu'elle est sous forme active (densité ρ). L'entité passe sous forme passive (densité $\overline{\rho}$) selon le taux μ. La forme active croît selon un taux qui dépend linéairement de la présence d'un nutriment, dont la concentration est notée S, qui diffuse lui même (linéairement) dans le susbtrat.

$$\begin{aligned} \frac{\partial \rho}{\partial t} - D\Delta\rho^{m+1} &= \kappa\rho S - \mu\rho \\ \frac{\partial S}{\partial t} - D_S\Delta S &= -\kappa\rho S \\ \frac{\partial \overline{\rho}}{\partial t} &= \mu\rho \end{aligned}$$

On notera que la variable $\overline{\rho}$ ne fait que stocker la quantité de ρ désactivé, elle n'est en fait utilisée que pour des raisons de représentation graphique : c'est la quantité totale d'entités $\rho + \overline{\rho}$ qui présente des motifs en dentrites.

58. Voir : S. Kitsunezaki, Interface Dynamics for Bacterial Colony Formation, *J. Phys. Soc. Japan* **66** (1997), 1544–1550. `http://www.complex.phys.nara-wu.ac.jp/~kitsune/paper/kitsune96B.ps`

10 Entropie

10.1 Entropie d'une variable aléatoire discrète

On considère une variable aléatoire discrète qui prend ses valeurs dans un ensemble de cardinal N. La loi de cette variable est décrite par

$$p = (p_1, p_2, \dots, p_N)\ ,\ p_i \geq 0\,,\ \sum p_i = 1.$$

Definition 10.1. *On définit*[59] *l'entropie de la loi discrète p comme*

$$S(p) = \sum p_i \log(p_i)$$

Dans ce contexte l'entropie est toujours négative, égale à 0 si et seulement si la variable est déterministe, et la valeur dans le cas uniforme $p_i \equiv 1/N$ est

$$S(p_u) = -\log N.$$

Montrons que cette valeur est un minimum. Pour toute fonction φ convexe, on a

$$\varphi\left(\frac{1}{N}\sum p_i\right) \leq \frac{1}{N}\sum \varphi(p_i),$$

d'où (avec $\varphi(a) = a\log a$),

$$S(p) \geq N\varphi(1/N) = -\log N.$$

L'entropie est donc minimale pour la loi uniforme, et seulement celle-là, et nulle dans les cas déterministe. Elle quantifie en effet l'information que la connaissance de la loi de probabilité donne sur le système.

Remarque 10.2. *On peut vérifier que cette entropie tend à diminuer pour un processus d'évolution de type diffusif*[60]*. Considérons par exemple une marche aléatoire sur un ensemble à N points, avec passages équiprobables aux points suivants et précédents, et périodicité. Notons ρ^n la loi de la position du point au temps n. A l'étape suivante, on a*

$$\rho_i^{n+1} = \frac{1}{2}\left(\rho_{i-1}^{n+1} + \rho_{i+1}^{n+1}\right).$$

On a alors

$$S(\rho^{n+1}) = \sum g(\rho_i^{n+1}) = \sum g\left(\frac{1}{2}\left(\rho_{i-1}^n + \rho_{i+1}^n\right)\right) \leq \frac{1}{2}\sum\left(g\left(\rho_{i-1}^n\right) + g\left(\rho_{i+1}^n\right)\right) = S(\rho^n),$$

pour toute fonction g convexe (en particulier $g(x) = x\log x$).

59. Dans ce contexte de théorie de l'information, on définit en général l'entropie comme l'*opposé* de cette quantité. Ce choix correspond à l'entropie thermodynamique, qui augmente toujours pour un système fermé, ce qui exprime le fait que le système évolue spontanément vers un état de désordre. On fait ici le choix de l'entropie mathématique, son opposé, qui aura tendance à décroître pour les systèmes fermés.

60. L'exemple proposé ici est un cas particulier d'une propriété plus générale de décroissance de l'entropie relative à la mesure stationnaire pour processus de markov diffusif, voir proposition 4.11.

Interprétation en termes de quantité d'information.
Dans le cas $N = 2^k$, et si l'on choisit le logarithme de base 2, on a $S_{min} = -k$, qui correspond au nombre de questions binaires qu'il faut poser pour localiser de façon sûre une valeur de x qui a été tirée selon la loi uniforme (avec une stratégie de dichotomie : est-elle dans la première moitié ? dans le premier quart de la première moitié ? etc ...). Dans le cas d'une probablité non uniforme, cette interprétation en terme de *bits* d'information est plus délicate. Considérons l'exemple de la distribution

$$p = \left(\frac{1}{2}, \frac{1}{2(N-1)}, \ldots, \frac{1}{2(N-1)}\right).$$

La variable a une chance sur deux de se trouver en première position, avec probablité uniforme sur le reste si ça n'est pas le cas. L'entropie de cette loi est

$$-\frac{1}{2} + \sum \frac{1}{2(N-1)} \log \frac{1}{2(N-1)} = -\frac{1}{2} - \frac{1}{2} - \frac{1}{2}\log(N-1) \approx -1 - \frac{k}{2}$$

si $N = 2^k$. Estimons maintenant le nombre de questions qu'il faut poser un moyenne pour localiser une variable suivant cette loi. On peut considérer un grand nombre de tirage de cette variable, avec à chaque fois la nécessité de la localiser en posant le minumum de questions binaires. La premiere question sera : est-elle en 1 ? cette question aura une réponse positive en moyennne une fois sur deux. Quand la réponse est négative, il faudra en gros k questions supplémentaires (dichotomie) pour la localiser. On a donc en moyenne

$$\frac{1}{2} + \frac{1}{2}(1+k) = 1 + \frac{k}{2}$$

qui correspond bien à l'opposé de l'entropie telle qu'on l'a définie.

Mesure de Gibbs. Un problème classique consiste à traduire sous forme de mesure de probabilité la connaissance marginale apportée par une information. Supposons par exemple que les états du système correspondent à des points de l'espace x_1, ..., x_N, et que l'on connaisse l'espérance β d'une certaine fonction f selon la loi p. On notera pour simplifier E_i la valeur de f en x_i, et $\bar{E}$ l'espérance. On s'intéresse alors au problème consistant à minimiser l'entropie $S(p)$ sous les contraintes

$$\sum_{i=1}^{N} p_i = 1\,,\ \sum_{i=1}^{N} p_i E_i = \bar{E}, \tag{10.1}$$

avec $\bar{E} \in]\min E_i, \max E_i[$. Notons que si $\bar{E}$ est égal à l'une des bornes de l'intervalle, par exemple $\max E_i$, alors p est concentré sur les indices qui réalisent ce maximum. S'il n'y en a qu'un, alors l'ensemble admissible est un singleton : le Dirac en ce point. S'il y en a plusieurs, le minimum de l'entropie sera la distribution uniforme sur le sous ensemble d'indices qui réalise le maximum. Bien entendu, si γ est à l'extérieur de l'intervalle fermé, alors l'ensemble admissible est vide.

Proposition 10.3. *On suppose $\bar{E} \in]\min E_i, \max E_i[$ et $N \geq 3$. L'entropie $p = (p_1, \ldots, p_N) \longmapsto S(p)$ admet un minimum unique sur $\mathbb{R}_+^N$, sous les contraintes* (10.1) *de la forme*

$$p_i = \frac{1}{Z} \exp\left(-\beta E_i\right).$$

Démonstration. Le minimum est atteint car la fonction est continue et l'ensemble admissible compact. L'unicité du minimiseur découle de la stricte convexité de la fonctionnelle. Si le minimiseur est atteint en un point de $]0,+\infty[^N$, alors on a

$$1 + \log p_i + \lambda_1 + \lambda_2 E_i = 0,$$

de telle sorte que p_i est de la forme

$$p_i = \frac{1}{Z} \exp(-\beta E_i).$$

On peut démontrer de deux manières que le minimum est bien de cette forme, ou bien en montrant que le minimum est bien atteint sur $]0,+\infty[^N$ (démonstration 1), ou alors en montrant qu'il existe bien un (p_i) de cette forme qui vérifie les contraintes, et en concluant par le théorème de Kuhn et Tucker. La deuxième démonstration est plus directe, mais la première utilise une démarche de calcul des variations praticable dans de nombreuses situations, nous développons donc ici ces deux approches.

Démonstration 1 : Supposons que le minimum ne soit pas dans $]0,+\infty[^N$, que par exemple $p_1 = 0$. S'il existe 2 indices i_1 et i_2 à poids > 0 (donc nécessairement < 1) associés à des valeurs de E_i distinctes, on considère une variation de p du type

$$h = \varepsilon\delta_1 + \varepsilon_1\delta_{i_1} + \varepsilon_2\delta_{i_2},$$

avec $\varepsilon > 0$. Les conditions pour que h soit admissible s'écrivent

$$\varepsilon_1 + \varepsilon_2 = -\varepsilon\,,\ E_{i_1}\varepsilon_1 + E_{i_2}\varepsilon_2 = -\varepsilon\beta_1.$$

Pour ε positif suffisament petit, il existe donc un unique couple $(\varepsilon_1, \varepsilon_2)$ tel que $p+h$ soit dans K. Comme la dérivée de $x \mapsto x \log x$ est $-\infty$ en 0, la variation effectuée dominue strictement l'entropie globale au voisinage de p, qui ne saurait donc être un minimiseur.

Si maintenant p charge un unique indice i (ou plusieurs indices associés à la même valeur de l'énergie), alors nécessairement il existe deux indices i_1 et i_2 tels que

$$0 < E_{i_1} < E_i < E_{i_2},$$

car $\bar{E}$ est dans l'intérieur de l'enveloppe convexe des E_i. (On a par ailleurs supposé que les E_i étaient positifs, ce qui ne nuit pas à la généralité du fait que l'on peut rajouter une même constante arbitraire aux E_i et à $\bar{E}$ sans changer la condition.) On considère alors une variation

$$h = -\varepsilon\delta_i + \varepsilon_1\delta_{i_1} + \varepsilon_2\delta_{i_2},$$

avec $\varepsilon > 0$. Les conditions pour que cette variation soit admissible s'écrivent

$$\varepsilon_1 + \varepsilon_2 = \varepsilon\,,\ E_{i_1}\varepsilon_1 + E_{i_2}\varepsilon_2 = \varepsilon E_i.$$

La valeur de $\varepsilon > 0$ étant fixée, le système ci-dessus admet une unique solution $(\varepsilon_1, \varepsilon_2)$, avec $\varepsilon_1, \varepsilon_2 > 0$, du fait que $E_i/E_2 < 1 < E_i/E_1$. La variation est donc admissible, et conduit pour les mêmes raisons que précédemment à une diminution stricte de l'entropie.

Démonstration 2 : Considérons la fonction

$$g\ :\ \beta \longmapsto \frac{\sum \exp(-\beta E_i) E_i}{\exp(-\beta E_i)}.$$

On a

$$g'(\beta) = \frac{-\left(\sum \exp(-\beta E_i) E_i^2\right)\left(\sum \exp(-\beta E_i)\right) + \left(\sum \exp(-\beta E_i) E_i\right)^2}{\left(\sum \exp(-\beta E_i) E_i\right)^2}$$

qui est strictement négatif d'après l'inégalité de Cauchy-Schwarz (si les E_i ne sont pas tous égaux, ce qui est le cas). La fonction g tend par ailleurs vers $\max E_i$ en $-\infty$, et vers $\min E_i$ en $+\infty$. L'équation $g(\beta) = \gamma \in]\min E_i, \max E_i[$ admet donc une solution unique. Le coefficient Z de normalisation est alors déterminé par

$$Z = \left(\sum \exp(-\beta E_i)\right)^{-1}.$$

Comme la fonction est convexe et le domaine convexe, la vérification des conditions de Kuhn et Tucker assurent que le p ainsi déterminé est bien le minimiseur de S sur l'ensemble admissible (Théorème 23.29, page 248). □

10.2 Entropie continue

Soit maintenant Ω un domaine de $\mathbb{R}^d$, et ρ une densité de probabilité définie sur Ω. On définit dans le même esprit son entropie par

$$S(\rho) = \int_\Omega \rho \log \rho \, dx.$$

On peut voir cette quantité comme une quantification de l'information que l'on a sur la position d'une variable aléatoire qui suit la loi associée à cette densité. Lorsque l'on a la densité uniforme $\rho \equiv 1/|\Omega|$ (absence complète d'information), on a

$$S(\rho) = \int_\Omega \frac{1}{|\Omega|} \log\left(\frac{1}{|\Omega|}\right) dx = -\log|\Omega|.$$

Conformément à l'intuition, cette valeur correspond à un minimum. En effet, pour toute fonction φ convexe, pour toute fonction g mesurable, l'inégalité de Jensen exprime que l'espérance par rapport à une mesure de proba μ de $\varphi \circ g$ est supérieure à φ de l'espérance de $g(x)$, i.e.

$$\varphi\left(\int_\Omega g(x)\, d\mu(x)\right) \leq \int_\Omega \varphi \circ g(x)\, d\mu(x).$$

On applique cette inégalité avec $d\mu = dx/|\Omega|$ (probabilité uniforme), $\varphi(a) = a \log a$, et $g(x) = \rho(x)$ pour obtenir

$$S(\rho) = |\Omega| \int_\Omega \rho \log \rho \frac{dx}{|\Omega|} \geq |\Omega| \frac{1}{|\Omega|} \log\left(\frac{1}{|\Omega|}\right) = -\log|\Omega|,$$

avec inégalité stricte dès que ρ n'est pas la mesure uniforme p.p.

Considérons maintenant l'équation de la chaleur dans le domaine Ω, avec condition aux limites de Neuman homogène (de façon à garder une masse 1 constante). On a

$$\frac{d}{dt}S(\rho) = \int_\Omega (1+\log\rho)\frac{\partial\rho}{\partial t} = \int_\Omega (1+\log\rho)\Delta\rho = -\int_\Omega \frac{1}{\rho}\nabla\rho\cdot\nabla\rho + \int_\Gamma \frac{\partial\rho}{\partial n}(1+\log\rho) \leq 0.$$

On trouve bien que l'entropie est décroissante. On notera qu'il en aurait été de même pour n'importe quelle fonction $S(\rho) = \int \varphi(\rho)$, avec φ convexe.

On considère l'équation d'évolution exprimant conjointement la diffusion et le transport par un champ de vecteur qui est l'opposé du gradient d'un potentiel Ψ :

$$\frac{\partial \rho}{\partial t} - D\Delta\rho + \nabla \cdot (\rho u) = 0, \quad u = -\nabla\Psi, \tag{10.2}$$

dans un domaine Ω borné, avec des conditions de bord qui assurent la conservation globale de la masse :

$$\partial\rho/\partial n = 0\,,\ u \cdot n = -\partial\Psi/\partial n = 0.$$

On peut l'écrire

$$0 = \frac{\partial \rho}{\partial t} - \nabla \cdot (D\nabla\rho + \rho\nabla\Psi) = \frac{\partial \rho}{\partial t} - D\nabla \cdot \rho\left(\frac{\nabla\rho}{\rho} + \frac{1}{D}\nabla\Psi\right) = \frac{\partial \rho}{\partial t} - D\nabla \cdot \rho\left(\nabla \log\left(\frac{\rho}{\pi}\right)\right),$$

avec $\pi = e^{-\Psi/D}$.

On obtient immédiatement que $\rho = \beta\pi$ est formellement solution stationnaire de l'équation. Si l'on se place dans le cas de condition de Neuman homogènes, avec un champ de vitesse tangent à la frontière, i.e. $u \cdot n = 0$, on a conservation de la masse totale, et $\beta\pi$ est bien solution stationnaire.

Vérifions que ρ tend bien vers cette mesure stationnaire en étudiant l'évolution de l'entropie relative de ρ par rapport à π :

$$S(\rho) = \int \rho \log\left(\frac{\rho}{\pi}\right). \tag{10.3}$$

On a

$$\frac{d}{dt}S(\rho) = \int (1 + \log\rho - \log\pi)\,\partial_t\rho = D\int (1 + \log(\rho/\pi))\left(\nabla \cdot \rho\left(\nabla \log\left(\frac{\rho}{\pi}\right)\right)\right)$$

$$= -D\int \rho\left|\nabla \log\left(\frac{\rho}{\pi}\right)\right|^2 + D\int_{\partial\Omega} (1 + \log(\rho/\pi))\,\rho\left(\nabla \log\left(\frac{\rho}{\pi}\right)\right) \cdot n.$$

Le terme de bord fait apparaître $\partial\rho/\partial n$ et $\partial\pi/\partial n$, qui sont tous les deux nuls. On obtient donc

$$\frac{d}{dt}S(\rho) = -D\int \rho\left|\nabla \log\left(\frac{\rho}{\pi}\right)\right|^2 \ \leq 0, \tag{10.4}$$

qui exprime la décroissance de l'entropie relative, décroissance stricte tant que ρ n'est pas proportionnel à la mesure stationnaire π.

11 Flots de gradient dans l'espace de Wasserstein

Cette section, très incomplète en l'état, décrit formellement la manière dont on peut interpréter certaines équations aux dérivées partielles comme des flots de gradient dans l'espace de Wasserstein. On se reportera à [7, 8] pour des développements plus approfondis des notions esquissées ici.

Le cadre mathématique usuel en modélisation est basé sur une vision eulérienne des choses : lorsque l'on considère une variation autour d'une fonction u, on a ajoute une perturbation v à u, et la mesure de l'éloignement est basé sur une mesure de cet ajout. Ainsi le gradient d'une fonctionnelle Ψ définie sur $L^2(\Omega)$ est le champ w qui vérifie

$$\Psi(u+\varepsilon v) = \Psi(u) + \varepsilon \int_\Omega wv + o(\varepsilon).$$

Faire varier u consiste donc à ajouter en chaque point x de Ω la quantité εv.

Cette approche très naturelle est pourtant biaisée : considérons sur l'intervalle $I =]0,1[$ une fonction ρ qui prend alternativement les valeurs 0 et 1 selon que l'on soit sur un sous-intervalle de type $]2k/2N, (2k+1)/2N$ ou $](2k+1)/2N, (2k+2)/2N$. Si l'on se place dans $L^2(I)$ (mais une démarche analogue pourrait être faire pour n'importe quelle distance "eulérienne", c'est à dire une distance basée sur la *différence* des fonctions), la distance entre ρ et $1-\rho$ est égale à la norme de ρ multipliée par $\sqrt{2}$. Elle reste donc de l'ordre de la norme de ρ même quand N tend vers $+\infty$. Or il est tentant de considérer les deux fonctions ρ et $1-\rho$ comme proches, selon deux points de vue. En premier lieu, leurs moyennes locales se rapprochent. Si l'on considère ces fonctions comme des images monodimensionnelles en niveau de gris (0 pour blanc, 1 pour noir), il est manifeste que toutes deux tendent (quand N tend vers $+\infty$) vers une image uniformément grise. Cette propriété peut se modéliser grâce à la notion de convergence faible, ou convergence au sens des mesures : ρ et $1-\rho$ tendent toutes deux vers la même mesure uniforme $1/2$. Une seconde manière de qualifier leur proximité, que nous allons développer dans ce qui suit, est la suivante : considérant ρ et $1-\rho$ comme des densités de matière sur l'intervalle $]0,1[$, on peut se demander s'il est coûteux de transporter l'une sur l'autre. Plus précisément, si l'on considère que le coût pour transporter une unité de matière d'un point x à un point y vaut une valeur prescrite $c(|y-x|)$ (fonction monotone de $|y-x|$, qui vaut 0 en 0), alors le coût total pour transporter ρ vers $1-\rho$ est de façon évidente $c(1/2N)/2$, qui tend bien vers 0 quand N tend vers $+\infty$. Nous privilégierons par la suite le coût quadratique $c(\alpha) = \alpha^2$, et nous définirons la distance associée comme la racine de ce coût, dont on peut vérifier qu'il s'agit effectivement d'une distance.

Pour définir la notion de flot gradient suivant cette approche, il nous faut définir ce que nous entendons par variation autour d'une densité donnée. Les développements qui suivent sont purement formels, en particulier nous supposons que tous les champs utilisés sont réguliers, et l'on pourra voir les mesures elles-mêmes comme des fonctions régulières. On se place dans $\mathbb{R}^d$, on considère une densité ρ donnée (positive) et un champ de vitesse w. Pour tout $\varepsilon > 0$ on considère l'application (ou *transport*)

$$T^\varepsilon \ : \ x \longmapsto x + \varepsilon w(x).$$

Pour ε assez petit (si w est lisse comme nous l'avons supposé), il s'agit d'une bijection régulière, est l'on peut définir ce que l'on appellera la mesure image, notée $\nu = T^\varepsilon_\sharp \rho$, comme la

mesure qui vérifie

$$\int f(T^\varepsilon(x))\rho(x)\,dx = \int f(y)\nu(y)\,dy,$$

pour toute fonction f régulière. La formule usuelle de changement de variable donne la valeur de la densité transportée en fonction du Jacobien de la tranformation :

$$T^\varepsilon{}_\sharp\rho(x+\varepsilon w) = \frac{\rho(x)}{|i+\varepsilon\nabla w|}\ .$$

Noter que, quand ε est petit (et si w est raisonnablement régulier), le jacobien de $i+\varepsilon\nabla w$ s'écrit $1+\varepsilon\nabla\cdot w+o(\varepsilon)$.

On notera que les variations considérées préservent la masse totale. De fait, cette approche conduit naturellement à considérer des familles de densités de masse totale fixée (la théorie est en général présentée pour des mesures de probabilité, donc de masse 1, mais la masse totale peut avoir une autre valeur).

Considérons maintenant une fonctionnelle Ψ dépendant de ρ. On appellera gradient [61] de Ψ en ρ un champ de vecteur v vérifiant

$$\Psi(T^\varepsilon{}_\sharp\rho) = \Psi(\rho) + \varepsilon\int v\cdot w\rho(x)\,dx + o(\varepsilon).$$

On écrira alors $v = \nabla^W\Psi(\rho)$, et l'on parlera de gradient au sens de Wasserstein, ou W-gradient.

La notion de flot gradient s'en déduit instantanément : on appellera flot gradient associé à Ψ une trajectoire de densités $t\mapsto\rho(\,\cdot\,,t)$ vérifiant l'équation de transport

$$\partial_t\rho + \nabla\cdot\,(\rho u) = 0,$$

où à chaque instant $u = -\nabla^W\Psi(\rho)$.

Flot potentiel. Considérons la situation où la fonctionnelle Ψ est donnée sous la forme

$$\Psi(\rho) = \int\varphi(x)\rho(x)\,dx.$$

On a

$$\Psi(T^\varepsilon{}_\sharp\rho) = \int\varphi(y)(T^\varepsilon{}_\sharp\rho)(y)\,dy = \int\varphi(x+\varepsilon w(x))\rho(x)\,dx = \Psi(\rho)+\varepsilon\int\nabla\varphi\cdot w\rho(x)\,dx,$$

de telle sorte que le gradient au sens où nous l'entendons maintenant s'identifie à $\nabla\varphi$. Le flot gradient associé correspond donc au transport par une vitesse $-\nabla\varphi$:

$$\partial_t\rho - \nabla\cdot(\rho\nabla\varphi) = 0.$$

61. On définit plus généralement la notion de sous-différentiel, qui correspond à l'ensemble des vecteurs v tels que

$$\Psi(\rho)+\varepsilon\int v\cdot w\rho(x)\,dx \le \Psi(T^\varepsilon{}_\sharp\rho)+o(\varepsilon),$$

pour des variations élémentaires du type $T^\varepsilon{}_\sharp\rho = i+\varepsilon w$. Cette notion permet de gérer des situations, non régulières, très courantes en pratique, où l'on ne peut pas définir le gradient au sens standard. La notion de flot gradient qui en résulte est basée sur l'appartenance du champ de vitesse u à l'opposé du sous-différentiel $\partial\Psi$ défini ci-dessus.

Considérons par exemple (pour $d = 1$) un potentiel $\varphi(x) = x^2$. Le champ de vitesse associé s'écrit $u = -2x$, donc les trajectoires sont des courbes $t \mapsto x(t) = x_0 e^{-2t}$. Le flot gradient au sens de Wasserstein aura donc tendance à concentrer la masse au voisinage de l'origine (on converge vers une masse de Dirac [62]). On peut vérifier aisément, sous réserve que l'on admette l'extension des ces notions aux cas de mesures non régulières, que si l'on prend comme condition initiale pour ρ une combinaison de masses de Dirac en différents points x_1^0, ..., $x_N^0 \in \mathbb{R}^d$, le W-flot gradient associé sera la somme des masses de Dirac affectées aux point $x_i(t)$, qui correspondent aux flots-gradient au sens usuel (euclidien)

$$\frac{dx_i}{dt} = -\nabla\varphi(x_i(t))\,,\ x_i(0) = x_i^0.$$

Ce flot gradient est donc une généralisation macroscopique des flots gradients ponctuels dans l'espace euclidien.

Remarque 11.1. *Noter que le flot gradient "eulérien" (dans L^2) se comporte de façon très différente. Dans le cas en dimension 1 évoqué ci-dessus, pour la fonctionnelle $\int x^2\rho(x)\,dx$, on a*

$$\Psi(\rho + \varepsilon\mu) = \int \varphi(x)(\rho + \varepsilon\mu)\,dx = \Psi(\rho) + \varepsilon \int x^2\mu(x)\,dx.$$

Le gradient au sens L^2 est donc la fonction $\varphi(x) = x^2$ elle-même. Le flot-gradient associé conduit donc à la trajectoire $t \mapsto \rho(x,t)$, avec $\rho(x,t) = \rho^0(x) - x^2 t$, qui n'a rien à voir avec le flot gradient euclidien associé à φ

Fonctionnelle d'énergie. Considérons maintenant le cas d'une fonctionnelle Ψ sous la forme suivante

$$\Psi(\rho) = \int \varphi(\rho(x))\,dx.$$

Cherchons à expliciter le W-gradient de la fonctionnelle (on suppose ici que les densités ne s'annulent pas) :

$$\begin{aligned}
\Psi({T^\varepsilon}_\sharp\rho) &= \int \varphi({T^\varepsilon}_\sharp\rho)(y)\,dy \\
&= \int \frac{\varphi({T^\varepsilon}_\sharp\rho)(y)}{{T^\varepsilon}_\sharp\rho(y)} {T^\varepsilon}_\sharp\rho(y)\,dy \\
&= \int \frac{\varphi({T^\varepsilon}_\sharp\rho)(x+\varepsilon w)}{{T^\varepsilon}_\sharp\rho(x+\varepsilon w)}\rho(x)\,dx.
\end{aligned}$$

Or la densité ${T^\varepsilon}_\sharp\rho(x+\varepsilon w)$ s'exprime à l'aide du Jacobien de la transformation

$${T^\varepsilon}_\sharp\rho(x+\varepsilon w) = \frac{\rho(x)}{|i + \varepsilon\nabla w|} = \rho(x)(1 - \varepsilon\nabla\cdot w + o(\varepsilon)).$$

62. De façon plus générale, pour une fonction régulière φ, le flot gradient aura tendance à concentrer la masse en des minimum locaux de la fonction, chacun concentrant la masse initialement présente dans son bassin d'attraction.

On obtient donc

$$\begin{aligned}
\Psi({T^\varepsilon}_\sharp\rho) &= \int \frac{\rho(x)(1-\varepsilon\nabla\cdot w+o(\varepsilon))}{\rho(1-\varepsilon\nabla\cdot w+o(\varepsilon)))}\rho(x)\,dx \\
&= \int \left(\varphi(\rho)-\varepsilon\rho\nabla\cdot w\varphi'(\rho)+o(\varepsilon)\right)\left(1+\varepsilon\nabla\cdot w+o(\varepsilon)\right)\,dx \\
&= \Psi(\rho)+\varepsilon\int\left(\varphi(\rho)-\rho\varphi'(\rho)\right)\nabla\cdot w\,dx+o(\varepsilon) \\
&= \Psi(\rho)+\varepsilon\int w\cdot\nabla\left(\rho\varphi'(\rho)-\varphi(\rho)\right)\,dx+o(\varepsilon) \\
&= \Psi(\rho)+\varepsilon\int w\cdot\left(\rho\nabla\varphi'(\rho)+\varphi'(\rho)\nabla\rho-\varphi'(\rho)\nabla\rho\right)+o(\varepsilon) \\
&= \Psi(\rho)+\varepsilon\int w\cdot\nabla\varphi'(\rho)\rho\,dx+o(\varepsilon),
\end{aligned}$$

ce qui permet de conclure que le W-gradient est $\nabla\varphi'(\rho)$.

Si l'on prend pour φ la fonction $\rho\mapsto\rho\ln\rho$, on obtient

$$u=-\nabla\varphi'(\rho)=-\frac{\nabla\rho}{\rho},$$

de telle sorte que le flot gradient associé à $\Psi=\int\rho\ln\rho$ vérifie l'équation de transport

$$\partial_t\rho-\nabla\cdot\rho\left(\frac{\nabla\rho}{\rho}\right)=\partial_t\rho-\Delta\rho=0,$$

c'est à dire l'équation de la chaleur.

12 Graphes

12.1 Définitions

Definition 12.1. *(Graphe orienté)*
*Un graphe orienté est défini par la donnée d'un ensemble V de sommets, et d'un ensemble d'*arcs *dans $V \times V$.*

Dans la définitions ci-dessus, les arcs sont *orientés* au sens où xy est différents de yx. Les deux peuvents être des arcs du graphe orienté, ou l'un des deux, ou aucun.

Definition 12.2. *(Cycle)*
On appelle cycle de (V, E) un n-uplet de sommets x_1, x_2,...,x_n (avec $n \geq 2$) tel que

$$(x_1, x_2) \in A\,,\ (x_2, x_3) \in E\,,\ \ldots (x_{n-1}, x_n) \in E\,,\ (x_n, x_1).$$

Definition 12.3. *(Graphe orienté acyclique)*
On dit que le graphe orienté (V, E) est acyclique s'il ne contient aucun cycle (Def. 12.2).

Théorème 12.4. *Soit (V, E) un graphe orienté acyclique fini. Il existe une numérotation des sommets compatible avec l'ordre partiel défini par le graphe, i.e.*

$$\exists \varphi \in \mathbb{N}^V\,,\ \textit{injective}\ ,\ (x, y) \in E \Longrightarrow \varphi(x) < \varphi(y).$$

12.2 Exemples

L'ensemble des utilisateurs (actifs ou non) de **Twitter** peut-être vu, à un instant donné, comme un graphe orienté, si l'on considère que tout "follower" pointe vers la personne qu'il suit.

Dans le même ordre d'idée, si l'on considère une **foule** à un instant donné, on peut voir chaque individu comme le sommet d'un graphe, qui pointe vers les personnes qui sont dans son cône de vision, et qui (si l'on s'en tient aux comportements sociaux, en excluant les contacts physiques) sont donc susceptibles d'influencer son comportement.

Si l'on considère un système d'équations différentielles exprimant l'évolution de concentrations d'**espèces chimiques** du fait des **réactions** entre les espèces, il est naturel de considérer le graphe dont les points sont les différentes espèces. Pour chaque espèce, on pointe vers les autres espèces (dont éventuellement elle-même) qui interviennent dans le seoncd membre de l'équation correspondante.

Une **chaîne alimentaire** peut aussi être considéré comme un graphe dont les points sont les espèces, chaque espèce pointant vers ses prédateurs.

On considère un système d'équations, impliquant n inconnues. On associe à ce système un graphe, considérant que chaque inconnue i pointe vers les inconnues qui apparaissent dans les équations impliquant i. Si le graphe est acyclique, on peut résoudre le système facilement en commençant par les éléments maximaux et en descendant la hiérarchie. Si le graphe contient

des cycles, on cherchera à transformer les équations (typiquement par élimination) de façon à obtenir un graphe acyclique.

Si l'on considère maintenant un **schéma** de type (pour fixer les idées) **différences finies**. On considère le graphe dont les nœuds sont les valeurs des inconnues aux pas de temps successifs, chaque nœud poitant vers le nœuds correspondant aux valeurs intervenant pour le calcul de la quantité concernée dans le schéma. Un schéma explicite sera typiquement acyclique, alors qu'un schéma implicite contiendra des cycles.

De façon générale, lorsque l'on s'intéresse à une collection d'*agents* (au sens le plus général), il est fécond de considérer le graphe d'*influence* associé, chaque agent pointant vers les agents qui l'influencent. Les modèles résultant d'une situation *acyclique* sont en général beaucoup plus simples à modéliser. Les éléments maximaux décident de ce qu'il font sans être influencés (d'un point de vue mathématique, il faudra donc décider de leur comportement, qui ne peut pas être donné par le modèle), et les effets se propagent dans la hiérarchie du réseau. Dans le cas où des cycles sont présents, la situaiton peut être beaucoup plus compliquée, générant en particulier des situations de non unicité. Cette situation se produira typiquement lorsque l'on s'intéresse à l'évolution d'une quantité afférente à chaque entité, qui dépend de l'*évolution* de la valeur instantanée de cette même quantité. Par exemple, dans le cas de foules, si l'on considère que chaque individu décide de sa vitesse en fonction de la position des personnes vers lesquels il pointe (i.e. qu'il voit), le problème pourra être bien posé même dans le cas cyclique. En revanche, si l'on considère que la vitesse d'une personne dépend aussi de la *vitesse* des gens qu'il voit, la présence de cycle va considérablement compliquer le problème, puisque le modèle n'est plus strictement *causal*. On pourra penser à l'exemple d'un cycle simple : deux personnes se font face, chacun souhaitant aller tout droit, en cherchant à décider de sa vitesse en fonction de la vitesse de l'autre.

Dans le contexte des schémas numérique pour les équations d'évolution, la présence de cycle dans les schémas implicite) nécessitera la résolution de systèmes linéaires (pour lesquels il faudra vérifier que la matrice associées est bien inversible). Dans le cas non linéaire, la présence de cycles peut invalider le caractère bien posé (en termes d'unicité, voire d'existence) du système à résoudre pour faire progresser l'algorithme de discrétisation en temps.

De façon générale, on pourra prendre en compte les paramètres du système, ou du modèle, comme des flèches pointant vers l'*extérieur* du graphe, vers un point abstrait qui représente l'ensemble des paramètres, que l'on peut voir comme un contrôle que l'on exerce sur le système. Dans le cas d'un graphe acyclique, une telle flèche ne permet, de façon évidente, de contrôler que les éléments qui sont inférieurs au point de départ de cette flèche dans la hiérarchie.

13 Convergence faible et compacité

Soient E et F deux e.v.n., et Ψ une forme bilinaire continue sur $E \times F$. On peut associer naturellement à Ψ une application (linéaire et continue) de F dans E' :

$$y \in F \longmapsto Ty \in E' \,,\ \langle Ty \,, x\rangle = \Psi(x,y) \quad \forall x \in E. \tag{13.1}$$

Proposition 13.1. *Soient E et F deux e.v.n. Si E est séparable*[63]*, alors de toute suite (y_n) bornée dans F on peut extraire une suite $(y_{n'})$ qui converge au sens suivant :*

$$\exists \varphi \in E' \,,\ Ty_{n'} \xrightarrow{\star} \varphi,$$

où T est définie par (13.1)*. Autrement dit, il existe $\varphi \in E'$ telle que*

$$\psi(x, y_{n'}) \longrightarrow \langle \varphi \,, x\rangle \quad \forall x \in E.$$

Démonstration. La suite extraire est construite par le procédé d'extraction diagonal de Cantor (voir preuve du théorème 20.32, page 204 dans le cas Hilbertien). □

On notera l'importance de la séparabilité de E dans la démonstration ci-dessus. Par ailleurs, le procédé construit une limite qui n'est pas un élément de F, mais une forme linéaire sur E', qui n'est pas nécessairement dans l'image de T.

La proposition précédente est très générale, et d'ailleurs très vide dans certains cas (prendre par exemple Ψ identiquement nulle, ou bien E de dimension finie alors que F est de dimension infinie). La propriété devient pertinente quand l'espace E et la forme Ψ sont suffisamment "riches" pour que la dualité soit *séparante*, c'est à dire (on privilégie ici l'espace E) que

$$\Psi(x,y) = 0 \quad \forall x \Longrightarrow y = 0.$$

Cette propriété assure l'*injectivité* de l'application T définie ci-dessus.

La richesse de l'espace F peut être formalisée par la condition symétrique de dualité séparante :

$$\Psi(x,y) = 0 \quad \forall y \Longrightarrow x = 0.$$

Si cette seconde condition est vérifiée, alors l'image de T est dense dans E' pour la topologie faible-$\star$ sur E' (i.e. en dualité avec E'). Dans le cas où E est réflexif, on aura bien densité de $T(F)$ dans E'. On prendra garde au fait que, si E n'est pas réflexif, on peut avoir E et F en dualité séparante sans que $T(F)$ ne soit dense dans E'. Considérer par exemple $E = \ell^\infty$, $F = \ell^1$, et Ψ la dualité canonique entre ces deux espaces. Elle est évidemment (doublement) séparante, mais $T(\ell^1)$ n'est pas dense dans ℓ^∞ : la forme linéaire qui à une suite de ℓ^∞ convergente associe sa limite, prolongée sur ℓ^∞ (par le théorème de Hahn-Banach analytique 19.1, page 191), est à distance au moins 1 de $T(\ell^1)$.

Corollaire 13.2. *Soit E un e.v.n. séparable. De toute suite bornée dans E' on peut extraire une sous-suite bornée qui converge pour la topologie faible-$\star$.*

63. Il admet une famille dénombrable dense.

On fera bien la distinction entre le corollaire précédent et le théorème de Banach-Alaoglu-Bourbaki, qui établit la compacité de la boule unité de E' pour la topologie faible-$\star$, sans hypothèse de séparabilité. Dans le cas où E n'est pas séparable, on a bien compacité, mais la topologie n'est *pas métrisable*, de telle sorte que la compacité ne peut pas se traduire en termes de suites extraites convergentes [64]. Ainsi la boule unité de ℓ^1 est bien compacte pour $\sigma(\ell^\infty, \ell^1)$, mais on ne peut par exemple extraire aucune sous suite convergente (faible-$\star$) de la suite (e_n).

Corollaire 13.3. *Soit E un espace de Banach dont le dual est séparable. De toute suite bornée dans E on peut extraire une sous-suite qui converge* [65] *dans E'' pour la topologie $\sigma(E', E'')$. Si E est réflexif, la sous-suite converge faiblement dans E.*

Dans le cas Hilbertien on peut supprimer la condition de séparabilité.

Corollaire 13.4. *Soit H un espace de Hilbert. De toute suite bornée dans H on peut extraire une sous-suite qui converge faiblement dans H*

Démonstration. Il suffit de se placer dans l'adhérence V de l'espace vectoriel engendré par les termes de la suite, qui est séparable par construction. On vérifie ensuite que l'on a bien convergence faible sur $H = V + V^\perp$ de la suite extraite. □

Espaces fonctionnels, mesures

On considère Ω un domaine de $\mathbb{R}^d$ (qui peut être l'espace tout entier).

Le corollaire 13.3 permet d'extraire d'une suite bornée une sous-suite faiblement convergente dès que l'espace considéré est réflexif, donc en particulier dans les espaces $L^p(\Omega)$ pour $1 < p < +\infty$, ainsi que dans les espaces de Sobolev $W^{m,p}(\Omega)$, pour tout $m \in \mathbb{N}$, tout $p \in]1, +\infty[$.

Pour les espaces non réflexifs (comme $L^1(\Omega)$ ou $L^\infty(\Omega)$, ou les espaces de Sobolev associés), la propriété est fausse en général, comme l'illustrent les exemples suivants.

Dans $L^1(\mathbb{R})$: la suite $f_n = \mathbb{1}_{]n,n+1[}$ est sur la sphère unité. Si une sous-suite converge faiblement vers f, alors f s'annule contre toute fonction régulière à support compact, elle est donc nécessairement nulle. Mais par ailleurs $\langle 1\ ,\ f_n\rangle$ est identiquement égale à 1, on doit donc avoir $\langle 1\ ,\ f\rangle = 1$, ce qui est impossible.

Dans L^∞, les choses sont un peu plus délicates, car le dual de cet espace n'est pas clairement identifié [66]. En particulier, le fait que l'on puisse (ou pas) extraire une sous-suite convergente de la suite définie précédemment n'est pas aisé à trancher. On peut néanmoins construire un contre-exemple analogue, en considérant par exemple la forme linéaire sur $L^\infty(\mathbb{R})$ qui à une fonction convergente en $+\infty$ associe sa limite, prolongée par le théorème de Hahn-banach analytique en $\varphi \in (L^\infty(\Omega))'$. On considère alors la suite $f_n = \mathbb{1}_{]n,+\infty[}$. Si elle

64. Autant dire qu'elle n'est pas commode à *utiliser*.

65. Plus précisément son image par la surjection canonique de E dans E''.

66. Montrer que le dual de L^∞ contient des formes qui ne peuvent pas se représenter par des fonctions de L^1 nécessite l'utilisation du théorème de Hahn-Banach analytique 19.1, page 191, donc indirectement de l'axiome du choix.

converge faiblement vers f, alors nécessairement f est nulle presque partout, donc tend vers 0 en $+\infty$, or on doit avoir $\langle \varphi \, , f \rangle = 1$, ce qui est absurde.

Convergence faible dans les cas non réflexifs. L'espace $L^\infty(\Omega)$ s'identifie au dual de $L^1(\Omega)$, qui est séparable, on peut donc, d'une suite bornée dans L^∞ extraire une sous-suite qui convergence (faible-$\star$) vers une limite de L^∞.

L'espace $L^1(\Omega)$, dont le dual L^∞ n'est pas séparable, peut être mis en dualité avec des espaces de fonctions continues (munis de la norme ∞) : espace C_c des fonctions continues à support compact, espace C_0 qui tendent vers 0 au bord de Ω, et l'espace C_b des fonctions bornées sur Ω. Noter que ces trois espaces s'identifient si l'on se place sur un compact. Dans le cas d'un domaine ouvert considéré ici, les 2 premiers espaces sont séparables, mais le troisième ne l'est pas. D'une suite bornée dans L^1 on pourra donc extraire une sous-suite qui converge vaguement (contre les fonctions de C_c) ou faiblement (contre les fonctions de C_0), mais la limite est définie comme une forme linéaire sur ces espaces, elle ne s'identifie pas forcément à une fonction de L^1 : il s'agit en toute généralité d'une mesure bornée. Par exemple la suite $f_n = n\mathbb{1}_{]0,1/n[}$ converge faiblement vers la masse de Dirac en 0. En l'occurrence, cette convergence est aussi étroite, mais on prendra garde au fait que l'on ne peut en général, d'une suite bornée de L^1, extraire une sous-suite qui converge étroitement (du fait de la non séparabilité de $C_b(\Omega)$). Ainsi la suite $f_n = n\mathbb{1}_{]n,n+1/n[}$ converge vaguement ou faiblement vers 0, il n'en existe aucune sous-suite qui convergerait étroitement.

Exercice 13.1. On considère l'espace E des fonctions continues sur $\mathbb{R}^d$ qui convergent vers une valeur finie lorsque $|x|$ tend vers $+\infty$. Montrer qu'il s'agit d'un espace complet (pour la norme ∞) séparable, et énoncer une propriété de compacité séquentielle faible-$\star$ pour $L^1(\mathbb{R}^d)$ mis en dualité avec E. Que peut on dire de la suite $f_n = n\mathbb{1}_{]n,n+1/n[}$ définie précédemment ? Proposer une généralisation de cette approche à des fonctions pour lesquelles la limite en $+\infty$ dépend de la direction $x/|x|$. (On pourra commencer par le cas $d = 1$, avec simplement 2 limites différentes en $+\infty$ et $-\infty$.)

14 Problème adjoint

Principe général.

On s'intéresse à un fonctionnelle qui dépend d'une variable de contrôle u par l'intermédiaire d'une variable d'état y, univoquement associée à u, i.e.

$$J(u) = G(y_u),$$

où y_u est reliée à u par une relation implicite

$$\Phi(y_u, u) = 0.$$

Les variables d'état y et de contrôle u vivent dans des espaces qui peuvent être de dimension infinie (il peut s'agir par exemple de fonctions de $[0, T]$ dans $\mathbb{R}^d$, comme on le verra plus loin).

On écrit la contrainte (lien entre u et y) de façon duale

$$\langle \Phi(y_u\ , u), p\rangle = 0,$$

pour tout p dans un espace en dualité séparante avec l'espace dans lequel vit $\Phi(y_u, u)$ (de façon à ce que l'identité ci-dessus implique $\Phi(y_u, u) = 0$). On introduit le Lagrangien

$$L(y, u, p) = G(y) + \langle \Phi(y\ , u), p\rangle,$$

qui est défini pour des couples (y, u) qui peuvent être indépendants (i.e. qui ne vérifient pas le lien $\Phi(y_u, u) = 0$). Pour tout y associé à u, le Lagrangien prend la valeur de la fonctionnelle, i.e.

$$J(u) = G(y_u) = L(y_u, u, p),$$

quel que soit p. On a alors

$$D_u J = D_y L \circ D_u y_u + (D_u \Phi)^\star p, \tag{14.1}$$

avec

$$D_y L = D_y G + (D_y \Phi)^\star p. \tag{14.2}$$

L'idée est alors de construire un p particulier qui annule $D_y L$, et donc le premier terme de (14.1). Il n'est donc pas nécessaire de connaître la différentielle de y_u par rapport à u : on obtient

$$DJ = (D_u \Phi)^\star p,$$

où p a été construit de façon à annuler $D_y L$ (expression donnée par (14.2)).

Contrainte statique linéaire. On considère ici le cas $y \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, et l'on cherche à minimiser

$$J(u) = \frac{1}{2} \left| C y_u - \bar{z} \right|^2,$$

où y_u est défini par

$$A y_u = B u,$$

avec $A \in \mathcal{M}_n(\mathbb{R})$, $B \in \mathcal{M}_{n,m}(\mathbb{R})$, $C \in \mathcal{M}_{p,n}(\mathbb{R})$, $\bar{z} \in \mathbb{R}^p$.

Le Lagrangien est défini par

$$(y,u,p) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n \longmapsto \frac{1}{2}\left|Cy - \bar{z}\right|^2 + (Bu - Ay) \cdot p.$$

On a dans ce cas

$$\nabla J = D_u L = B^T p$$

pour p solution du problème adjoint

$$A^T p = C^T \left(Cy - \bar{z}\right).$$

Problème adjoint dans le cas d'une EDO.

On considère l'équation différentielle suivante, dans $\mathbb{R}^n$,

$$\left\{ \begin{array}{rcl} \dot{y} & = & f(y,u,t) \\ y(0) & = & y_0 \end{array} \right. \tag{14.3}$$

où u est un paramètre de contrôle qui vit dans l'espace $U = \mathbb{R}^m$. On s'intéresse à la dépendence d'une fonction de y (et éventuellement de u lui même) vis-à-vis de la variable de contrôle u.

Démarche générale. On s'intéresse dans un premier temps au cas où la fonctionnelle mesure l'écart entre l'état final et un point cible donné :

$$J(u) = \frac{1}{2}\left|y_u(T) - \bar{y}_T\right|^2.$$

L'objectif est de calculer la différentielle de J.

On introduit le Lagrangien

$$L(y,u,p) = \frac{1}{2}\left|y(T) - \bar{y}_T\right|^2 + \int_0^T \left(f(y,u,t) - \dot{y}(t)\right) \cdot p(t)\, dt,$$

où p est une fonction définie sur $[0,T]$.

Lorsque y est associé à u par (14.3), on le note y_u. On a, pour tout u et tout p,

$$J(u) = L(y_u, u, p).$$

On prend la différentielle de cette identité :

$$DuJ = D_y L \circ D_u y_u + D_u L.$$

L'approche consiste à trouver un p particulier qui annule $D_y L$ (et donc le premier terme), de telle sorte qu'il ne sera pas nécessaire d'expliciter $D_u y_u$. La différentielle de J se réduira alors au second terme, qui s'écrira en fonction du p particulier

On a

$$\langle D_u L \, , \, \delta u \rangle = \int_0^T \left(\frac{\partial f}{\partial u}(y,u,t) \delta u \right) \cdot p,$$

où $\partial f / \partial u$ est linéaire de $\mathbb{R}^m$ dans $\mathbb{R}^n$. On peut identifier $D_u L$ à un vecteur de $\mathbb{R}^m$:

$$D_u L = \int_0^T \left(\frac{\partial f}{\partial u}(y,u,t) \right)^\star p.$$

Pour la différentielle par rapport à y, on réécrit tout d'abord le Lagrangien en intégrant par partie le second terme :

$$L(y,u,p) = \frac{1}{2} \left| y(T) - \bar{y}_T \right|^2 + \int_0^T \left(f(y,u,t) \cdot p(t) + y(t) \cdot \dot{p}(t) \right) \, dt - y(T) \cdot p(T) + y(0) \cdot p(0).$$

On a donc

$$\langle D_y L \, , \, \delta y \rangle = (y(T) - \bar{y}_T) \cdot \delta y + \int_0^T \left(\dot{p} + \frac{\partial f}{\partial y}(y,u,t) p(t) \right) \cdot \delta y - \delta y(T) \cdot p(T).$$

On introduit maintenant le problème adjoint, à valeur *finale* prescrite :

$$\left\{ \begin{array}{rcl} -\dot{p} & = & \dfrac{\partial f}{\partial y}(y,u,t) p(t) \\ p(T) & = & y(T) - \bar{y}_T. \end{array} \right. \tag{14.4}$$

Pour un tel p, $D_y L = 0$, et donc

$$DuJ = D_y L \circ D_u y_u + D_u L = D_u L = \int_0^T \left(\frac{\partial f}{\partial u}(y,u,t) \right)^\star p,$$

où p est solution de (14.4).

Fonctionnelle plus générale. On considère maintenant le cas

$$J(u) = \int_0^T F(y,u,t) \, dt.$$

Une démarche analogue conduit à

$$\langle DJ \, , \, \delta u \rangle = \int_0^T g(t) \delta u(t) \, dt,$$

ce qui permet d'identifier le gradient de J comme la fonction $t \mapsto g(t)$ donnée par

$$g = \frac{\partial F}{\partial u}(y,u,t) \, \frac{\partial f}{\partial u}(y,u,t) \, p(t),$$

où p est solution du problème adjoint

$$\left\{ \begin{array}{rcl} -\dot{p} & = & \dfrac{\partial f}{\partial y}(y,u,t) \, p(t) + \dfrac{\partial F}{\partial y}(y,u,t) \\ p(T) & = & 0. \end{array} \right. \tag{14.5}$$

15 Transport optimal (cas discret)

15.1 Problème d'affectation

Le problème d'affectation se formule comme suit :

Problème 15.1. *On considère* 2 *ensembles de même cardinal* $N \in \mathbb{N}$*, tous deux identifiés à* $\{1, \dots, N\}$*, et l'on se donne une collection de* coûts $c_{ij} \in \mathbb{R}$*. Le problème consiste à trouver une bijection* φ *qui minimise la quantité*

$$\sum_{i=1}^{N} c_{i\varphi(i)}.$$

Le problème ci-dessus ne présente pas d'intérêt théorique particulier : l'ensemble des bijections (groupe symétrique S_N) est fini, le problème admet bien (au moins) une solution. Mais la recherche effective de ce minimum peut extrêmement laborieuse, car le cardinal de l'ensemble des candidats croît comme $N!$.

15.2 Problème de Monge Kantorovich discret

Nous allons considérer une version relaxée de ce problème, qui peut se formuler intuitivement de la façon suivante, dans un contexte de transport : on considère le premier ensemble comme contenant des positions dans un certain espace (il n'est pas nécessaire de préciser lequel ici), et le second ensemble aussi comme une collection de positions dans un espace (éventuellement le même, mais pas forcément). On note c_{ij} ce que celà coûte de transporter une quantité de matière unitaire de x_i vers y_j. Le problème précédent consistant à considérer que l'on avait une même quantité de matière en chaque point (par exemple $1/N$), et que l'on cherchait à transporter cette matière vers le second ensemble en envoyant toute la matière de chaque point vers une destination unique. Nous allons considérer maintenant qu'il est possible de distribuer la matière venant d'un point vers plusieurs destination. Cette relaxation du problème permet de lever la contrainte d'avoir le même nombre de points au départ et à l'arrivée. Dans ce qui suit on notera γ_{ij} la quantité de matière allant de i vers j. On appellera $\gamma = (\gamma_{ij})$ un *plan de transport*.

Problème 15.2. *(Monge Kantorovich discret)*
On considère 2 *ensembles*[67] *finis* X *et* Y*, de cardinaux respectifs* N *et* $M \in \mathbb{N}$ *et l'on se donne une collection de coûts* $c_{ij} \in \mathbb{R}$*. On considère deux mesures de probabilités discrètes* μ *et* ν *sur* X *et* Y*, respectivement (*μ_i *est la masse portée par* i*, avec* $\sum \mu_i = 1$*, de même pour* ν*). On supposera tous les poids strictement positifs*[68]*. On cherche à minimiser le coût total*

$$C(\gamma) = \sum_{i,j} c_{ij}\gamma_{ij},$$

67. Il n'y a pas lieu de préciser ici les points d'arrivée et points de départ. Nous nous intéresserons plus loin au transport entre points d'un espace euclidien, mais ici on peut tout aussi bien effectuer un transport d'une essoreuse vers le concept de néant chez Sartre.

68. On peut toujours se ramener à cette situation en supprimant de X et / ou Y les points non chargés.

sous la contrainte que γ transporte μ vers ν, i.e.

$$\gamma_{ij} \geq 0\,,\ \sum_j \gamma_{ij} = \mu_i \quad \forall i\,,\ \sum_i \gamma_{ij} = \nu_j \quad \forall j, \tag{15.1}$$

ce que l'on écrira $\gamma \in \Pi(\mu, \nu)$, ou simplement $\gamma \in \Pi$ quand il n'y a pas d'ambiguïté.

Remarque 15.1. *On peut formuler ce problème en termes probabilistes, en considérant γ comme une loi de probabilité sur l'espace produit $X \times Y$, dont les mesures images par les projections sur X et Y sont respectivement μ et ν. Parmi de telles lois , on cherche celle(s) qui minimise(nt) l'espérance de la "fonction" $c = (c_{ij})$ sur $X \times Y$.*

Remarque 15.2. *L'ensemble admissible est non vide, il contient en particulier le plan correspondant à une loi de probabilité sur $X \times Y$ pour deux variables indépendantes, qui s'écrit*

$$\gamma_{ij} = \ \mu_i\ \nu_j.$$

On verra que c'est le plan qui minimise l'entropie de la loi γ (voir définition 10.1, page 101).

Proposition 15.3. *Le problème 15.2 admet un minimiseur.*

Démonstration. Les γ_{ij} sont positifs, et chacun d'eux est majoré par le max des μ_i, l'ensemble Π est donc borné, il est évidemment fermé donc compact : la fonction continue (car linéaire) $C(\,\cdot\,)$ admet donc un minimiseur sur Π. □

Remarque 15.4. *Dans le cas d'un coût du type $c_{ij} = a_i + b_j$, le problème est fortement dégénéré, puisque tout transport de μ vers ν réalise le même coût. Inversement, pour deux ensembles de même cardinal N, avec μ et ν lois uniformes sur X et Y, si l'on se donne une bijection φ de S_n, on peut construire une famille de coûts telle que le plan associé à la bijection*[69] *soit l'unique minimiseur, en prenant par exemple $c_{i\varphi(i)} = -1$, et $c_{ij} = 0$ si $j \neq \varphi(i)$.*

Question 15.1. (? ?)
Étant donnée une collection de coût (c_{ij}), existe-t-il des ensembles X et Y de points de $\mathbb{R}^d$ tels que $c_{ij} = |y_j - x_i|$? (on pourra aussi considérer $c_{ij} = |y_j - x_i|^p$, $c_{ij} = \psi(|y_j - x_i|)$ avec ψ croissante et nulle en 0.)

Question 15.2. (?)
Le problème 15.2 admet-il une solution unique "en général" ? (on s'attachera à exprimer précisément ce que l'on entend par unicité générique.)

Lien avec le problème d'affectation. Dans le cas où les cardinaux sont les mêmes, et les mesures équidistribuées, on peut préciser le lien entre le modèle relaxé basé sur les plans de transports et le problème d'affectation. Pour simplifier les notations, on considère ici la situation où chaque point porte une masse unitaire, de telle sorte que la masse totale des mesures considérées est égale au nombre de points.

Proposition 15.5. *On se place dans le cas $N = M$ (même nombre de points de part et d'autre, et $\mu_i = \nu_j \equiv 1$), et l'on note Π_S l'ensemble des plans de transports associés à une affectation, i.e. $\gamma_{ij} = \delta_{i\varphi(i)}$, où φ est une permutation du groupe symétrique. L'ensemble Π des plans de transport admissibles est l'enveloppe convexe de Π_S.*

69. C'est à dire : $\gamma_{i\varphi(i)} = 1/N$, et $\gamma_{ij} = 0$ si $j \neq \varphi(i)$.

Démonstration. Il s'agit d'une conséquence du théorème de Krein-Milman en dimension finie, qui assure que tout convexe compact d'un espace affine de dimension finie est l'enveloppe convexe de ses points extrêmaux [70]. Tout point de S_N est de façon évidente extrémal pour Π. Réciproquement, considérons un plan générique (i.e. qui n'est pas associé à une bijection) γ. On considère dans un premier temps les indices i pour lesquels γ_{ij} est nul pour tous les indices j sauf un (qui vaut donc 1). Cette sous-famille des points de départ est en bijection avec les points d'arrivées j correspondants, pour lesquels, symétriquement, γ_{ij} est nul pour tous les i sauf 1. On note I (resp. J) l'ensemble des indices non concernés dans l'espace de départ (resp. d'arrivée). Les ensemble I et J sont de même cardinal, et non vides par hypothèse. Avec des notations évidentes, la restriction du plan γ à $X_I \times Y_J$ est diffuse, au sens que pour tout i, $\gamma_{ij} \in]0,1[$ pour au moins 2 indices $j \in J$, et pour tout $j \in J$, on a $\gamma_{ij} \in]0,1[$ pour au moins 2 indices $i \in I$. On part d'un indice $i_0 \in I$, et l'on choisit j_0 tel que $\gamma_{i_0 j_0} > 0$. On choisit ensuite $i_1 \neq i_0$ tel que $\gamma_{i_1 j_0} > 0$, puis $j_1 \neq j_0$ tel que $\gamma_{i_1 j_1} > 0$. On construit ainsi une suite d'indices

$$i_0\,,\ j_0\,,\ i_1\,,\ \ldots\,,\ i_{n-1}\,,\ i_n,$$

que l'on peut voir comme un chemin dans le graphe sur $I \cup J$ associé au plan γ, chemin qui ne contient pas d'aller-retour. L'ensemble des indices étant fini, il existe forcément un n tel que i_n correspond à un indice $i_\ell \neq i_{n-1}$ déjà visité. On considère alors la variation

$$h = \sum_{k=\ell}^{n-1} \left(\pi_{i_k,j_k} - \pi_{i_{k+1},j_k}\right),$$

avec $i_n = i_\ell$, et où $\pi_{i,j}$ est l'élément de $\mathbb{R}^{NM}$ qui vaut 1 sur la composante (i,j), et qui est nul pour les autres couples. Pour η suffisamment petit, $\gamma \pm \eta h$ est positif, et par construction $\gamma \pm \eta h$ vérife les contraintes de marginales, les deux perturbations sont donc dans $\Pi_{\mu,\nu}$, et γ est moyenne non triviale de ces deux plans de transport, il ne s'agit donc pas d'un point extrémal.

Les seuls points extrêmaux correspondent donc aux permutations. □

Proposition 15.6. *On se place comme précédemment dans la situation de mesures équidistribuées sur des ensembles de même cardinal. Le problème de Monge Kantorovich discret 15.2 admet au moins une solution dans S_N, i.e. une solution optimale du type permutation.*

Démonstration. D'après la proposition 15.3, le problème 15.2 admet un minimiseur γ. D'après la proposition 15.5, ce minimiseur s'écrit comme combinaison convexe de plans associés à des permutations φ_1, …, φ_K :

$$\gamma = \sum \theta_k \gamma^k$$

(on ne garde dans la somme ci-dessus que les termes non triviaux, de telle sorte que $\theta_k > 0$ pour tout k). Le coût étant linéaire, on a

$$C(\gamma) = \sum \theta_k C(\gamma^k).$$

Comme chaque $C(\gamma^k)$ est supérieur ou égal à $C(\gamma)$, et que $\theta^k > 0$ pour tout k, la combinaison convexe ci-dessus implique que $C(\gamma^k)$ est égal à $C(\gamma)$ pour tout k. Chaque permutation impliquée dans la combinaison réalise donc le minimum.

□

70. On dit que $\gamma \in \Pi \subset \mathbb{R}^d$ est point extrêmal de Π si $\gamma = (\gamma^1 + \gamma^2)/2$, avec $\gamma^1, \gamma^2 \in \Pi$, implique $\gamma^1 = \gamma^2 = \gamma$.

15.3 Formulation duale du problème de MK discret

La formulation duale du problème 15.2 est basée sur l'expression duale des contraintes de marginales :

$$\sum_j \gamma_{ij} = \mu_i \quad \forall i \quad \Longleftrightarrow \quad \sum_{i=1}^{N} p_i \left(\mu_i - \sum_j \gamma_{ij} \right) = 0 \quad \forall p \in \mathbb{R}^N,$$

et l'on exprime de même les contraintes de destination à l'aide de $q \in \mathbb{R}^M$. On introduit donc (conformément à la définition 23.27, page 246) le Lagrangien

$$(\gamma, p, q) \in V \times \Lambda \longmapsto \sum_{i,j} c_{ij}\gamma_{ij} + \sum_{i=1}^{N} p_i \left(\mu_i - \sum_j \gamma_{ij} \right) + \sum_{j=1}^{M} q_j \left(\nu_j - \sum_i \gamma_{ij} \right), \qquad (15.2)$$

avec $V = \mathbb{R}_+^{NM}$ et $\Lambda = \mathbb{R}^N \times \mathbb{R}^M$. Noter que cette définition du Lagrangien correspond à un choix qui est fait (et qui peut sembler arbitraire) de dualiser les contraintes d'égalité (correspondant aux contraintes de marginales), mais pas les contraintes de positivité.

Le problème primal (voir definition 23.24, page 246) est le problème consistant à minimiser la fonctionnelle

$$F(\gamma) = \sup_{p,q} L(\gamma, p, q) = \left| \begin{array}{lll} \displaystyle\sum_{i,j} c_{ij}\gamma_{ij} & \text{si} & \gamma \in \Pi \\ +\infty & \text{sinon} & \end{array} \right.$$

Minimiser cette fonctionnelle revient bien à résoudre le problème de minimisation sous contrainte 15.2.

Le problème dual (voir toujours la définition 23.24, page 246) consiste à maximiser la fonctionnelle duale $G(p, q) = \inf_\gamma L(\gamma, p, q)$. Cette fonctionnelle s'exprime (on ordonne différemment les sommes dans l'expression de $L(\gamma, p, q)$) :

$$\begin{aligned} G(p,q) &= \inf_{\gamma \in V} \left(\sum_{i,j} (c_{ij} - p_i - q_j)\, \gamma_{ij} + \sum_{i=1}^{N} p_i\mu_i + \sum_{j=1}^{M} q_j\nu_j \right) \\ &= \sum_{i=1}^{N} p_i\mu_i + \sum_{j=1}^{M} q_j\nu_j + \inf_{\gamma \in V} \left(\sum_{i,j} (c_{ij} - p_i - q_j)\, \gamma_{ij} \right). \end{aligned}$$

Comme γ parcourt $V = \mathbb{R}_+^{NM}$, l'infimum ci-dessus vaut $-\infty$ à moins que l'on ait $p_i + p_j \leq c_{ij}$ pour tous i, j, et 0 dans ce dernier cas. On a donc

$$G(p, q) = \inf_{\gamma \in V} L(\gamma, p, q) = \left| \begin{array}{lll} \displaystyle\sum_{i=1}^{N} p_i\mu_i + \sum_{j=1}^{M} q_j\nu_j & \text{si} & p_i + q_j \leq c_{ij} \quad \forall\, i\,,\ j, \\ -\infty & \text{sinon}\ . & \end{array} \right.$$

On écrira $p \oplus q \leq c$ la contrainte d'inégalité sur les p_i et q_j. Le problème dual (il est immédiat que l'ensemble des p, q, vérifiant la contrainte est non vide) s'écrit donc

$$\sup_{p \oplus q \leq c} (p \cdot \mu + q \cdot \nu)\,.$$

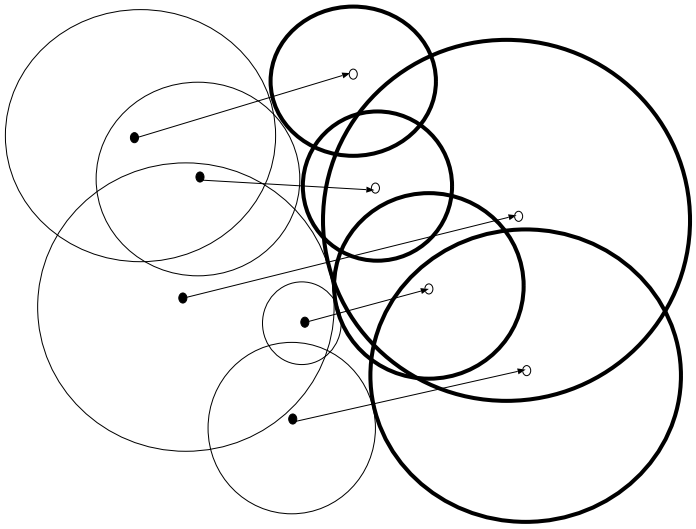


FIGURE 15.1 – Interprétation géométrique des potentiels de Kantorovich pour la distance 1.

Il s'agit de montrer que le Lagrangien défini ci-dessus admet un point selle ou, de façon équivalente (voir proposition 23.26, page 246), que le problème dual admet une solution, et que sa valeur maximale est la valeur minimale du problème initial. La remarque suivante permet de se ramener à la construction de vecteurs de multiplicateurs de Lagrange vérifiant une propriété très simple.

Remarque 15.7. *Soit γ un plan de transport entre μ et ν. Si (p,q) vérife $p \oplus q \leq c$, avec égalité sur le support de γ, i.e.*

$$\gamma_{ij} > 0 \Longrightarrow p_i + q_j = c_{ij},$$

alors (γ, p, q) est point-selle pour le Lagrangien L (défini par (15.2)). En effet, (p,q) vérifie alors la contrainte du problème dual, et on a

$$G(p,q) = \sum_i \mu_i p_i + \sum_j \mu_j q_j = \sum_{ij} \gamma_{ij}(p_i + q_j) = \sum_{ij} \gamma_{ij} c_{ij} = F(\gamma).$$

Comme on a $G(\tilde{p},\tilde{q}) \leq F(\tilde{\gamma})$, cela implique que (p,q) (resp. γ) est solution du problème dual (resp. primal).

Remarque 15.8. *Dans le cas où X et Y sont des collections d'un même nombre N de points de $\mathbb{R}^d$, et que $c_{ij} = |y_j - x_i|$, la remarque précédente peut s'interpréter géométriquement : pour trouver un minimiseur du coût, il suffit*[71] *de trouver $2N$ cercles (ou sphères pour $d \geq 3$) Σ_i^x et Σ_j^y centrés en les points x_i et y_j, respectivement, de telle sorte qu'il existe une bijection φ telle que Σ_i^x est tangent à $\Sigma_{\varphi(i)}^y$, et que les autres couples de cercles (Σ_i^x, Σ_j^y) ne se chevauchent pas strictement. Selon cette vision du problème dual, les p_i (resp. q_j) sont les rayons des cercles Σ_i^x (resp. Σ_j^y). La figure 15.1 donne un exemple d'une telle construction, pour $d = 2$ et $N = 5$.*

71. Il s'agit essentiellement d'une interprétation géométrique des potentiels de Kantorovich, il n'est pas clair que ce nouveau problème soit plus facile à résoudre que le problème de minimisation initial.

15.4 Existence d'une solution au problème dual

Bien qu'il soit d'usage, en programmation linéaire, de conserver la contrainte de positivité du γ sous forme "essentielle" (l'espace primal intègre cette contrainte, sans expression duale), la construction d'un nouveau Lagrangien qui dualise ces contraintes permet ici (dans le cas de la dimension infinie) de montrer rapidement l'existence d'un point-selle.

L'approche consiste simplement, comme dans la définition 23.27, page 246, à ajouter un terme du type $-\sum \mu_{ij}\gamma_{ij}$ au Lagrangien défini précédemment :

$$\tilde{L} \ : \ (\gamma, p, q, \mu) \in \mathbb{R}^{NM} \times \mathbb{R}^N \times \mathbb{R}^M \times \mathbb{R}_+^{NM} \longmapsto \tilde{L}(\gamma, p, q, \mu) = L(\gamma, p, q) - \sum \mu_{ij}\gamma_{ij}.$$

Proposition 15.9. *Le Lagrangien $L(\,\cdot\,,\cdot\,,\cdot\,)$ admet un point selle (γ, p, q) ou, de façon équivalente,*

$$G(p, q) = \max_{\tilde{p}\oplus\tilde{q}\leq c} G(\tilde{p}, \tilde{q}) = \min_{\tilde{\gamma}\in\Pi} F(\tilde{\gamma}) = F(\gamma).$$

D'après la proposition 23.22, page 244 (en notant que les contraintes d'égalité affines peuvent se traiter comme deux contraintes d'inégalité affines [72], pour lesquelles la question de qualification ne se pose pas comme le précise la définition 23.21), il existe p , q, et $\mu \geq 0$ tels que

$$c_{ij} - p_i - q_j - \mu_{ij} = 0,$$

avec $\mu_{ij} = 0$ dès que $\gamma_{ij} > 0$ (contrainte non activée). Le couple (p, q) vérifie donc la contrainte d'inégalité, avec égalité sur le support de γ, ce qui implique (voir remarque 15.7) que (γ, p, q) est point-selle du Lagrangien.

L'existence d'un point-selle peut aussi être obtenue, de façon plus laborieuse, à partir de la régularisée entropique du problème de minimisation (voir section 15.11, page 132).

15.5 Exemples d'applications

Sous sa forme la plus générale, le problème est entièrement déterminé par les mesures d'arrivée et de départ, et les coûts c_{ij}. Dans un grand nombre de situations, X et Y sont des ensembles de points de l'espace euclidien, et c_{ij} est une certaine mesure de la distance entre eux.

Ainsi, la version discrète du problème de Monge correspond à la donnée d'une mesure de départ μ supportée par N points (x_i), du plan, la mesure d'arrivée ν est supportée par M points (y_j), et les coût sont donnés par $c_{ij} = |y_j - x_i|$. Le problème envisagé par Monge concernait des déblais et des remblais, on peut étendre ce cadre est des lieux de production et de distribution : N boulangeries produisent des quantités de pain journalières $\mu_1, \ldots, \mu_N$, destinées à M dépôts de pains qui distribuent respectivement $\nu_1, \ldots, \nu_M$. Si l'on suppose que le coût de transport d'une quantité de pain peut être calculée en multipliant la quantité par un coût unitaire [73], et que ce coût unitaire est lui même proportionnel à la distance entre

72. On n'a bien sûr alors aucun information sur le signe du multiplicateur de Lagrange (ici p_i ou q_j), dont le signes final dépendra de laquelle des deux contraintes est réellement activée.

73. Cette hypothèse qui est assez discutable, et donc problématique puisque toute l'approche est basée sur cette hypothèse.

point de départ et point d'arrivé (on peut penser au coût de l'essence), minimiser le coût total correspond au problème considéré précédemment.

Une généralisation immédiate de ce problème consiste à considérer des coûts du type $c_{ij} = |y_j - x_i|^p$, le cas $p = 2$ jouant un rôle extrêmement important dans de multiples domaines. Une "application" dans le cas quadratique est la suivante : on considère deux systèmes de N points du plan, que l'on cherche à connecter deux à deux par des ressorts de longueur au repos nulle. Minimiser l'énergie élastique (quadratique en les positions) revient à choisir les couples que l'on va connecter.

Exercice 15.3. (Matching) Montrer que, dans le cas où X et Y sont des points d'un espace euclidien, et dans le cas quadratique $c_{ij} = |y_j - x_i|^2$, minimiser le coût global revient à maximiser la somme des $\gamma_{ij}\, x_i \cdot y_j$. Considérer la situation où X correspond à un ensemble d'*agents*, représenté par un vecteur de nombres réels (par exemple entre 0 et 1 pour fixer les idées) correspondant à l'intérêt que chacun porte aux caractéristiques d'un produit, l'ensemble Y (vecteurs de même type) représentant l'ensemble des produits offerts au "marché" X. Interpréter alors le problème de transport optimal de X vers Y au vu de la remarque précédente.

Interprétation des q_j comme prix. Dans un esprit proche de ce qui précéde, on considère un ensemble d'agents X, et l'on suppose que chaque agent est doté d'un capital μ_j. L'ensemble des biens [74] est noté Y, et la quantité de chaque bien (mesurée dans la même unité que les μ_j) vaut ν_j. On note u_{ij} l'*utilité* que représente le bien j pour l'agent i, de telle sorte que ηu_{ij} mesure en quelque sorte la satisfaction apportée à i s'il consacre une partie η de son capital à l'acquisition du bien j. Maximiser la satisfaction globale correspond à un problème de type Monge-Kantorovich discret

$$\max_{\gamma \in \Pi} \sum_{ij} \gamma_{ij} u_{ij}.$$

Ce contexte conduit à une interprétation limpide des potentiels de Kantorovich, ou multiplicateurs de Lagrange associés aux contraintes de marginale. On considère que le bien j a un prix q_j, et que les utilités sont exprimées dans une unité telle que $u_{ij} - q_j$ quantifie l'attrait effectif de j pour i (qui diminue bien sûr lorsque le prix augmente). On a alors une interprétation très claire de la remarque 15.7, qui dans le contexte présent exprime que le problème de maximisation est équivalent à la recherche d'un système de prix pour les différents biens, et d'un plan décrivant le comportement effectifs des agents, de façon à ce que chaque agent n'ait aucun intérêt à changer son choix. Supposons plus précisemment que l'on connaisse un plan de marché γ (qui encode l'ensemble des choix des agents) et un système de prix q tel que, pour tout i, pour tout j dans le support de γ (c'est à dire que i achète une quantité non nulle de j), on ait

$$u_{ij} - q_j = \max_k (u_{ik} - q_k),$$

ce qui signifie simplement que, le système de prix étant ce qu'il est, l'agent i perd tout intérêt pour les biens qui ne correspondent pas à son choix courant, il est *content*, ou tout du moins, en l'état actuel du reste de l'univers, il ne peut pas augmenter sa satisfaction en changeant ses choix. Si l'on pose $p_i = \max_k u_{ik} - q_k$, on dispose d'un plan de transport, et d'un couple (p, q)

74. Les biens sont considérés ici comme des quantités sécables, et pas comme des biens discrets tels que l'achat ou le non achat se représenterait de façon binaire.

qui vérifie $p \oplus q \geq u$ avec égalité sur le support de γ, on a donc une solution du problème. Les q_j, associés au contraintes sur les produits, s'interprètent donc comme des prix, et les p_i, de la forme $u_{ij} - q_j$, à une certaine forme de satisfaction effective des différents agents.

Exercice 15.4. a) Dans le cas du coût ℓ^1 (i.e. $c_{ij} = |y_j - x_i|$), donner des exemples de situations pour lesquels on n'a pas unicité du minimiseur.

b) Même question pour le coût quadratique $c_{ij} = |y_j - x_i|^2$.

15.6 Interpolation

On note $\mathcal{A}(\mathbb{R}^d)$, ou simplement $\mathcal{A}$, l'ensemble des mesures atomiques sur $\mathbb{R}^d$ à support fini, c'est à dire l'ensemble des μ de la forme

$$\mu = \sum_{i=1}^{N} \mu_i \delta_{x_i}\,,\ \mu_i > 0\,,\ \sum_{i=1}^{N} \mu_i = 1\,,\ \mu_i \geq 0.$$

Si l'on se donne deux mesures ρ_0 et ρ_1 de $\mathcal{A}$, l'existence d'un plan de transport optimal de ρ_0 vers ρ_1 permet de définir une notion d'interpolée entre ces deux mesures. Précisons qu'il existe une première manière canonique, eulérienne en quelque sorte, d'interpoler les deux mesures, en définissant simplement

$$\tilde{\rho}_t = (1-t)\rho_0 + t\rho_1.$$

Pour tout $t \in [0,1]$, $\tilde{\rho}_t$ est une mesure de probabilité, et la courbe $t \mapsto \rho_t$ relie les deux mesures dans un certain sens, ce qui assure à peu de frais la convexité de l'espace des mesures (de probabilité) atomiques. Le support de ρ_t est la réunion des deux supports, pour $t \in]0,1[$.

Si l'on considère maintenant 2 points x_0 et x_1 de $\mathbb{R}^d$, on peut construire, de façon tout aussi canonique, un segment reliant ces points par interpolation affine : $x_t = (1-t)x_0 + tx_1$. On peut définir pour les mesures une notion d'*interpolation par déplacement* plus respectueuse de ce second point de vue (Lagrangien en quelque sorte). Cette notion a été introduite par R. McCann [75] en 1997, et on parle parfois d'interpolation *au sens de McCann*.

Cette notion est particulièrement féconde dans un contexte où l'on a unicité d'un plan de transport optimal (dans un sens qui peut dépendre du contexte), mais elle est basée sur la possibilité d'associer à tout plan de tranport admissible une interpolée canonique. C'est ce choix que nous faisons de définir ci-dessous une notion, non pas d'interpolée entre deux mesures, mais d'interpolée associée à un plan de transport.

Definition 15.10. *Soient ρ_0 et ρ_1 deux mesures de $\mathcal{A}$, et $\gamma \in \Pi_{\rho_0,\rho_1}$ un plan de transport entre ρ_0 et ρ_1. On associe à γ l'interpolée par déplacement définie de la façon suivante :*

$$\rho_t^\gamma = \sum_{ij} \gamma_{ij} \delta_{(1-t)x_i + ty_j}.$$

On parle dans la litérature de l'interpolée entre deux mesures en privilégiant la construction associée au plan de transport optimal entre les deux mesures (lorsque celui-ci est unique).

75. Robert J. McCann, A Convexity Principle for Interacting Gases, Advances in Mathematics 128, 153 179 (1997),
`http://www.math.toronto.edu/mccann/papers/advances.pdf`

L'ensemble des mesures de probabilités atomiques sur $\mathbb{R}^d$ reste convexe pour cette nouvel acception de l'interpolation : pour tout plan de transport, la courbe $t \mapsto \rho_t^\gamma$ associée reste dans $\mathcal{A}$, on parlera de convexité par déplacement (*displacement convexity*).

Noter en revanche que, si l'on se restreint à l'ensemble $\mathcal{A}(K)$ des mesures supportées dans un compact K donné, on perd la convexité de $\mathcal{A}(K)$ dès que K n'est plus convexe.

Remarque 15.11. *Si Ψ est une fonction strictement convexe de $\mathbb{R}^d$ dans $\mathbb{R}$, et ρ_t la courbe d'interpolation associée à un transport γ entre deux mesures atomiques ρ_0 et ρ_1 distinctes, la fonction*

$$t \mapsto \langle \rho_t \, , \, \Psi \rangle = \int_{\mathbb{R}^d} \Psi(x) \, d\rho_t$$

est strictement convexe. Noter que la même fonction définie à partir de l'interpolée eulérienne $\tilde{\rho}_t$ est simplement l'interpolée affine entre les deux valeurs extrêmes, elle est donc convexe, mais aussi concave, quelles que soient les proprités de convexité de la fonction Ψ.

15.7 Métrique induite sur l'ensemble des mesures atomiques

On note comme précédemment $\mathcal{A} = \mathcal{A}(\mathbb{R}^d)$ l'ensemble des mesures de probabilités atomiques sur $\mathbb{R}^d$ à support fini, c'est à dire l'ensemble des μ de la forme

$$\mu = \sum_{i=1}^N \mu_i \delta_{x_i} \, , \ \mu_i > 0 \, , \ \sum_{i=1}^N \mu_i = 1.$$

L'entier N n'est pas fixé, mais on ne considère ici que des sommes finies. Pour $p \geq 1$ fixé, μ et ν dans A_d, on note

$$W_p(\mu, \nu) = \left(\inf_{\gamma \in \Pi(\mu,\nu)} \sum \gamma_{ij} \, |y_j - x_i|^p \right)^{1/p},$$

où l'infimum correspond au problème de MK discret 15.2, pour lequel l'existence d'un plan minimisant est établie dans 15.3. On se propose de montrer que W_p est une distance sur A_d.

Théorème 15.12. *La fonction $W_p(\,\cdot\,,\,\cdot\,)$ définie ci-dessus sur $\mathcal{A} \times \mathcal{A}$ est une distance.*

Démonstration. On a de façon évidente $W_p(\mu, \nu) = 0$ si et seulement si $\mu = \nu$, et la distance est symétrique par construction (le problème de recherche d'un plan de coût minimal est symétrique par rapport aux mesures). Pour l'inégalité triangulaire, on considère trois mesures μ^1, μ^2, et μ^3 de $\mathcal{A}$. On note γ^{12} et γ^{23} des plans qui réalisent la distance de 1 vers 2 et de 2 vers 3, respectivement. On note γ^{123} le "plan à trois" défini de la façon suivante [76]

$$\gamma^{123}_{i_1 i_2 i_3} = \frac{1}{\mu^2_{i_2}} \gamma^{12}_{i_1 i_2} \gamma^{23}_{i_2 i_3}.$$

On note γ^{13} le plan défini de façon naturelle par

$$\gamma^{13}_{i_1 i_3} = \sum_{i_2} \gamma^{123}_{i_1 i_2 i_3}.$$

76. On peut voir γ^{123} comme la loi d'une variable aléatoire sur $\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d$ dont les projections ont pour lois respectives μ^1, μ^2 et μ^3.

On a

$$W_p(\mu^1,\mu^3) \le \left(\sum_{i_1 i_3} \gamma^{13}_{i_1 i_3} \left|x^3_{i_3} - x^1_{i_1}\right|^p\right)^{1/p} = \left(\sum_{i_1 i_2 i_3} \gamma_{i_1 i_2 i_3} \left|x^3_{i_3} - x^1_{i_1}\right|^p\right)^{1/p}$$
$$\le \left(\sum_{i_1 i_2 i_3} \gamma^{123}_{i_1 i_2 i_3} \left|x^2_{i_2} - x^1_{i_1}\right|^p\right)^{1/p} + \left(\sum_{i_1 i_2 i_3} \gamma^{123}_{i_1 i_2 i_3} \left|x^3_{i_3} - x^2_{i_2}\right|^p\right)^{1/p}$$

d'après l'inégalité de Hölder, d'où finalement

$$W_p(\mu^1,\mu^3) \le \left(\sum_{i_1 i_2} \gamma^{12}_{i_1 i_2} \left|x^2_{i_2} - x^1_{i_1}\right|^p\right)^{1/p} + \left(\sum_{i_2 i_3} \gamma^{23}_{i_2 i_3} \left|x^3_{i_3} - x^2_{i_2}\right|^p\right)^{1/p} = W_p(\mu^2,\mu^3) + W_p(\mu^1,\mu^2),$$

ce qui termine la preuve. □

Exercice 15.5. Montrer que l'espace $\mathcal{A}$ défini ci-dessus n'est pas complet, même si l'on contraint les supports des mesures à demeurer dans un compact de $\mathbb{R}^d$. Identifier des sous-ensembles stricts de A_d qui sont complets pour la même métrique.

Exercice 15.6. On considère l'espace $\mathcal{A}^N$ des mesures atomiques de $\mathbb{R}^d$ à N points (non nécessairement distincts), avec équidistribution de masse sur les N points. Identifier l'espace métrique $\mathcal{A}^N$ muni de la distance précédemment définie.

15.8 Approche de Benamou-Brenier

Cette section présente les principes d'une formulation alternative du problème de Monge Kantorovich proposée par Benamou et Brenier à la fin du siècle dernier [77]. Cette approche s'est révélée extrêmement féconde sur le plan de la résolution numérique de tels problèmes, mais aussi sur le plan abstrait. Soient x_0 et x_1 deux points de $\mathbb{R}^d$. Pour toute vitesse $v(t)$ régulière donnée sur l'intervalle $[0,1]$ telle que la trajectoire associée x_t relie x_0 et x_1, la longueur ℓ de la courbe vérifie

$$|x_1 - x_0|^2 \le \ell^2 = \left(\int_0^1 |v(s)|\right)^2 ds \le \int_0^1 |v(s)|^2 \, ds.$$

Par ailleurs, si l'on prend la vitesse constante égale à $(x_1 - x_0)$, on a égalité entre les deux extrémités de la chaine précédente d'inégalités. On a donc

$$|x_1 - x_0|^2 = \min_{x_1 = x_0 + \int v} \int_0^1 |v(s)|^2 \, ds.$$

On peut généraliser cette approche à deux mesures atomiques supportées par des nuages de points (x_i) et (y_j), en considérant pour chaque couple (x_i, y_j) une vitesse v_{ij} sur $[0,1]$ susceptible de les relier. On notera W l'ensemble des vitesses admissibles correspondant à cette condition. Le problème de transport optimal avec coût quadratique s'écrit alors

$$\min_{v \in W, \gamma \in \Pi} \left(\sum_{ij} \int_0^1 \gamma_{ij} \, |v_{ij}(s)|^2 \, ds\right)$$

77. J.D. Benamou, Y. Brenier, A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem, Numerische Mathematik January 2000, Volume 84, Issue 3, pp 375-393, http://link.springer.com/article/10.1007/s002110050002

On peut écrire différemment ce problème en utilisant la notion de solution faible de l'équation de transport. On se ramène ainsi à la recherche d'un champ de vitesse v_t qui est ρ_t-mesurable pour tout $t \in [0,1]$, qui transporte ρ_0 vers ρ_1, i.e. (ρ_t, v_t) est solution faible sur $\mathbb{R}^d \times [0,1]$ de l'équation de transport

$$\partial_t \rho_t + \nabla \cdot (\rho_t v_t) = 0,$$

avec données initiales et finales ρ_0 et ρ_1, et qui minimise la quantité

$$\int_0^1 \int_{\mathbb{R}^d} |v_t|^2 \, d\rho_t.$$

Cette approche se généralise à des mesures quelconques sur $\mathbb{R}^d$.

15.9 Étude de W_1

Dans le cas $p = 1$, la distance peut s'exprimer de façon particulière, qui exprime un premier lien entre ce type de métrique et la convergence faible des mesures. On note comme précédemment $\mathcal{A}$ l'ensemble des mesures de probabilités atomiques sur $\mathbb{R}^d$ à support fini, c'est à dire l'ensemble des μ de la forme

$$\mu = \sum_{i=1}^N \mu_i \delta_{x_i} \,, \ \mu_i > 0 \,, \ \sum_{i=1}^N \mu_i = 1.$$

Proposition 15.13. *(Distance W_1 sur les mesures atomiques.)*
Pour toutes mesures μ et ν de $\mathcal{A}(\mathbb{R}^d)$ (mesures atomiques à support fini), on a

$$W_1(\mu,\nu) = \inf_{\gamma \in \Pi(\mu,\nu)} \sum_{ij} \gamma_{ij} \, |y_j - x_i| = \max_{\varphi \in \mathrm{Lip}_1} \left(\sum_i \mu_i \varphi(x_i) - \sum_j \nu_j \varphi(y_j) \right),$$

où Lip_1 *est l'ensemble des fonctions 1-Lipschitziennes.*

Démonstration. On note γ_{ij} un plan optimal entre μ et ν. On a, pour toute fonction 1-Lipschitzienne,

$$\sum_i \mu_i \varphi(x_i) - \sum_j \nu_j \varphi(y_j) = \sum_{i,j} \gamma_{ij} \left(\varphi(x_i) - \varphi(y_j) \right) \leq \sum_{i,j} \gamma_{ij} \, |y_j - x_i| \ = W_1(\mu,\nu). \tag{15.3}$$

Réciproquement, considérons une solution (p,q) du problème dual :

$$\sum_i p_i \mu_i + \sum_j q_j \nu_j = W_1(\mu,\nu) \text{ avec } \ p_i + q_j \leq c_{ij} \,, \ p_i + q_j = c_{ij} \text{ sur } \ \mathrm{supp}(\gamma).$$

On a, pour tout i, $p_i \leq c_{ij} - q_j$ pour tout j, avec égalité pour au moins un indice j, donc

$$p_i = \min_j (c_{ij} - q_j).$$

Considérons maintenant la fonction

$$\varphi \, : \, x \longmapsto \inf_j \left(|y_j - x| - q_j \right).$$

Cette fonction est 1-Lipschitzienne comme infimum de fonctions 1-Lipschitziennes[78]. Par ailleurs φ prend les valeurs du potentiel de Kantorovitch sur le support de μ :

$$\varphi(x_i) = \inf_j \left(|y_j - x_i| - q_j\right) = p_i.$$

Enfin, on a

$$\varphi(y_j) = \inf_k \left(|y_k - y_j| - q_j\right) \leq -q_j,$$

donc $-\varphi(y_j) \geq q_j$. Pour cette fonction φ particulière, on a donc

$$\sum_i \mu_i \varphi(x_i) - \sum_j \nu_j \varphi(y_j) \geq \sum_i \mu_i p_i + \sum_j \nu_j q_j = W_1(\mu, \nu).$$

On a donc, d'après (15.3),

$$\sup_{\varphi \in \mathrm{Lip}_1} \left(\sum_i \mu_i \varphi(x_i) - \sum_j \nu_j \varphi(y_j) \right) = \max_{\varphi \in \mathrm{Lip}_1} (\,\cdot\,) = W_1(\mu, \nu).$$

□

15.10 Complétion de l'espace de Wasserstein discret

On définit maintenant $\mathcal{A} = \mathcal{A}(K)$ comme l'ensemble des mesures de probabilités atomiques supportées dans un compact K de $\mathbb{R}^d$, c'est à dire l'ensemble des μ de la forme

$$\mu = \sum_{i=1}^N \mu_i \delta_{x_i}\,,\ \mu_i > 0\,,\ \sum_{i=1}^N \mu_i = 1\,,\ x_1, \ldots, x_N \in K,$$

avec toujours $N \in \mathbb{N}$ non fixé (il dépend de μ, et n'est pas borné). Pour $p \geq 1$ fixé, μ et ν dans $\mathcal{A}$, on note comme précédemment

$$W_p(\mu, \nu) = \left(\inf_{\gamma \in \Pi(\mu,\nu)} \sum \gamma_{ij} \, |y_j - x_i|^p \right)^{1/p}.$$

Proposition 15.14. *Le complété de $\mathcal{A}$ pour la distance W_p s'identifie à l'espace $\mathcal{P}(K)$ des mesures de probabilité sur K.*

Démonstration. Le complété abstrait de $\mathcal{A}$ est l'espace des suites de Cauchy pour W_p quotienté par la relation d'équivalence

$$(\mu^n) \sim (\nu^n) \Longleftrightarrow W_p(\mu^n, \nu^n) \longrightarrow 0.$$

78. On a $\varphi(x) = \inf_j \varphi_j(x)$. Pour tous x, y, on a $\varphi(x) = \varphi_j(x)$ pour un certain j, d'où

$$\varphi(y) = \inf_k \varphi_k(y) \leq \varphi_j(y) \leq \varphi_j(x) + |y - x| = \varphi(x) + |y - x|\,,$$

et ainsi $\varphi(y) - \varphi(x) \leq |y - x|$. On a de la même manière $\varphi(x) - \varphi(y) \leq |y - x|$.

De toute suite (μ^n) de Cauchy dans $\mathcal{A}$ (en fait, pour toute suite de $\mathcal{A}$), on peut extraire une sous-suite qui converge faiblement [79] dans $\mathcal{P}(K)$. Montrons que la limite ne dépend pas du représentant dans la classe d'équivalence. Soient μ^n et ν^n deux suites adjacentes ($\mu \sim \nu$), et φ une fonction Lipschtzienne sur K. On a (en notant γ^n un plan optimal de μ^n vers ν^n)

$$\langle \nu^n - \mu^n \, , \varphi \rangle = \sum_j \nu_j^n \varphi(y_j^n) - \sum_i \mu_i^n \varphi(x_i^n) = \sum_j \sum_i \gamma_{ij}^n \left(\varphi(y_j^n) - \varphi(x_i^n) \right)$$

$$\leq L \sum_j \sum_i \gamma_{ij}^n \left| \; y_j^n - x_i^n \right| \leq L \left(\sum_j \sum_i \gamma_{ij}^n \left| \; y_j^n - x_i^n \right|^p \right)^{1/p}$$

$$= W_p \left(\mu^n, \nu^n \right) \longrightarrow 0 \text{ quand } \; n \to +\infty. \tag{15.4}$$

On a bien sûr la même inégalité pour $\langle \nu^m - \mu^n \, , \varphi \rangle$, d'où la convergence de $\langle \nu^n - \mu^n \, , \varphi \rangle$ vers 0. Par densité des fonctions Lipschitziennes dans les fonctions continues (K est compact), les mesures limites sont donc les mêmes.

Montrons que toute mesure de probabilité $\mu \in \mathcal{P}(K)$ peut être approchée faiblement par une telle suite. On suppose dans un premier temps que K est un (hyper-)cube. Pour $n \in \mathbb{N}$, on décompose K de façon régulière en n^d petits cubes (C_i^n), de centres x_i^n. On associe à μ une mesure atomique portée par les x_i^n, en prenant pour masse μ_i^n la μ-mesure de C_i^n (si μ charge les faces entre les cubes, on choisit arbitrairement d'associer la masse d'une face à l'une des cellules adjacentes). Par construction, le p-coût entre μ^n et μ^m (avec $n \leq m$) est de l'ordre de $1/n^p$: la suite est donc bien de Cauchy. Si K n'est pas un cube, on suit le même procédé avec un cube contenant K, en projetant sur K les centres des cellules qui seraient à l'extérieur.

□

Remarque 15.15. *Toute mesure μ de $\mathcal{P}(K)$ est ainsi limite (pour W_p) d'une suite (μ^k) d'éléments de $\mathcal{A}(K)$. En appliquant la chaîne d'inégalités* (15.4) *à μ^k et μ^ℓ, et en faisant tendre ℓ vers l'infini, on montre par ailleurs, en suivant un raisonnement analogue à ce qui précède, que*

$$\langle \varphi \, , \mu - \mu_k \rangle \longrightarrow 0$$

pour toute fonction φ continue sur K.

Proposition 15.16. *La métrique W_p induite sur $\mathcal{P}(K)$ par la complétion décrite précédemment métrise la topologie de la convergence faible sur $\mathcal{P}(K)$, i.e.*

$$\mu_n \rightharpoonup \mu \Longleftrightarrow W_p(\mu_n, \mu) \longrightarrow 0.$$

Démonstration. On montre dans un premier temps que l'équivalence est vérifiée pour $p = 1$. On considère une suite $\mu_n \in \mathcal{P}(K)$ qui converge vers μ pour W_1. On approche les μ_n et μ par des suites (μ_n^k) et μ^k de $\mathcal{A}$. Pour toute fonction 1-Lipschitzienne φ, on a, pour tout n,

$$\langle \mu_n - \mu \, , \varphi \rangle = \lim_k \left\langle \mu_n^k - \mu^k \, , \varphi \right\rangle \leq W_1(\mu_n^k, \mu^k)$$

79. Comme K est compact, il n'y a pas lieu de distinguer ici la convergence étroite (contre les fonctions continues bornées), la convergence vague (contre les fonctions continues à support compact), ou convergence faible (contre l'adhérence de ces dernières pour la norme uniforme).

d'après la remarque 15.15 et la proposition 15.13. Par convergence de $W_1(\mu_n^k, \mu^k)$ vers $W_1(\mu_n, \mu)$, on a donc

$$\langle \mu_n - \mu \, , \, \varphi \rangle \leq W_1(\mu_n, \mu).$$

On a la même inégalité en prenant $-\varphi$, donc

$$|\langle \mu_n - \mu \, , \, \varphi \rangle| \ \leq W_1(\mu_n, \mu).$$

On a donc convergence vers 0 de $\langle \mu_n - \mu \, , \, \varphi \rangle$, pour toute fonction φ 1-Lipschtizienne, donc pour toute fonction Lipschitienne par linéarité, donc pour toute fonction continue par densité des fonctions Lipchitziennes dans les fonctions continues sur le compact K, d'où la convergence faible de μ_n vers μ.

Réciproquement, on considère une suite (μ_n) qui converge faiblement vers μ. On a

$$W_1(\mu_n, \mu) = \lim_k W_1(\mu_n^k, \mu^k)$$

On fixe n. Pour tout k, la distance $W_1(\mu_n^k, \mu^k)$ est réalisée faiblement sous la forme $\left\langle \mu_n^k - \mu^k \, , \, \varphi^k \right\rangle$, pour une fonction φ^k qui est 1-Lipschitzienne. Quitte à supposer que toutes ces fonctions valent 0 en un point fixé de K (on peut leur rajouter une constante arbitraire du fait que μ_n^k et μ^k ont même masse), on en extrait une sous-suite qui converge uniformément vers une fonction φ continue (théorème d'Arzelà-Ascoli). On a donc (pour la suite extraite, pour laquelle on conserve l'indice k par commodité d'écriture)

$$\lim_k W_1(\mu_n^k, \mu^k) = \lim_k \left\langle \mu_n^k - \mu^k \, , \, \varphi^k \right\rangle = \lim_k \left(\left\langle \mu_n^k - \mu^k \, , \, \varphi \right\rangle + \left\langle \mu_n^k - \mu^k \, , \, (\varphi^k - \varphi \right\rangle \right)$$

$$= \langle \mu_n - \mu \, , \, \varphi \rangle,$$

qui tend vers 0 quand n tend vers $+\infty$.

On en déduit la propriété pour $p > 1$ en notant que, pour toute mesure atomique (γ ci-dessous désigne le plan optimal pour le p-coût)

$$W_p(\mu, \nu)^p = \sum \gamma_{ij} \, |y_j - x_i|^p \geq \left(\sum \gamma_{ij} \, |y_j - x_i| \right)^p \geq W_1(\mu, \nu)^p.$$

Par ailleurs, pour tout $p \geq 1$, on a sur le borné K une inégalité $|y - x|^p \leq C \, |y - x|$ uniforme en $(x, y) \in K \times K$. On a donc (γ désigne maintenant le plan optimal pour le 1-coût)

$$W_p(\mu, \nu)^p \leq \sum \gamma_{ij} \, |y_j - x_i|^p \leq C \sum \gamma_{ij} \, |y_j - x_i| = C \, W_1(\mu, \nu).$$

On a donc finalement, pour toute mesure de probabilité atomique, et donc pour toute mesure de $\mathcal{P}(K)$ (les suites de Cauchy sont les mêmes dans W_p et W_1 du fait même des inégalités démontrées dans le cas atomique),

$$W_1(\mu, \nu) \leq W_p(\mu, \nu) \leq C^{1/p} \, W_1(\mu, \nu)^{1/p}.$$

□

Exercice 15.7. Décrire, dans $\mathcal{A}(K)$, le cercle dont le centre est un Dirac centré à l'origine, et de rayon 1. On considérera que K est une boule fermée de $\mathbb{R}^d$ centrée en l'origine.

15.11 Régularisation entropique

On propose ici une démonstration alternative de l'existence d'un point-selle, plus laborieuse, mais qui permet d'étudier une méthode effectivement utilisée en pratique. Cette méthode est basée sur la *régularisée entropique* de la fonctionnelle $C(\gamma)$, définie par

$$\gamma \in \mathbb{R}_+^{NM} \longmapsto C_\varepsilon(\gamma) = \sum_{i,j} c_{ij}\gamma_{ij} + \varepsilon \sum_{i,j} \gamma_{ij} \log \gamma_{ij} = C(\gamma) + \varepsilon S(\gamma), \tag{15.5}$$

où S est l'entropie de la probabilité γ sur $\mathbb{R}^N \times \mathbb{R}^M$ (voir définition 10.1, page 101).

Lemme 15.17. *On suppose que μ et ν chargent tous les points de X et Y, respectivement. La fonctionnelle C_ε définie par* (15.5) *admet un minimiseur γ^ε unique sur Π (défini par* (15.1)*), avec $\gamma_{ij}^\varepsilon > 0$ pour tous i, j).*

Démonstration. La fonction C_ε est continue sur le compact Π, elle admet un minimiseur γ^ε, qui est unique par convexité de Π et stricte convexité de C_ε.

Montrons que ce minimiseur a pour support $X \times Y$, c'est à dire que tous les γ_{ij} sont strictement positifs. Cette propriété vient du fait que la fonction choisie, $x \log x$, a une dérivée qui vaut $-\infty$ en 0, de telle sorte qu'il est très défavorable, en termes de minimisation, de s'approcher de cette limite. Pour utiliser ce fait et montrer qu'un tel point ne peut pas être minimiseur, il faut simplement vérifier que l'on peut faire de petites variations admissibles [80].

Supposons par exemple que γ_{11} soit nul. Comme $\mu_1 > 0$, il existe un j tel que $\gamma_{1j} > 0$, et de la même manière un i tel que $\gamma_{i1} > 0$. On perturbe alors γ de la façon suivante : on rajoute ε à γ_{11}, on enlève ε à γ_{i1}, on enlève ε à $\gamma_{1j} > 0$, et pour compenser le gain de i et la perte de j, on rajoute ε à γ_{ij}. Pour ε suffisamment petit ($< \min(\gamma_{i1}, \gamma_{1j})$), cette perturbation est admissible. Elle affecte linéairement la partie linéaire de la fonctionnelle, et linéairement au premier ordre les termes d'entropies sur les liens $1 \to j$ et $i \to 1$. Pour le terme d'entropie correspondant à $1 \to 1$, on a une variation négative qui domine les variations linéaires au voisinage de 0, du fait que la dérivée en 0 de $x \log x$ est $-\infty$. Si γ_{ij} était initialement non nul, la variation correspondante est linéaire, s'il était nul, on renforce la variation négative surlinéaire. □

Lemme 15.18. *Le Lagrangien associé au problème de minimisation régularisé :*

$$L_\varepsilon \ : \ (\gamma, p, q) \in V \times \Lambda \longmapsto \sum_{i,j} c_{ij}\gamma_{ij} + \varepsilon S(\gamma) + \sum_{i=1}^{N} p_i \left(\mu_i - \sum_j \gamma_{ij} \right) + \sum_{j=1}^{M} q_j \left(\nu_j - \sum_i \gamma_{ij} \right),$$

admet un point-selle $(\gamma^\varepsilon, p^\varepsilon, q^\varepsilon)$, où γ^ε est le minimiseur du lemme 15.17.

80. Cela pourrait ne pas être le cas comme l'illustre l'exemple suivant. Un problème classique consiste à minimiser l'entropie de la densité d'une loi de probabilité en imposant son espérance. Si l'espérance est prise égale à la valeur maximale que peut prendre la variable aléatoire, la densité va nécessairement charger cette valeur uniquement, et pourra donc prendre la valeur 0 sur les autres valeurs possibles.

Démonstration. La fonctionnelle C_ε réalise son minimum sur l'ouvert $]0,+\infty[^{NM}$, sous les contraintes de marginales, en γ^ε. Comme les contraintes sont affines on a, d'après la proposition 23.5, page 238, existence de multiplicateurs de Lagrange $(p^\varepsilon, q^\varepsilon) \in \mathbb{R}^N \times \mathbb{R}^M$ tels que

$$c_{ij} + \varepsilon(1 + \log \gamma_{ij}^\varepsilon) - p_i^\varepsilon - q_j^\varepsilon = 0. \tag{15.6}$$

On applique alors le corollaire 23.30 du théorème 23.29, page 248, qui assure que $(\gamma^\varepsilon, p^\varepsilon, q^\varepsilon)$ est point-selle du Lagrangien L_ε. □

Lemme 15.19. *Le problème dual associé au Lagrangien L_ε admet un maximum unique $(p^\varepsilon, q^\varepsilon)$ tel que la moyenne de p^ε est nulle.*

Démonstration. La fonctionnelle duale est définie par

$$G_\varepsilon(p,q) = \sum_{i=1}^N p_i\mu_i + \sum_{j=1}^M q_j\nu_j + \inf_{\gamma\in V}\left(\sum_{i,j}\left(c_{ij} - p_i - q_j + \varepsilon\log\gamma_{ij}\right)\gamma_{ij}\right). \tag{15.7}$$

La fonctionnelle de γ ci-dessus est strictement convexe, et admet un minimiseur caractérisé par

$$\gamma_{ij} = e^{-1}e^{-\frac{c_{ij}-p_i-q_j}{\varepsilon}},$$

ce qui donne

$$G_\varepsilon(p,q) = \sum_{i=1}^N p_i\mu_i + \sum_{j=1}^M q_j\nu_j - \varepsilon e^{-1}\sum_{i,j} e^{-\frac{c_{ij}-p_i-q_j}{\varepsilon}}. \tag{15.8}$$

Montrons que la matrice Hessienne de G_ε est semi-définie négative, et de noyau la droite engendrée par $(1,-1) \in \mathbb{R}^N \times \mathbb{R}^M$ (ajouter un élément de cette droite à (p,q) revient à ajouter une constante aux éléments de p, et enlever cette même constante aux éléments de q). On considère pour cela la matrice Hessienne de $(p,q) \longmapsto \sum e^{p_i+q_j}$ (on prend momentanément $\varepsilon = 1$ pour alléger l'écriture). Cette matrice H peut se décrire par blocs : 2 blocs diagonaux du type

$$D_p = \operatorname{diag}\left(e^{p_i}\sum_j e^{q_j}\right)_i,\ D_q = \operatorname{diag}\left(e^{q_j}\sum_i e^{p_i}\right)_j,$$

et un bloc extra-diagonal supérieur $B = (e^{p_i+q_j})_{ij}$ (le bloc inférieur est ${}^tB$). On a

$$(\overline{p},\overline{q})\cdot H\begin{pmatrix}\overline{p}\\ \overline{q}\end{pmatrix} = \sum_i e^{p_i}\overline{p}_i^2\sum_j e^{q_j} + \sum_j e^{q_j}\overline{q}_j^2\sum_i e^{p_i} + 2\sum_{ij}\overline{p}_i\overline{q}_j e^{p_i+q_j}.$$

On a $2\overline{p}_i\overline{q}_j \geq -\overline{p}_i^2 - \overline{q}_j^2$, avec inégalité stricte dès que $\overline{q}_j \neq -\overline{p}_i$. Si l'on prend $(\overline{p},\overline{q})$ non nul dans l'orthogonal de $(1,-1)$, on aura nécessairement $\overline{q}_j \neq -\overline{p}_i$ pour au moins l'un des couples (i,j), d'où

$$(\overline{p},\overline{q})\cdot H\begin{pmatrix}\overline{p}\\ \overline{q}\end{pmatrix} > 0.$$

La Hessienne de G_ε (qui est essentiellement l'opposé de la matrice H) est donc définie négative, G_ε admet donc un maximiseur unique dans l'orthogonal du noyau. Elle admet par suite un maximiseur unique tel que la moyenne des p_i est nullle, c'est ce minimiseur particulier que nous noterons $(p^\varepsilon, q^\varepsilon)$ dans la suite. □

Lemme 15.20. *La suite des $(p^\varepsilon, q^\varepsilon)$ construite ci-dessus est bornée.*

Démonstration. On note δ_{ij} le vecteur de $\mathbb{R}^N \times \mathbb{R}^M$ dont tous les éléments sont nuls, sauf le i-ème sur $\mathbb{R}^N$, et le j-ième sur $\mathbb{R}^M$, et C le cône convexe engendré par les δ_{ij} :

$$C = \left\{ \sum \gamma_{ij} \delta_{ij} \ , \ \gamma_{ij} \geq 0 \right\}.$$

On a $(\mu, \nu) \in C$. Plus précisément, (μ, ν) peut s'écrire comme une combinaison des δ_{ij} dont tous les coefficients sont strictement positifs (prendre par exemple pour γ_{ij} le transport qui distribue chaque masse μ_i selon la loi ν).

D'autre part, d'après (15.6), il existe une constante C telle que $p^\varepsilon \oplus q^\varepsilon \leq C$.

Enfin, comme $(p^\varepsilon, q^\varepsilon)$ maximise la fonctionnelle duale G_ε définie par (15.8), on a (on écrit simplement $G_\varepsilon(p^\varepsilon, q^\varepsilon) \geq G_\varepsilon(0,0)$) :

$$(p^\varepsilon, q^\varepsilon) \cdot (\mu, \nu) \geq (p^\varepsilon, q^\varepsilon) \cdot (\mu, \nu) - \varepsilon e^{-1} \sum_{i,j} e^{-\frac{c_{ij} - p_i - q_j}{\varepsilon}} \geq -\varepsilon e^{-1} \sum_{i,j} e^{-\frac{c_{ij}}{\varepsilon}} \geq \beta,$$

uniformément en ε (on peut supposer les c_{ij} positifs car le problème ne minimisation ne change pas si l'on rajoute une même constante à tous les c_{ij}).

Supposons maintenant que $(p^\varepsilon, q^\varepsilon)$ ne soit pas bornée, on peut extraire une sous-suite telle que la suite normalisée $(p^\varepsilon, q^\varepsilon)/\left|(p^\varepsilon, q^\varepsilon)\right|$ converge vers un (p, q) de norme 1, avec la moyenne des p_i égale à 0. Comme $p^\varepsilon \oplus q^\varepsilon \leq c$, on a à la limite $(p, q) \cdot \delta_{ij} \leq 0$ pour tous i, j, donc (p, q) est dans C°, cône polaire de C. On a aussi d'après ce qui précède $(p, q) \cdot (\mu, \nu) \geq 0$. Comme (μ, ν) est dans C, on a nécessairement $(p, q) \cdot (\mu, \nu) = 0$. Mais (voir début de la preuve), (μ, ν) s'écrit comme une combinaison de δ_{ij} à coefficients > 0, on a donc

$$0 = (p, q) \cdot (\mu, \nu) = \sum_{ij} \gamma_{ij} \delta_{ij} \cdot (p, q) = \sum_{ij} \gamma_{ij} (p_i + q_j).$$

Comme (p, q) est dans le polaire de C, il s'agit d'une somme de termes négatifs, qui sont donc tous nuls. Comme les γ_{ij} sont tous non nuls, on a finalement $p_i + q_j = 0$ quels que soient i et j. Les p_i sont donc tous identiques, donc (comme leur somme est nulle) tous nuls, de même pour les q_j, ce qui est absurde puisque (p, q) est de norme 1. □

Proposition 15.21. *Le minimiseur γ^ε construit au lemme 15.17 converge (à sous-suite extraite près) vers un minimiseur γ^0 de $C(\,\cdot\,)$, et toute valeur d'adhérence de la suite est minimiseur. Les multiplicateurs de Lagrange $(p^\varepsilon, q^\varepsilon)$ convergent eux mêmes (à sous-suite extraite près) vers un couple (p^0, q^0), et (γ^0, p^0, q^0) est point-selle du Lagrangien L.*

Démonstration. La suite (γ^ε), est bornée, on peut donc en extraire une sous-suite qui converge dans le fermé Π vers γ^0 , et l'on a

$$C(\gamma^\varepsilon) + \varepsilon S(\gamma^\varepsilon) \leq C(\gamma) + \varepsilon S(\gamma) \quad \forall \gamma \in \Pi,$$

d'où, par passage à la limite, $C(\gamma^0) \leq C(\gamma)$ pour tout $\gamma \in \Pi$. De plus, $(p^\varepsilon, q^\varepsilon)$ étant borné, on a convergence à sous-suite extraite près vers $(p^\varepsilon, q^\varepsilon)$. En passant à la limite dans (15.6), on obtient $p^0 \oplus q^0 \leq c$, avec

$$\gamma^0_{ij} > 0 \Longrightarrow p_i + q_j = \gamma_{ij},$$

d'où la conclusion (voir remarque 15.7). □

Remarque 15.22. *Si, faisant fi des bons usages, on fait tendre ε vers $+\infty$, on a convergence vers le minimiseur de l'entropie sous les contraintes de marginale, le coût n'intervient plus. Le minimiseur s'écrit*

$$\gamma_{ij} = Ce^{p_i+q_j} = Ce^{p_i}e^{q_j},$$

où C est une constante de normalisation (γ est une loi de probabilité sur $X \times Y$) . Du fait de l'écriture tensorielle ci-dessus, on peut voir γ comme une loi sur $X \times Y$ pour un couple de variables aléatoires indépendantes*.*

Remarque 15.23. *Noter que notion d'entropie permet de retrouver une certaine forme d'unicité dans le cas d'un problème de départ qui admet des solutions multiples : on peut choisir de privilégier parmi toutes les solutions celle qui minimise l'entropie, dont on peut montrer que c'est la limite des solutions aux problèmes régularisés quand ε tend vers* 0 *(voir proposition ci-dessous). Noter aussi que cette manière de sélectionner une solution n'est pas forcément légitime dans certains contextes. Lorsque les cardinaux sont les mêmes, et les mesures uniformes, on peut s'intéresser au contraire aux solutions du type bijection, qui sont celles qui maximisent au contraire l'entropie mathématique (i.e. qui minimisent l'entropie physique).*

Proposition 15.24. *On se donne deux mesures (μ_i), et (ν_j), une collection de coûts (c_{ij}), on note γ une solution du problème de MK discret 15.2, i.e. γ minimise*

$$C(\gamma) = \sum_{ij} \gamma_{ij} c_{ij},$$

sur $\Pi_{\mu,\nu}$ (défini par (15.1)*), et γ^ε le minimiseur du problème régularisé (voir lemme 15.17), qui minimise*

$$C_\varepsilon(\gamma) = \sum_{ij} \gamma_{ij} c_{ij} + \varepsilon \sum_{ij} \gamma_{ij} \log \gamma_{ij},$$

sur $\Pi_{\mu,\nu}$. Alors γ^ε converge vers $\overline{\gamma}$, plan qui minimise l'entropie parmi tous les minimiseurs admissibles de $C(\,\cdot\,)$.

Démonstration. On note C_{opt} la valeur du minimum de C sur Π. On ne change rien à un problème de minimisation en multipliant la fonctionnelle par une constante > 0 quelconque, et en rajoutant une constante arbitraire. On peut donc définir γ^ε comme le minimiseur sur Π d'une nouvelle fonctionnelle (on garde la notation C_ε par commodité)

$$C_\varepsilon(\gamma) = \frac{1}{\varepsilon}\left(C(\gamma) - C_{opt}\right) + S(\gamma)$$

L'ensemble admissible Π étant compact, on peut extraire de (γ_ε) une sous-suite qui converge vers un élément γ^0 de Π. Du fait que $C(\gamma^\varepsilon) \geq C_{opt}$, que γ^ε minimise C^ε, on a la chaîne d'inégalité suivante

$$S(\gamma^\varepsilon) \leq C_\varepsilon(\gamma^\varepsilon) \leq C_\varepsilon(\overline{\gamma}) = S(\overline{\gamma}),$$

où $\overline{\gamma}$ est le minimiseur de l'entropie parmi les minimiseurs du coût, qui est bien unique par stricte convexité de l'entropie sur l'ensemble convexe des minimiseurs du coût. On a donc à la limite $S(\gamma^0) \leq S(\overline{\gamma})$. Par ailleurs, d'après l'inégalité $C_\varepsilon(\gamma^\varepsilon) \leq S(\overline{\gamma})$ ci-dessus, la quantité

$$\frac{1}{\varepsilon}\left(C(\gamma) - C_{opt}\right) + S(\gamma)$$

est bornée, avec $S(\gamma)$ minoré, et $C(\gamma^\varepsilon) - C_{opt} \geq 0$. On a donc

$$C(\gamma^\varepsilon) \longrightarrow C_{opt},$$

d'où $C(\gamma^0) = C_{opt}$. Le plan limite γ^0 est donc minimiseur du coût, et il minimise l'entropie parmi ses confrères, γ^0 est donc bien le minimiseur de l'entropie parmi les minimiseurs du coût. On en conclut la convergence de toute la suite γ^ε vers $\overline{\gamma}$. □

15.12 Calcul effectif par Régularisation entropique

On considère deux mesures μ et ν supportées par des ensembles X et Y finis, de cardinaux respectifs N et M. Pour une matrice de coûts $c = c_{ij}$ donnée, on cherche à approcher une solution du problème 15.2, qui consiste à minimiser le coût

$$C(\gamma) = \sum_{i,j} c_{ij}\gamma_{ij},$$

sur l'ensemble Π des plans de transport admissibles (voir equation (15.1)), i.e. dont les marginales sont μ et ν.

Une méthode consiste à chercher un minimiseur pour la régularisée entropique de C, définie par

$$C_\varepsilon(\gamma) = \sum_{i,j} c_{ij}\gamma_{ij} + \varepsilon \sum_{i,j} \gamma_{ij}\log\gamma_{ij} = C(\gamma) + \varepsilon S(\gamma).$$

On a

$$\gamma_{ij}c_{ij} = -\varepsilon\gamma_{ij}e^{-c_{ij}/\varepsilon},$$

de telle sorte que

$$C_\varepsilon(\gamma) = \varepsilon \sum_{i,j} \gamma_{ij}\log\left(\frac{\gamma_{ij}}{\eta_{ij}}\right), \quad \text{avec } \eta_{ij} = e^{-c_{ij}/\varepsilon}.$$

Le coût régularisé est donc (au facteur ε près) l'entropie relative de γ (vu comme une loi de probabilité sur $X \times Y$) vis-à-vis de la loi [81] η. Cette entropie relative est aussi appelée *divergence de Kullback-Leibler*, et notée en conséquence $\mathrm{KL}(\gamma|\eta)$. Les conditions d'optimalité s'écrivent

$$1 + \log(\gamma_{ij}/\eta_{ij}) + p_i + q_j = 0.$$

Un plan γ est optimal si et seulement si (la condition est suffisante d'après le théorème 23.29, page 248) il peut se mettre sous la forme

$$\gamma_{ij} = a_i b_j \eta_{ij}\,,\ a_i > 0\,,\ b_j > 0, \tag{15.9}$$

tout en vérifiant bien sûr les conditions de marginales :

$$a_i \sum_j b_j\eta_{ij} = \mu_i\,,\ b_j \sum_i a_i\eta_{ij} = \nu_j. \tag{15.10}$$

L'approche itérative proposée ci-dessous s'appuie sur le caractère explicite de la minimisation de l'entropie relative lorsque l'on ne considère que l'une des deux contraintes (marginale

81. La densité η n'est pas nécéssairement de masse 1, mais la renormaliser conduit à rajouter une constante à C_ε, ce qui ne change pas le problème de recherche d'un minimiseur.

sur X ou sur Y). Considérons un plan $\bar{\gamma}$, et le problème consistant à minimiser l'entropie relative de γ relativement à $\bar{\gamma}$, sous la contrainte de marginale sur X :

$$\inf_{\gamma\in\Pi_\mu}\left(\sum\gamma_{ij}\log\left(\frac{\gamma_{ij}}{\bar{\gamma}_{ij}}\right)\right)\ ,\ \Pi_\mu=\left\{\ \gamma\in\mathbb{R}_+^{NM}\ ,\ \sum_j\gamma_{ij}=\mu_i\quad\forall i\right\}\ .$$

Du fait de la présence du log, les contraintes $\gamma_{ij}\geq 0$ ne sont pas activées (voir démonstration du lemme 15.17), et l'on a des multiplicateurs de Lagrange p_1, ..., p_N, tels que

$$\gamma_{ij}=\bar{\gamma}_{ij}e^{-p_i}\quad\forall i\ ,\ j.$$

On en déduit à l'aide des contraintes l'expression explicite

$$\gamma_{ij}=\bar{\gamma}_{ij}\frac{\mu_i}{\sum_j\bar{\gamma}_{ij}}.$$

Le problème de minimisation d'une fonctionnelle du même type avec contrainte de marginale sur Y peut évidemment se traiter de la même manière.

Algorithme 15.25. *On construit de façon itérative $\gamma^0=\eta$, $\gamma^{1/2}$, γ^1, ..., γ^k, $\gamma^{k+1/2}$, γ^{k+1}, ...de la façon suivante :*

$$\gamma_{ij}^{k+1/2}=\gamma_{ij}^k\frac{\mu_i}{\sum_j\gamma_{ij}^k}\quad\left(\gamma^{k+1/2}=\arg\min_{\Pi_\mu}KL(\gamma|\gamma^k)\right)$$

$$\gamma_{ij}^{k+1}=\gamma_{ij}^{k+1/2}\frac{\nu_j}{\sum_i\gamma_{ij}^{k+1/2}}\quad\left(\gamma^{k+1}=\arg\min_{\Pi_\nu}KL(\gamma|\gamma^{k+1/2})\right).$$

On peut voir cet algorithme de "projections"[82] alternées comme un algorithme de point fixe sur le problème en a_i, b_j donné par les équations (15.9)-(15.10). En effet, si l'on prend pour a^0 et b^0 des vecteurs qui ne contiennent que des 1, et qu'on pose

$$\gamma_{ij}^0=a_i^0b_j^0\eta_{ij}\ ,\ \gamma_{ij}^k=a_i^kb_j^k\eta_{ij}$$

une étape de l'algorithme précédent peut s'écrire

$$\gamma_{ij}^{k+1/2}=\gamma_{ij}^k\frac{\mu_i}{\sum_j\gamma_{ij}^k}=a_i^kb_j^k\eta_{ij}\frac{\mu_i}{\sum_j a_i^kb_j^k\eta_{ij}}=b_j^k\underbrace{\left(\frac{\mu_i}{\sum_j b_j^k\eta_{ij}}\right)}_{a_i^{k+1}}\eta_{ij},$$

$$\gamma_{ij}^{k+1}=\gamma_{ij}^{k+1/2}\frac{\nu_j}{\sum_i\gamma_{ij}^{k+1/2}}=a_i^{k+1}b_j^k\eta_{ij}\frac{\nu_j}{\sum_i a_i^{k+1}b_j^k\eta_{ij}}=a_i^{k+1}\underbrace{\left(\frac{\nu_j}{\sum_j a_i^{k+1}\eta_{ij}}\right)}_{b_j^{k+1}}\eta_{ij}.$$

L'algorithme se ramène finalement au calcul des a^1, b^1, ..., a^k, b^k, ..., selon la procédure

$$a_i^{k+1}=\frac{\mu_i}{\sum_j b_j^k\eta_{ij}}\ ,\ b_j^{k+1}=\frac{\nu_j}{\sum_i a_i^{k+1}\eta_{ij}}.$$

82. Il ne s'agit pas à strictement parler de projection, car la divergence de Kulback-Leibler n'est pas une distance.

Remarquons en premier lieu que, si l'algorithme en (a^k, b^k) converge vers (a, b), alors le plan limite $\gamma_{ij} = a_i b_j \eta_{ij}$ vérifie (15.9)-(15.10), c'est donc le minimiseur recherché.

Convergence de l'algorithme [83].

Implémentation effective en Python de l'approche par régularisation entropique.

Il est naturel de stocker la collection des coûts sous la forme d'une matrice (format `c = np.zeros((N,N))`). On peut calculer le plan initial η en écrivant simplement `eta = np.exp(-cc/eps)`.

15.13 Calcul effectif par l'algorithme des enchères

On considère ici deux ensembles X et Y de même cardinal N, et l'on s'intéresse au problème de *maximisation* de $\sum u_{i\varphi(i)}$. La quantité u_{ij} désigne ici l'*utilité* d'un agent i (acheteur potentiel) pour le produit j. On cherche ainsi à maximiser la satisfaction globale de la population X en trouvant une stratégie d'affectation adaptée à la distribution des utilités.

Remarquons en premier lieu que si l'on trouve une bijection $\varphi \in S_N$ et un système de prix (q_j) tels que

$$u_{i\varphi(i)} - q_{\varphi(i)} = \max_j \left(u_{ij} - q_j\right), \tag{15.11}$$

on a, en notant $p_i = u_{i\varphi(i)} - q_{\varphi(i)}$, un couple (p, q) et un transport γ (associé à φ) tel que

$$p_i \geq u_{ij} - q_j \quad \forall i\,,\ j,$$

avec égalité sur le support de γ, et donc (d'après la remarque 15.7) que le plan γ^φ associé à φ est optimal.

Algorithme 15.26. *(Algorithme des enchères)*
On se donne q^0, φ^0. Si, à l'étape n, la collection de prix q^n et la bijection φ^n vérifient (15.11)*, c'est terminé. Dans le cas contraire, on sélectionne un $i^\star$ pour lequel la relation est invalidée, i.e. tel que*

$$u_{i^\star\varphi^n(i^\star)} - q_{\varphi^n(i^\star)} < \max_j \left(u_{i^\star j} - q_j\right).$$

On note $j^\star$ un indice qui réalise le max ci-dessus [84] *:*

$$u_{ij^\star} - q_{j^\star} = \max_j \left(u_{i^\star j} - q_j\right).$$

On attribue alors $j^\star$ à $i^\star$, et $\varphi^n(i^\star)$ à $(\varphi^n)^{-1}(j^\star)$, i.e.

$$\varphi^{n+1}(i^\star) = j^\star\,,\ \varphi^{n+1}\left((\varphi^n)^{-1}(j^\star)\right) = \varphi^n(i^\star)$$

ou, exprimé différemment,

$$\varphi^{n+1} = \varphi^n \circ \tau_{i^\star,(\varphi^n)^{-1}(j^\star)},$$

83. Thèse de Julie Champion, page 53.
`http://thesesups.ups-tlse.fr/2036/1/2013TOU30083.pdf`

84. L'agent $i^\star$ préfèrerait l'objet $j^\star$ qui, en l'état courant des prix, lui apporterait plus de satisfaction (= utilité - prix) que $\varphi^n(i^\star)$.

où τ_{i_1,i_2} est la transposition qui échange i_1 et i_2. On augmente enfin le prix de $j^\star$ d'une quantité qui ramène l'attrait de $j^\star$ pour $i^\star$ au niveau du second produit le plus attractif :

$$q_{j^\star}^{n+1} = q_{j^\star}^n + \max_j (u_{i^\star j} - q_j) - \max_{j \neq j^\star} (u_{i^\star j} - q_j).$$

Cet algorithme est susceptible de patiner dans certains cas, lorsque plusieurs produit réalisent le maximum d'attrait pour un agent (le prix reste alors stationnaire).

On utilise en pratique une version modifiée de l'algorithme, qui visent à trouver une bijection φ et une gamme de prix (q) tels que chaque agent i soit ε-satisfait, c'est à dire que

$$u_{i\varphi(i)} - q_{\varphi(i)} \geq \max_j (u_{ij} - q_j) - \varepsilon. \tag{15.12}$$

Algorithme 15.27. *(Algorithme des enchères modifié)*
On se donne q^0, φ^0. Si, à l'étape n, la collection de prix q^n et la bijection φ^n vérifient (15.12)*, on s'arrête. Dans le cas contraire, on sélectionne un $i^\star$ pour lequel la relation est invalidée, i.e. tel que*

$$u_{i^\star \varphi^n(i^\star)} - q_{\varphi^n(i^\star)} < \max_j (u_{i^\star j} - q_j) - \varepsilon.$$

On note $j^\star$ un indice qui réalise le max ci-dessus

$$u_{ij^\star} - q_{j^\star} = \max_j (u_{i^\star j} - q_j).$$

On attribue alors $j^\star$ à $i^\star$, et $\varphi^n(i^\star)$ à $(\varphi^n)^{-1}(j^\star)$, i.e.

$$\varphi^{n+1}(i^\star) = j^\star\,,\ \varphi^{n+1}\left((\varphi^n)^{-1}(j^\star)\right) = \varphi^n(i^\star).$$

On augmente enfin le prix de $j^\star$ du montant maximum qui préserve son ε-satisfaction :

$$q_{j^\star}^{n+1} = q_{j^\star}^n + \max_j (u_{i^\star j} - q_j) - \max_{j \neq j^\star} (u_{i^\star j} - q_j) + \varepsilon \geq q_{j^\star}^n + \varepsilon.$$

Remarque 15.28. *Noter que, dans cette ε-version de l'algorithme, le bien $j^\star$ choisi par $i^\star$ après une étape n'est pas forcément son meilleur choix (après augmentation du prix de $j^\star$), mais l'agent est tout de même ε-satisfait avec son $j^\star$, et a augmenté les chances de le garder en proposant un prix supérieur (ce qui tendra à écarter les autres agents de ce choix). Les prix des autres produits ne pouvant que croître, la seule chose qui pourrait lui faire renoncer à $j^\star$ est qu'un autre agent s'en empare.*

Cet algorithme, contrairement au précédent, assure une croissance stricte d'un prix à chaque étape. Par ailleurs, lorsqu'un produit est choisi au cours des itérations, il est susceptible de changer ensuite de propriétaire, mais il fera toujours par construction l'ε-bonheur de ce dernier. La non convergence de l'algorithme ne peut donc se produire que si certains produits ne sont jamais considérés. Mais le prix de tels produits resterait alors constant, les autres augmentant strictement, de telle sorte qu'ils finissent à terme par devenir compétitifs, même si leur utilité brute était très faible :

Proposition 15.29. *L'algorithme 15.27 converge après un nombre fini d'itérations.*

Démonstration. Considérons un scénario dans lequel l'algorithme continuerait indéfiniment. D'après la remarque ci-dessus, cela signifie qu'un sous ensemble non vide Y_1 de biens ne fait jamais l'objet d'un choix. On note Y_3 l'ensemble des biens qui sont considérés une infinité de fois, et par Y_2 l'ensemble des biens visités un nombre fini de fois. On se place au-delà de la dernière itération qui a vu un bien de Y_2 pris en compte. Les prix des biens de Y_3 tendent vers $+\infty$, donc, pour tout i, tout j dans Y_3, la quantité $u_{ij} - q_j$ tend vers $-\infty$, donc les biens de Y_3 deviennent uniformément moins compétitifs que les biens de Y_1, ce qui est absurde. □

Montrons que cet algorithme conduit, à convergence, à une approximation d'ordre ε (plus précisément inférieure à $N\varepsilon$) de l'utilité maximale. Rappelons que l'on considère ici un problème de MK renversé, dans le cas de deux ensembles de même cardinal N, et des mesures uniformes (de masse totale N). On cherche en effet ici à maximiser l'utilité globale

$$U(\gamma) = \sum \gamma_{ij} u_{ij},$$

sur Π. Le problème dual consiste à minimiser

$$\sum p_i + \sum q_j$$

sous les contraintes $p_i + q_j \geq u_{ij}$. Si l'on note F la fonction correspondant au problème primal (définie maintenant à partir du lagrangien comme un inf en (p, q)), et G la fonction duale (définie comme un sup en γ), on a une situation renversée par rapport au lemme 23.23, page 245, i.e.

$$F(\gamma) \leq G(p, q) \quad \forall \gamma \in (\mathbb{R}_+)^{N^2},\ (p, q) \in \mathbb{R}^N \times \mathbb{R}^N.$$

Du fait de l'existence d'un point selle démontré au début de cette section (proposition 15.9), on a bien sûr

$$\sup F(\gamma) = \max F(\gamma) = \inf G(p, q) = \min G(p, q).$$

Proposition 15.30. *Pout tout $\varepsilon > 0$, on considère une bijection φ de S_N et un système de prix (q_j) qui vérifient*[85]

$$u_{i\varphi(i)} - q_{\varphi(i)} \geq \max_j (u_{ij} - q_j) - \varepsilon.$$

Alors l'utilité associée à la bijection φ approche l'utilité maximale à $N\varepsilon$ près, i.e.

$$U(\gamma^S) \geq \max_{\Pi} U_\gamma - N\varepsilon.$$

Démonstration. On définit

$$p_i = u_{i\varphi(i)} - q_{\varphi(i)}.$$

On a par hypothèse

$$p_i \geq u_{ij} - q_j - \varepsilon$$

de telle sorte que le couple $(p + \varepsilon, q)$ est admissible. On a donc

$$\max F = \min G \leq G(p + \varepsilon, q) = \sum (p_i + \varepsilon) + \sum q_j = \sum_i \left(p_i + q_{\varphi(i)} \right) + N\varepsilon$$

85. On écrit exactement ici que (φ, q) est un point d'arrêt de l'algorithme des enchères modifié.

$$= \sum_i u_{i\varphi(i)} + N\varepsilon. \leq \max F + N\varepsilon.$$

On a donc $F(\gamma^\varphi) \geq \max F - N\varepsilon$.

□

Implémentation effective en Python de l'algorithme des enchères.

On définit en premier lieu une matrice d'utilités (u_{ij}). Pour le cas du transport optimal (problème d'affectation), on se donne par exemple deux familles de points de $\mathbb{R}^2$, et l'on définit

$$u_{ij} = -\left|y_j - x_i\right|^p.$$

La matrice correspondante est initialisée en Python par `uu = np.zeros((N,N))`. On définit le vecteur des prix comme `q = np.zeros((1,N))`. On peut construire alors la matrice `mm` correspondant à $u_{ij} - q_j$ de la façon suivante :

```
e = np.ones((N,1))
qq = np.matmul(e,q)
mm =  uu-qq
```

Pour une telle matrice, la commande `jjmax = np.argmax(mm,axis=1)` permet de calculer un tableau d'indices correspondant, pour chaque ligne, à la colonne qui réalise le maximum des valeurs. Si l'on dispose d'un vecteur, par exemple la ligne de `mm` correspondan au $i^\star$ sélectionné, on peut récupérer les indices correspondant aux deux plus grands éléments par la commande

```
[next_to_jstar,jstar] = np.argsort(mm[istar,:])[-2:]
```

On encodera l'affectation courante par un tableau d'entiers, initialisé par exemple à `phi = range(N)`.

Remarque 15.31. *On prendra garde au fait que, à chaque itération, l'agent $i^\star$ choisit le (ou un) bien $j^\star$ qui maximise sa satisfaction, mais qu'il en augmente ensuite le prix (pour en écarter les autres) d'un montant qui le rend très exactement $\varepsilon-$ satisfait, mais* pas mieux. *On aura toujours (mathématiquement), du fait de l'augmentation du prix,*

$$u_{i^\star\varphi^{n+1}(i^\star)} - q_{varphi^{n+1}(i^\star)} = \max_j (u_{i^\star j} - q_j) - \varepsilon,$$

où $i^\star$, rappelons-le, est l'agent actif à l'itération n. Si l'on compte à l'itération suivante $n+1$ le nombre de gens ε-satisfaits[86]*, en comptant le nombre d'indices i tels que*

$$u_{i\varphi^n(i)} - q_{\varphi^n(i)} \geq \max_j (u_{ij} - q_j) - \varepsilon,$$

en effectuant un test du type `...>= - eps`*, il est possible que la propriété pour $i^\star$ soit fausse, alors qu'elle devrait être vraie, du fait des erreurs d'arrondis. Même si la réalité mathématique*

86. Il est naturel d'arrêter l'algorithme lorsque ce nombre vaut le nombre total d'agents.

est $a = b$, il est possible qu'informatiquement la propriété `a >= b` *soit fausse (au zéro machine près, c'est à dire autour de 10^{-14}). On pourra contourner cette difficulté en incrémentant le prix d'une quantité légèrement inférieure à ε, par exemple $0.99\,\varepsilon$. De façon générale, on se gardera d'effectuer sur des nombres réels des tests d'égalité, ou d'inégalité large ou stricte lorsque les cas d'égalités sont sensibles*[87].

87. Dans le cas présent il est assez aisé d'identifier la difficulté, puisque en gros une fois sur deux le test sera négatif alors qu'il devrait être positif. Dans d'autres situations, l'égalité n'est pas générique, de telle sorte que, pour des tests portant sur des nombres d'ordre un, on a de l'ordre d'une chance sur 10^{14} de tomber sur un cas ambigu de quasi-égalité. C'est alors évidemment beaucoup plus vicieux, puisque le problème risque de ne se poser qu'après un très grand nombre de tests de l'algorithme.

Troisième partie

Aspects numériques

16 Différences finies

16.1 La méthode

La méthode dite des différences finies, destinée à construire des approximations de solutions d'équations aux dérivées partielles, est basée sur une discrétisation naturelle des dérivées partielles, à partir de la simple expression

$$f'(x) = \frac{f(x+\varepsilon) - f(x)}{\varepsilon} + o(\varepsilon).$$

Considérons par exemple l'**équation de la chaleur** sur l'intervalle $I =]0,1[$, avec conditions de Dirichlet aux extrémités de l'intervalle, sur l'intervalle de temps $[0,T]$:

$$\partial_t u - D\partial_{xx} u = 0\,,\ u(0,\,\cdot\,) = u^0(\,\cdot\,) \text{ donné.}$$

On introduit une discrétisation uniforme de l'intervalle I, de pas $\Delta x = 1/J$:

$$0 = x_0\,,\ x_1 = \Delta x\,,\ \ldots\,,\ x_j = j\Delta x\,,\ \ldots\,,\ x_{J-1} = (J-1)\Delta x\,,\ x_J = J\Delta x, \tag{16.1}$$

et de même pour l'intervalle en temps (de pas $\Delta t = T/N$)

$$0 = t_0\,,\ t_1 = \Delta t\,,\ t_n = n\Delta t\,,\ t_N = N\Delta t = T.$$

On cherche alors à construire des nombres u_j^n qui ont vocation à approcher les valeurs de $u(j\Delta t, n\Delta x)$. On définit tout d'abord les u_j^0 par interpolation de la condition initiale sur le maillage, le cœur de l'approche consiste alors à écrire des relations entre les u_j^n qui permettent de construire sans ambiguïté toutes les valeurs à partir des u_j^0.

Une approche naturelle consiste par exemple à écrire

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} - D\frac{u_{j-1}^n - 2u_j^n + u_{j+1}^n}{(\Delta x)^2} = 0 \quad \forall j = 1,\ldots,J-1, \tag{16.2}$$

ce qui peut s'écrire matriciellement, avec des notations évidentes

$$u^{n+1} = \left(\mathrm{Id} + \frac{D\Delta t}{\Delta x} A\right) u^n,$$

où A est la matrice du Laplacien discret (avec condition de Dirichlet) définie par (A.13). On parle d'un schéma *explicite*, car la discrétisation de l'opérateur de dérivée en espace est basée sur des valeurs déjà calculées. De fait, l'expression ci-dessus permet de calculer les u_j^{n+1} directement, sans résolution d'un système linéaire.

Le schéma *implicite*, dont nous verrons qu'il présente de meilleures propriétés de stabilité, s'écrit

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} - D\frac{u_{j-1}^{n+1} - 2u_j^{n+1} + u_{j+1}^{n+1}}{(\Delta x)^2} = 0 \quad \forall j = 1,\ldots,J-1, \tag{16.3}$$

qui peut s'écrire, avec les même notations que précédemment.

Remarque 16.1. *On peut associer un graphe orienté à chacun des schémas numériques introduits ci-dessus (voir figure 16.1). Le graphe associé au schéma explicite est acyclique, ce qui exprime le fait que les calculs peuvent être faits explicitement en partant des valeurs correspondants aux points maximaux du graphe (condition initiale). Le graphe associé au schéma implicite contient des cycles, ce qui exclut la possibilité de calculer directement les valeurs inconnues. Ce schéma fait en effet intervenir un système linéaire qu'il s'agira de résoudre (de façon exacte ou approchée). Noter que, si l'on connait l'inverse de la matrice impliquée dans le schéma, il devient de fait explicite, avec un graphe de dépendance représenté en bas de la figure 16.1 (chaque point de l'étape $n+1$ est alors relié à chaque point de l'étape n, ce qui exprime le caractère non local de l'inverse du Laplacien discret).*

Considérons maintenant l'équation de transport à vitesse constante $V > 0$ sur $I =]0, 1[$, avec conditions périodiques

$$\partial_t u + V \partial_x u = 0.$$

On considère la discrétisation en espace (16.1), en identifiant maintenant le point 0 et le point J. Le schéma dit *décentré amont* s'écrit

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + V \frac{u_j^n - u_{j-1}^n}{\Delta x} = 0 \quad \forall j = 1, \dots, J \quad (\text{avec } 0 \equiv J), \tag{16.4}$$

le décentré aval est obtenu en discrétisant la dérivée en espace à l'aide de $u_{j+1}^n - u_j^n$. Le schéma centré est basé sur les valeurs de part et d'autre du point considéré : $(u_{j+1}^n - u_{j-1}^n)/2$. On peut aussi considérer des versions implicites de ces différents schéma.

Comme nous le verrons plus loin, ces approches ont des propriétés très différentes en termes de stabilité. On peut en particulier vérifier que le schéma explicite centré est complètement inutilisable en pratique, car instable : il produit génériquement des densitées négatives, et la densité maximale augmente au fil des itérations.

16.2 Consistance, stabilité, convergence

On considère ici une équation aux dérivées partielles d'ordre 1 en temps :

$$\partial_t u + L(u) = f.$$

où L est un opérateur différentiel en espace (typiquement opérateur de transport, ou de diffusion, ou la somme des deux, pour ce qui nous intéresse ici).

Un schéma numérique à deux niveaux consiste en la donnée de relations entre les valeurs $(u^n)_j$ et $(u^{n+1})_j$, qui permet de calculer de façon univoque les secondes à partir des premières :

$$F_j(u^{n+1}, u^n, \Delta t, \Delta x) = 0 \tag{16.5}$$

où l'index j parcours l'ensemble des degrés de liberté en espace. Nous ne considérerons ici que des schémas *linéaires*, qui peuvent s'écrire de façon matricielle [88]

$$u^{n+1} = A u^n. \tag{16.6}$$

88. La matrice A n'est pas nécessairement donnée explicitement ; dans le cas des schémas implicite, cette matrice ne sera d'ailleurs jamais construite (on se contentera en pratique de résoudre des systèmes linéaires pour différents membres de droite).

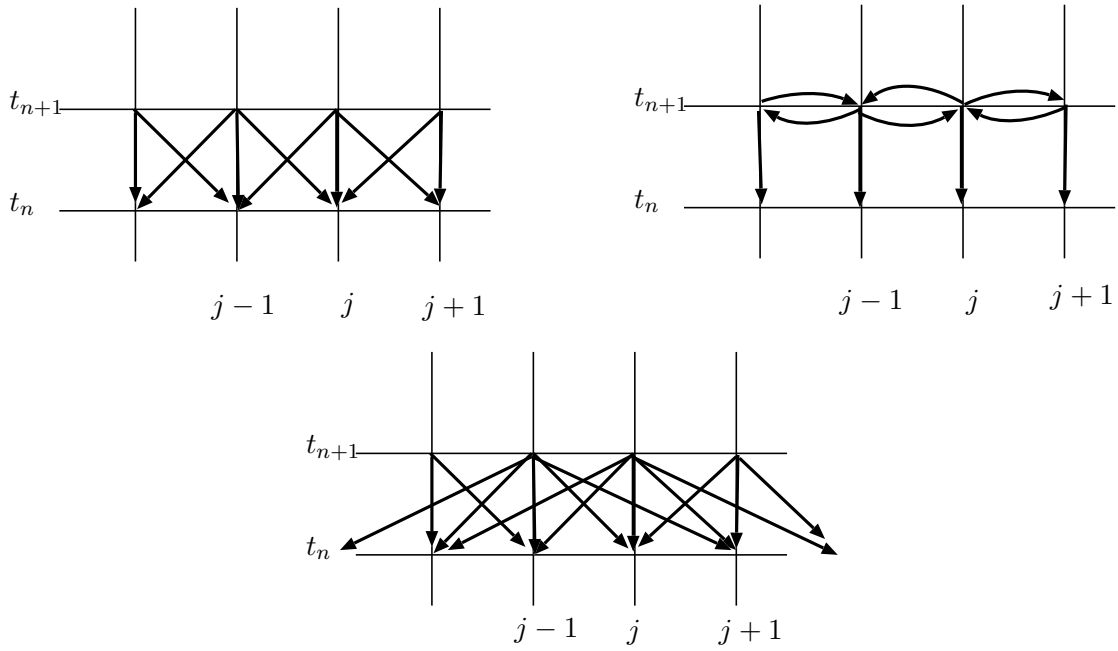


FIGURE 16.1 – Graphes de dépendance associés aux schéma explicite (gauche) et implicite (droite) pour l'équation de la chaleur.

Dans tous les exemples donnés ci-dessus, le schéma est obtenu en remplaçant les dérivées par des expressions faisant intervenir les variables discrètes et les pas de temps et d'espace. Le lien entre l'équation et le schéma peut se préciser grâce à la notion de consistance :

Definition 16.2. *(Consistance)*
On considère un schéma de discrétisation (16.5) *pour une équation aux dérivées partielles. Soit u une solution exacte, régulière, de l'équation. Pour une discrétisation donnée, on note $\tilde{u}$ l'interpolée de la solution exacte aux points de discrétisation, i.e.*

$$\tilde{u}_j^n = u(j\Delta x, n\Delta t).$$

Si

$$F_j(\tilde{u}^{n+1}, \tilde{u}^n, \Delta t, \Delta x) = \mathcal{O}((\Delta x)^q) + \mathcal{O}((\Delta t)^r),$$

uniformément en j et n, on dit que le schéma est consistant, d'ordre q en espace, et r en temps[89].

Remarque 16.3. *Pour lever le flou sur la régularité requise, précisons la démarche l'élaboration d'un schéma de consistance : on considère une solution exacte de l'équation, on lui "applique le schéma". Plus précisément, on applique la relation $F(\cdot)$ à son interpolée, et on*

89. Une petite ambiguïté réside dans le fait que l'on peut multiplier l'ensemble des relations d'un schéma par des puissances de Δt et Δx sans changer les dépendances, tout en affectant l'ordre obtenu dans la définition de la consistance. Nous nous placerons toujours dans le cas où le schéma est de type (16.4) ou (16.3), c'est à dire que, si l'on injecte dans le schéma (comme on l'a fait dans la définition de consistance un fonction régulière en espace temps qui n'est pas la solution exacte, on trouve une quantité finie (ni nulle ni infinie) lorsque Δx et Δt tendent vers 0 .

fait des développements de Taylor-Lagrange de façon à faire apparaître l'équation vérifiée par u, et des restes impliquant Δt, Δx, et des dérivées en espace et en temps de la solution exacte. Ce sont ces dérivées qui vont fixer la régularité requise pour u. Noter que cette définition est d'une certaine manière formelle, elle est afférente au schéma lui-même, on pourrait imaginer un schéma d'ordre très élevé qui discrétise une équation considérée dans un contexte où les solution ne sont jamais aussi régulières qu'il le faudrait pour que les développements soient licites. Cela ne remet pas en question l'ordre du schéma en temps que schéma, en revanche la consistance d'ordre élevé ne permettra pas de montrer une convergence effective de la méthode globale d'approximation d'une solution. Concrètement, les solutions moins régulières seront approchées avec une précision moindre. La consistance correspond ainsi à un ordre de précision indépassable[90].

Nous aurons besoin pour comparer la solution approachée à la solution exacte de définir une distance. Une première étape consiste à construire à partir de la "solution approchée" (qui pour l'instant n'est qu'une collection de valeurs ponctuelles aux points de la discrétisation en espace-temps) une fonction définie partout (ou au moins presque partout). On associe ainsi à une collection u^n de valeurs aux points de discrétisation x_j la fonction constante, égale à u_j^n sur l'intervalle $]x_j - \Delta x/2, x_j + \Delta x/2[$. On notera $\bar{u}^n$ cette fonction.

On peut alors exprimer la norme $\|\bar{u}^n\|_p$ en fonction des valeur discrètes, par exemple pour $p = 1,\ 2,\ +\infty$,

$$\|\bar{u}^n\|_1 = \Delta x \sum_j \left|u_j^n\right| \ , \ \|\bar{u}^n\|_2 = \left(\Delta x \sum_j \left|u_j^n\right|^2\right)^{1/2} \ , \ \|\bar{u}^n\|_\infty = \max_j \left|u_j^n\right| .$$

Noter que toutes les normes p sont dominées par la norme ∞ (uniformément par rapport au nombre de points de discrétisation), et que la consistance a été définie par une majoration uniforme.

Definition 16.4. *(Stabilité)*
On considère un schéma de discrétisation d'une EDP sur un intervalle de temps $[0, T]$. Un schéma numérique est dit (inconditionnellement) stable *(pour la norme p) s'il existe une constante C telle que*

$$\|\bar{u}^n\|_p \leq K \left\|\bar{u}^0\right\|_p \quad \forall n = 1, \ldots, N = T/\Delta t,$$

pour toute donnée initiale discrète $\bar{u}^0$. On parlera de stabilité conditionnelle si la propriété ci-dessus est conditionnée à la vérification d'une relation liant Δt et Δx.

Remarque 16.5. *Il est sous-entendu dans la définition précédente que, dans le cas de stabilité conditionnelle, la condition imposée sur Δt et Δx doit autoriser un "chemin" du couple vers 0, c'est à dire que l'on peut construire une suite du couple $(\Delta t, \Delta x)$ de pas de temps et d'espace vérifiant la condition de stabilité, et telle que $(\Delta t, \Delta x)$ tende vers $(0, 0)$.*

A une solution exacte de l'équation considérée, on associe maintenant une collection de fonctions constantes par moceaux : $\tilde{u}^n$ est la fonction constante par morceaux qui prend la valeur $u(x_j, t^n)$ sur $]x_j - \Delta x/2, x_j + \Delta x/2[$.

90. Sous réserve que les développements de Taylor aient été effectués de façon optimale.

Le théorème suivant établit qu'un schéma consistant et stable est convergent, à l'ordre de consistance.

Théorème 16.6. *(Lax)*
On considère une équations aux dérivées partielles linéaire. On note (u^n) les valeurs approchées obtenues par application d'un schéma numérique consistant à l'ordre q en espace et r en temps vis à vis de cette équation, et stable (pour la norme p). Soit $u(\,\cdot\,,\,\cdot\,)$ une solution de l'équation associée à une condition initiale u_0, définie sur $[0,L]\times[0,T]$. On suppose que u a la régularité en temps et en espace requise pour que l'estimation de consistance soit effective.

On note $(\bar u^n)$ la famille de fonctions constantes par morceaux obtenues par application du schéma numérique, avec $\bar u^0 = \tilde u^0$, et $e^n = \tilde u^n - \bar u^n$. On a convergence de la méthode numérique au sens suivant

$$\lim_{\Delta t, Dx\to 0} \sup_n \|e^n\|_p = 0.$$

On a plus précisément

$$\sup_n \|e^n\|_p \le C\left((\Delta x)^q + (\Delta t)^r\right).$$

Démonstration. Le schéma s'écrit $u^{n+1} = Au^n$. Comme il est consistant, la solution exacte le vérifie approximativement :

$$\tilde u^{n+1} = A\tilde u^n + \Delta t\varepsilon^n \quad \|\varepsilon^n\| \le C\left((\Delta x)^q + (\Delta t)^r\right)$$

(la consistance porte une estimation uniforme de valeurs ponctuelles, elle implique donc bien la même majoration pour toute norme de type L^p). On obtient donc, en faisant la différence, $e^{n+1} = Ae^n - \Delta t\varepsilon^n$, d'où

$$e^n = A^n e^0 - \Delta t \sum_{k=1}^{n} A^{n-k}\varepsilon^{k-1} \le CK\left((\Delta x)^q + (\Delta t)^r\right)$$

□

Stabilité L^2

La stabilité L^2 peut parfois s'établir par une localisation du spectre des matrices impliquées dans le schéma. Mais il existe une méthode très générale qui permet de contourner l'analyse spectrale de la matrice. Cette approche est basée sur la transformée de Fourier, que l'on présente pour simplifier sur l'intervalle $]0,1[$ avec conditions périodiques. À une collection de valeurs $(u_j^n)_j$ on associe comme précédemment une fonction $\bar u^n$ constante par morceaux sur les intervalles centrés en

$$0\,,\ \Delta x\,,\ 2\Delta x\,,\ \ldots\,,\ J\Delta x = 1,$$

(avec identification du dernier point au premier). Cette fonction de L^2 peut s'écrire comme la somme de sa série de Fourier

$$\bar u^n(x) = \sum_{k\in\mathbb{Z}} \hat u^n(k)\exp(2i\pi kx) \ \ p.p. \quad \text{avec} \quad \hat u^n(k) = \int_0^1 \exp(-2i\pi kx)\bar u^n(x)\,dx,$$

et la formule de Plancherel s'écrit

$$\|\bar{u}^n\|_{L^2} = \int_0^1 |\bar{u}^n(x)|^2 \, dx = \sum_{k\in\mathbb{Z}} |\hat{u}^n(k)|^2 .$$

Maintenant, pour $x = j\Delta x$, on a $u_j^n = \bar{u}^n(x)$,

$$u_{j+1}^n = \sum_{k\in\mathbb{Z}} \exp(2i\pi kx)\hat{u}^n(k)\exp(2i\pi k\Delta x),$$

Et une expression similaire pour u_{j-1}^n. Considérons par exemple le schéma explicite (16.2) pour l'équation de la chaleur, en remplaçant dans le schéma les variables discrète par les expressions impliquant la série de Fourier. On obtient une combinaison infinie des $\exp(2i\pi kx)$, qui sont orthogonaux dans L^2. On peut donc écrire que chaque coefficient est nul, i.e. pour tout k on a

$$\hat{u}^{n+1}(k) = \hat{u}^n(k)\left(1 + \frac{D\Delta t}{(\Delta x)^2}\left(\exp(2i\pi k\Delta x) - 2 + \exp(-2i\pi k\Delta x)\right)\right)$$

$$= \hat{u}^n(k)\left(1 + \frac{D\Delta t}{(\Delta x)^2}\left(\exp(i\pi k\Delta x) - \exp(-i\pi k\Delta x)\right)^2\right) = \underbrace{\left(1 - 4\frac{D\Delta t}{(\Delta x)^2}\sin\left(\pi k\Delta x\right)^2\right)}_{A(k)}\hat{u}^n(k).$$

On a appelle $A(k)$ le *coefficient d'amplification* . On a de façon évidente stabilité dès que

$$|A(k)| \ \leq 1 \quad \forall k,$$

ce qui conduit ici à la condition de stabilité

$$\frac{D\Delta t}{(\Delta x)^2} \leq \frac{1}{2}.$$

Cette condition est suffisante, et l'on énoncera en général le résultat de stabilité conditionnelle associé.

Remarque 16.7. *Noter que la condition $|A(k)| \leq 1$ n'est pas nécessaire à strictement parler. Certes, si l'un des coefficient est de module strictement plus grand que 1 et éloigné de 1 uniformément par rapport au pas de temps, on peut trouver une condition initiale (qui excite le mode correspondant) qui soit telle que le schéma ne soit pas stable. Mais il pourrait arriver que le coefficient d'amplification soit majoré par une quantité du type $1 + c\Delta t$, auquel cas on peut avoir stabilité, du fait que*

$$(1 + c\Delta t)^n = (1 + cT/N)^n \leq (1 + cT/N)^N \leq e^{cT}.$$

Dans le cas considéré ici, une telle majoration n'est pas possible lorsque la condition sur le pas de temps est violé, à cause du facteur du type $\sin(2\pi k\Delta x)^2/(\Delta x)^2$, qui est bien majoré pour k petit, mais d'ordre $1/(\Delta x)^2$ pour $k \approx 1/(4\Delta x)$.

16.3 Analyse des principaux schémas numériques

Équation de transport

Proposition 16.8. *Le schéma décentré amont est consistant (d'ordre 1 en temps et 1 en espace) et stable (en norme L^∞ et en norme L^2), donc convergent pour ces deux normes, sous la condition CFL*

$$\Delta t \leq \frac{\Delta x}{V}.$$

Démonstration. On vérifie immédiatement la consistance du schéma. Montrons la stabilité L^∞ (conditionnelle). On a

$$u_j^{n+1} = u_j^n - \frac{V\Delta t}{\Delta x}(u_j^n - u_{j-1}^n) = u_j^n \left(1 - \frac{V\Delta t}{\Delta x}\right) + \frac{V\Delta t}{\Delta x} u_{j-1}^n.$$

Il s'agit d'une combinaison barycentrique des valeurs précédentes dès que $V\Delta t/\Delta x \leq 1$, c'est à dire que l'on a la condition dite CFL :

$$\Delta t \leq \frac{\Delta x}{V}.$$

Sous cette condition, on a stabilité L^∞.

Pour la stabilité L^2, on utilise l'approche décrite précédemment, on a

$$\hat{u}^{n+1}(k) = \hat{u}^n(k) \left(1 - \frac{V\Delta t}{\Delta x}\left(1 - \exp(-2i\pi k \Delta x)\right)\right)$$

qui est bien de module inférieur à 1 pour tout k sous la même condition CFL $\Delta t \leq \Delta x/V$.

□

Le schéma de transport centré est très particulier [91] bizarrement stable pour la norme L^2, mais instable pour la norme L^∞.

Proposition 16.9. *Le schéma centré pour l'équation de transport*

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + V \frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} = 0 \tag{16.7}$$

est instable en norme L^∞, mais stable en norme L^2 sous la condition $\Delta t = \mathcal{O}((\Delta x)^2)$.

Démonstration. Le schéma s'écrit

$$u_j^{n+1} = u_j^n - \lambda u_{j+1}^n + \lambda u_{j-1}^n$$

avec $\lambda = V\Delta t/(2\Delta x)$. On n'a donc pas stabilité L^∞. L'étude de stabilité L^2 conduit à

$$\hat{u}^{n+1}(k) = \hat{u}^n(k)(1 - \lambda \exp(2i\pi k\Delta x) + \lambda \exp(-2i\pi k\Delta x)) = \hat{u}^n(k)\left(1 - 2i\,\lambda \sin\left(2\pi k \Delta x\right)\right)$$

Le coefficient d'amplification est donc de module inférieur à $1 + 2\lambda^2 = 1 + V^2\Delta t^2/2(\Delta x)^2$. Sous une condition du type $\Delta t = \mathcal{O}((\Delta x)^2)$, le coefficient est donc inférieur à $1 + c\Delta t$, d'où la stabilité L^2 (voir remarque 16.7). □

Équation de la chaleur

Proposition 16.10. *Le schéma explicite est consistant (d'ordre 1 en temps et 2 en espace) et stable (en norme L^∞ et en norme L^2), donc convergent pour ces deux normes, sous la condition*

$$\Delta t \leq \frac{(\Delta x)^2}{2D}.$$

91. Il est souvent indiqué comme inconditionnellement instable dans la litérature, et de fait peut utilisé en pratique pour l'équation de transport simple.

Démonstration. Le schéma explicite pour l'équation de la chaleur s'écrit

$$u_j^{n+1} = u_j^n \left(1 - \frac{2D\Delta t}{(\Delta x)^2}\right) + \frac{D\Delta t}{(\Delta x)^2} u_{j-1}^n + \frac{D\Delta t}{(\Delta x)^2} u_{j+1}^n,$$

qui est bien une combinaison barycentrique des veleurs précédentes sous la condition $\Delta t \leq (\Delta x)^2/2D$.

Pour la stabilité L^2, on écrit

$$\hat{u}^{n+1}(k) = \hat{u}^n(k) \left(1 + \frac{D\Delta t}{(\Delta x)^2} \left(\exp(2i\pi k\Delta x) - 2 + \exp(-2i\pi k\Delta x)\right)\right)$$

$$= \hat{u}^n(k) \left(1 + \frac{D\Delta t}{(\Delta x)^2} \left(\exp(i\pi k\Delta x) - \exp(-i\pi k\Delta x)\right)^2\right) = \hat{u}^n(k) \left(1 - \frac{4D\Delta t}{(\Delta x)^2} \sin\left(\pi k\Delta x\right)^2\right)$$

qui est bien de module ≤ 1 sous la même condition sur le pas de temps. □

Proposition 16.11. *Le schéma implicite est consistant (d'ordre 1 en temps et 2 en espace) et inconditionnellement stable en norme L^2 et en norme L^∞, donc convergent pour ces deux normes.*

Démonstration. Stabilité L^∞ : on a, pour tout j,

$$u_j^{n+1} + \lambda(u_j^{n+1} - u_{j-1}^{n+1}) + \lambda(u_j^{n+1} - u_{j-1}^{n+1}) = u_j^n.$$

On en déduit que le plus petit u_j^{n+1} est supérieur à u_j^n, donc supérieur au plus petit des u_ℓ^{n+1}, et que le plus grand u_j^{n+1} est de la même manière inférieur au plus grand u_ℓ^{n+1} (principe du maximum), d'où la stabilité L^∞.

Pour la stabilité L^2, on a

$$\hat{u}^{n+1}(k) = \hat{u}^n(k) \left(1 + \frac{4D\Delta t}{(\Delta x)^2} \sin\left(\pi k\Delta x\right)^2\right)^{-1},$$

d'où l'inconditionnelle stabilité L^2.

□

Exercice 16.1. Étudier (consistance et stabilité L^2) le θ-schéma pour l'équation de la chaleur

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + \theta D \frac{-u_{j-1}^{n+1} + 2u_j^{n+1} - u_{j+1}^{n+1}}{(\Delta x)^2} + (1-\theta) D \frac{-u_{j-1}^n + 2u_j^n - u_{j+1}^n}{(\Delta x)^2} = 0 \qquad (16.8)$$

en fonction de la valeur de θ. Montrer en particulier que le schéma est inconditionnellement stable pour tout $\theta \in [1/2, 1]$.

16.4 Symboles discret et continu des opérateurs différentiels

Considérons une équation d'évolution du type

$$\partial_t u + Lu = 0,$$

sur l'intervalle $]0,1[$ périodique, où L est un opérateur différentiel linéaire (combinaison linéaire de dérivées partielles en espace de u). On écrit la solution sous la forme de sa série de Fourier

$$u(x,t) = \sum_{\mathbb{Z}} \hat{u}_t(k) \exp(2i\pi kx),$$

avec, pour chaque coefficient de Fourier, l'équation différentielle

$$\frac{d}{dt}\hat{u}_t(k) + \hat{L}(k)u(k) = 0,$$

où $\hat{L}(k)$ est le symbole de l'opérateur L. Pour l'équation de la chaleur, on a par exemple

$$Lu = -D\partial_{xx}u\,,\ \hat{L}(k) = D(2\pi)^2k^2,$$

et pour le transport

$$Lu = V\partial_x u\,,\ \hat{L}(k) = 2i\pi kV.$$

Si l'on discrétise en temps (par un schéma d'Euler explicite) l'équation différentielle sur $\hat{u}_t(k)$, on obtient

$$\hat{u}^{n+1}(k) = \hat{u}^n(k)\left(1 - \Delta t\hat{L}(k)\right).$$

Il apparaît qu'un tel schéma est génériquement instable pour les modes grands ($\hat{L}(k)$ est un polynôme en k). La seule possibilité pour qu'un tel schéma soit stable est que $\hat{L}(k)$ soit de degré zéro, donc constant, c'est à dire que l'opérateur ne soit en fait pas un opérateur différentiel. Pour la méthode des différences finies, on peut espérer avoir stabilité dans les cas non triviaux car la discrétisation en espace fait disparaître les hautes fréquences. Par exemple, dans le cas de la chaleur $L = -D\partial_{xx}$, ce qui joue le rôle du symbole de l'opérateur est

$$\frac{D}{(\Delta x)^2}\left(\exp(2i\pi k\Delta x) - 2 + \exp(-2i\pi k\Delta x)\right) = 4\frac{D}{(\Delta x)^2}\sin(\pi k\Delta x)^2$$

qui est bien équivalent à $4\pi^2k^2$, symbole de l'opérateur $-D\partial_{xx}$, quand Δx tend vers 0 (on retrouve la notion de *consistance* dans le domaine spectral). En revanche le symbole discret n'est pas un polynôme, ou plutôt c'est un polynôme en $\exp(2i\pi k\Delta x)$ et $\exp(-2i\pi k\Delta x)$. Il est donc uniformément borné par rapport au mode k, et l'on peut espérer avoir stabilité dès que $1 - \Delta t\hat{L}(k)$ est dans le disque unité pour tout k (cette condition n'est pas nécessaire à strictement parler, voir remarque 16.7, mais la plupart des schémas stables explicites rencontrés vérifieront de fait cette condition). Pour l'équation de la chaleur, le symbole est réel, avec $0 \leq \hat{L}(k) \leq 4D/(\Delta x)^2$, on a donc stabilité sous condition sur le pas de temps, comme vu précédemment (voir figure 16.2).

Pour le transport, la situation est la suivante : le symbole de l'opérateur continu est imaginaire pur, il vaut $2i\pi k$, de telle sorte que $\left|1 - \Delta t\,\hat{L}(k)\right| > 1$ pour tout $k \neq 0$. Une discrétisation en espace appropriée (schéma décentré amont en l'occurrence) permet de "tordre" le symbole de façon à se ramener dans le disque unité, ce qui assure la stabilité sous condition sur le pas de temps. Plus précisément, pour le schéma décentré amont, le symbole discret est

$$\Lambda(k) = \frac{V}{\Delta x}\left(1 - \exp(-2i\pi k\Delta x)\right)$$

qui est bien équivalent au symbole continu, à k fixé, quand Δx tend vers 0. Mais il n'est pas imaginaire pur, il fait un angle $2\pi k\Delta x$ avec le symbole continu, de telle sorte que

$$|1 - \Delta t\Lambda(k)| \leq 1 \text{ dès que } \Delta t \leq V/\Delta x.$$

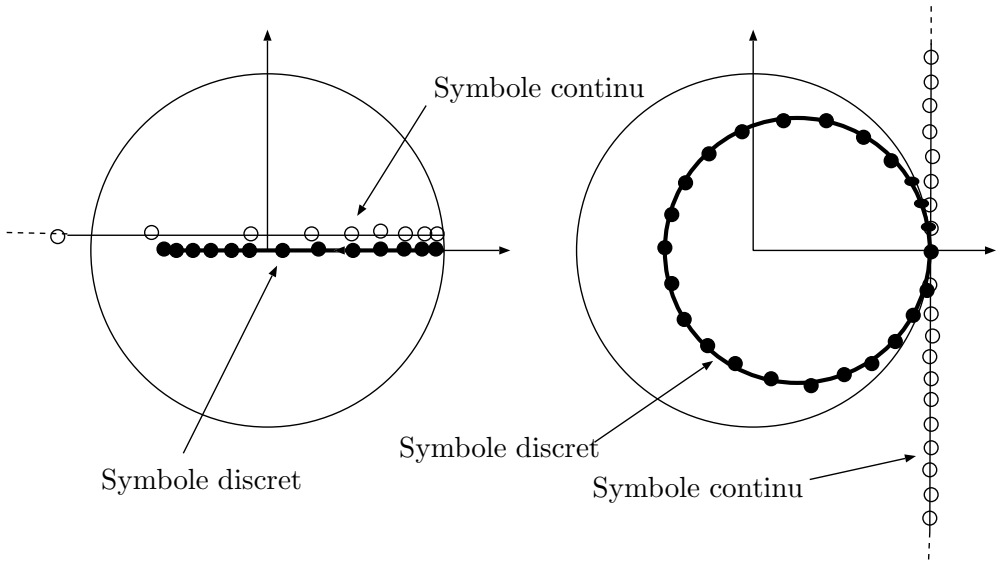


FIGURE 16.2 – Image des symboles discrets (ronds noirs) et continus (ronds blancs) pour l'équation de la chaleur (gauche) et l'équation de transport (droite)

Cette stabilisation par discrétisation s'accompagne d'un phénomène dit de *diffusion numérique*, qui apparaît clairement au niveau spectral. Le symbôle de l'opérateur continu, $2i\pi k$, est imaginaire pur, ce qui reflète le transport sans déformation des modes associés à toutes les fréquences : la solution de

$$\frac{d}{dt}\hat{u}_t(k) = -\hat{L}(k)\,\hat{u}_t(k) = -2i\pi kV\,\hat{u}_t(k)$$

est bien de module constant. Par discrétisation en espace, chaque mode $2i\pi kV$ est remplacé par un mode tourné $V\left(1-\exp(-2i\pi k\Delta x)\right)/\Delta x$, qui stabilise l'évolution, mais qui n'est plus imaginaire pur, on a une partie réelle non triviale

$$\mathrm{Re}(\Lambda) = \frac{V}{\Delta x}\left(1-\cos(2\pi k\Delta x)\right).$$

Le pendant discrétisé en espace de l'équation différentielle ci-dessus est

$$\frac{d}{dt}\hat{u}_t(k) = -\Lambda(k)\,\hat{u}_t(k) = -\frac{V}{\Delta x}\left(1-\exp(-2i\pi k\Delta x)\right)\,\hat{u}_t(k),$$

qui correspond à une décroissance exponentielle vers 0 pour les modes non triviaux : tous les modes oscillants sont amortis.

Dans le processus d'évolution des modes de Fourier de la solution discrète, celà conduit au fait que les coefficients d'amplification $A(k) = (1-\Delta t\Lambda(k))$ sont de module strictement inférieur à un, ce qui entraine une diminution des poids des modes correspondants. Cet amortissement des poids, d'autant plus important que la fréquence est élevée, induit une régularisation de la solution discrète au fil des itérations (alors que l'équation de transport n'est pas elle-même régularisante).

On peut quantifier plus précisément ce phénomène de diffusion numérique, ainsi que la manière dont la discrétisation en espace modifie la vitesse de transport des modes de Fourier

de haute fréquence. Pour le problème continu, on a

$$\frac{d}{dt}\hat{u}_t(k) = -\hat{L}(k)\,\hat{u}_t(k) = -2i\pi kV\,\hat{u}_t(k).$$

Pour le mode k, i.e. $\exp(2i\pi kx)$, l'évolution du coefficient est donnée par $\hat{u}_t(k) = \exp(-2i\pi kVt)$, d'où, pour la fonction elle-même

$$\exp(2i\pi kVt)\exp(2i\pi kx) = \exp(2i\pi k(x - Vt)),$$

qui correspond bien à un transport à vitesse V. Pour le problème discrétisé en espace, on a

$$\frac{d}{dt}\hat{u}_t(k) = -\Lambda(k)\,\hat{u}_t(k) = -\frac{V}{\Delta x}\left(1 - \exp(-2i\pi k\Delta x)\right)\,\hat{u}_t(k)$$

$$= -\frac{V}{\Delta x}\exp(-i\pi k\Delta x)\left(\exp(i\pi k\Delta x) - \exp(-i\pi k\Delta x)\right) = -2i\frac{V}{\Delta x}\exp(-i\pi k\Delta x)\sin(\pi k\Delta x).$$

La solution selon ce mode k s'écrira donc

$$\exp(-\Lambda(k)t)\exp(2i\pi kx).$$

La partie *réelle* de $-\Lambda(k)t$, qui vaut

$$\mathrm{Re}(-\Lambda(k)) = -2\frac{V}{\Delta x}\sin(\pi k\Delta x)^2 < 0,$$

correspond à l'amortissement parasite (phénomène de diffusion numérique). Noter que cet amortissement est asymptotiquement nul si l'on fait tendre Δx, à k fixé, vers 0, ce qui reflète le caractère non diffusif de l'équation de départ. La partie *imaginaire* de $-\Lambda(k)$ encode la propagation dans l'espace du mode considéré :

$$\mathrm{Im}(-\Lambda(k)) = -2\frac{V}{\Delta x}\cos(\pi k\Delta x)\sin(\pi k\Delta x) = -\frac{V}{\Delta x}\sin(2\pi k\Delta x).$$

La partie de la solution associée à ce mode imaginaire s'écrit en effet

$$\exp\left(-i\frac{V}{\Delta x}\sin(2\pi k\Delta x)t\right)\exp(2i\pi kx) = \exp\left(2i\pi k\left(\underbrace{x - \frac{V}{2\pi k\Delta x}\sin(2\pi k\Delta x)t}_{=x-V_kt}\right)\right),$$

qui correspond, pour le mode k, à une propagation à vitesse constante

$$V_k = \frac{V}{2\pi k\Delta x}\sin(2\pi k\Delta x).$$

On retrouve bien la vitesse V lorsque, à k fixé, Δx tend vers 0 (ce qui traduit un nouvelle fois, dans le domaine spectral, la consistance du schéma vis-à-vis de l'équation), mais la vitesse est réduite pour les hautes fréquences (phénomène de *dispersion* numérique).

Remarque 16.12. *Noter que cette étude de l'évolution des modes de Fourier est analogue à l'étude de la propagation des perturbations pour le modèle de trafic routier ou piéton linéarisé autour de la solution d'équilibre, dans le cas d'une route périodique.*

Remarque 16.13. *(Supériorité des schémas implicites)*
Il semble intuitif qu'un schéma implicite possède de meilleures propriétés de stabilité qu'un schéma explicite. Le cadre présenté ci-dessus permet de formaliser cette tendance. Nous limiterons le cadre de cette remarque à des opérateurs différentiels nativement stabilisant dans L^2, c'est à dire ceux dont le symbole reste dans le demi plan complexe $Re(z) \geq 0$ (ce qui est bien le cas pour les opérateurs de diffusion et de transport). On a en effet, pour le mode k,

$$\frac{d}{dt}\hat{u}_t(k) = -\hat{L}(k)\,\hat{u}_t(k),$$

et donc décroissance du (module du) coefficient correspondant au mode k dès que $Re(\hat{L}(k)) \geq 0$. Pour le problème semi-discrétisé en temps, l'approche explicite s'écrit

$$\hat{u}^{n+1}(k) = \left(1 - \Delta t\hat{L}(k)\right)\hat{u}^n(k)$$

d'où, comme on l'a vu précédemment, une instabilité inconditionnelle sauf dans les cas triviaux. Le schéma implicite s'écrit

$$\hat{u}^{n+1}(k) = \left(1 + \Delta t\hat{L}(k)\right)^{-1}\hat{u}^n(k),$$

avec $\left(1 + \Delta t\hat{L}(k)\right)$ à l'extérieur du disque unité, donc stabilité inconditionnelle.

Pour le problème discrétisé en espace par différences finies, on peut énoncer les faits suivants. Si la discrétisation en espace préserve la propriété de positivité de la partie réelle du symbole, i.e. $Re(\Lambda(k)) \geq 0$, le schéma explicite (discrétisé en espace temps, exprimé sur les modes de Fourier) s'écrit

$$\hat{u}^{n+1}(k) = (1 - \Delta t\Lambda(k))\,\hat{u}^n(k),$$

et l'on a au mieux *une stabilité conditionnelle*[92]*. Toujours sous l'hypothèse $Re(\Lambda(k)) \geq 0$, le schéma implicite*

$$\hat{u}^{n+1}(k) = (1 + \Delta t\Lambda(k))^{-1}\,\hat{u}^n(k),$$

assure la décroissance des coefficients de tous les modes, donc stabilité sans condition sur le pas de temps.

Les choses sont un peu plus troubles pour un schéma qui ne vérifierait pas la propriété de symbole à partie réelle positive. Disons que, dans ce cas, l'implicitation ne suffit pas en général pour stabiliser le schéma. Considérons par exemple le schéma décentré aval pour l'équation de transport ; le schéma explicite s'écrit

$$\hat{u}^{n+1}(k) = (1 - \Delta t\Lambda(k))\,\hat{u}^n(k)\,,\ \Lambda(k) = \frac{V}{\Delta x}\ (\exp(2i\pi k\Delta x) - 1)\,,$$

on a cette fois instabilité inconditionnelle : le symbole discret pointe dans la mauvais direction (vers les parties réelles positives), la situation est donc désespérée. Le schéma implicite s'écrirait

$$\hat{u}^{n+1}(k) = (1 + \Delta t\Lambda(k))^{-1}\,\hat{u}^n(k)$$

92. Stabilité conditionnelle avec décroissance de la norme L^2 si l'on peut assurer que $(1 - \Delta t\Lambda(k))$ reste dans le disque unité pour tout k, ou éventuellement stabilité conditionnelle avec condition renforcée, et perte de la propriété de décroissance de la norme L^2, dans le cas où $(1 - \Delta t\Lambda(k))$ sort du disque unité tout en restant dans le demi-espace $\mathrm{Re}(z) \leq 1$ (comme pour le schéma centré explicite, voir proposition 16.7.

Ici, pour les pas de temps grands*, on peut espérer avoir stabilité, mais pour Δt tendant vers 0 on aura toujours apparition de coefficients d'amplification de module > 1. Le fait que le schéma soit stable pour de grands pas de temps n'est évidemment d'aucun intérêt, puisqu'il exclut toute convergence du schéma (voir remarque 16.5).*

16.5 Interprétation probabiliste de schémas explicites

Certains schémas de discrétisation par différences finies peuvent s'interpréter de façon probabliste. L'équation de la chaleur pouvant exprimer un processus de diffusion, il n'est pas surprenant que sa discrétisation puisse être interprétée comme une marche aléatoire. C'est plus inattendu pour l'équation de transport, dont la discrétisation conduit à un phénomène de *diffusion numérique*, dont on propose ici une interprétation stochastique.

Schéma explicite pour la chaleur. On se place dans le cadre périodique, avec $x_0 = 0$ identifié à $x_J = 1$. Le schéma (16.2), page 144, peut s'écrire

$$u_j^{n+1} = \left(1 - \frac{D\Delta t}{(\Delta x)^2}\right) u_j^n + \frac{D\Delta t}{(\Delta x)^2} u_{j-1}^n + \frac{D\Delta t}{(\Delta x)^2} u_{j+1}^n \quad \forall j = 0, \dots, J-1,$$

(avec la convention naturelle $0 \equiv J$ et $-1 \equiv J-1$). Considérons $u^n = (u_j^n)_{0 \le j \le J-1}$ comme une mesure discrète de probabilité, le schéma s'écrit

$$u^{n+1} = {}^tP u^n,$$

avec [93]

$${}^tP = \begin{pmatrix} 1-2\lambda & \lambda & 0 & \cdot & \cdot & \lambda \\ \lambda & 1-2\lambda & \lambda & 0 & \cdot & \cdot \\ 0 & \lambda & \cdot & \cdot & & \cdot \\ \cdot & & \cdot & \cdot & \cdot & \cdot \\ \cdot & & & \cdot & 1-2\lambda & \\ \lambda & \cdot & \cdot & 0 & \lambda & 1-2\lambda \end{pmatrix}$$

Pour $\lambda \le 1/2$ (condition de stabilité L^∞), la matrice P est une matrice stochastique : tous ses éléments sont positifs ou nuls, et la somme des éléments de chaque ligne vaut 1). On peut interpréter les éléments de la ligne i comme des probablités de transition partant de i. La marche aléatoire sous-jacente est définie comme suit : partant de i la probabilité de rester sur place est $1 - 2\lambda$, et la probabilité résiduelle 2λ se partage équitablement entre $i-1$ et $i+1$ (en tenant compte de la périodicité). Cette chaine de Markov est irréductible et réversible, et la mesure stationnaire associée est la mesure discrète uniforme, qui minimise l'entropie (voir section 10, page 101).

Schéma explicite pour le transport. On se place dans le cadre périodique, avec $x_0 = 0$ identifié à $x_J = 1$. Le schéma (16.4), page 145, peut s'écrire

$$u^{n+1} = {}^tP u^n,$$

93. nous écrivons ${}^tP$ bien que la matrice soit symétrique, car c'est bien ${}^tP$ qui interviendra dans les cas non symétriques.

avec

$$ {}^tP = \begin{pmatrix} 1-\lambda & 0 & 0 & \cdot & \cdot & \lambda \\ \lambda & 1-\lambda & 0 & 0 & \cdot & \cdot \\ 0 & \lambda & \cdot & \cdot & & \cdot \\ \cdot & & \cdot & \cdot & \cdot & \cdot \\ \cdot & & & \cdot & 1-\lambda & \\ 0 & \cdot & \cdot & 0 & \lambda & 1-\lambda \end{pmatrix} $$

La matrice P est stochastique pour $\lambda V \Delta t/\Delta x \leq 1$ (condition CFL). La marche aléatoire sous-jacente est définie comme suit : partant de i la probabilité de rester sur place est $1-\lambda$, et la probabilité d'avancer d'une case est λ, avec $\lambda = V\Delta t/\Delta x$.

Cas général. De façon générale, considèrons une équation de conservation, du type

$$\partial_t u + L(u) = 0$$

où L est un opérateur différentiel linéaire exprimant une conservation, i.e. de la forme $\partial_x F(u)$, où F est lui-même un opérateur différentiel linéaire (d'ordre 0 dans le cas du transport simple).

On considère maintenant un schéma de discrétisation par différences finies, du type (explicite)

$$u^{n+1} = (\mathrm{Id} + \Delta t\, A)\, u^n,$$

où A est une discrétisation consistante de l'opérateur $\partial_x(F(u))$. Si le schéma respecte la propriété de conservation, i.e. la somme des u_j^n se conserve [94], alors [95] la somme des éléments d'une colonne de A vaut 0 : le schéma se met sous la forme

$$u^{n+1} = {}^tP u^n,$$

où P est une matrice stochastique.

Dans les cas considérés précédemment, la matrice $\mathrm{Id} + \Delta t A = {}^tP$ est en fait bistochastique, les sommes des éléments d'une ligne valent également 1. Cette propriété reflète simplement une propriété commune aux deux équations considérées, qui admettent (dans le cas périodique) toute fonction constante comme solution stationnaire. Le pendant stochastique de cette propriété est que la mesure stationnaire associée à la chaîne de Markov représentée par la matrice P est la mesure uniforme.

Plans de transport

Les matrices ${}^tP$ associés aux schémas explicites rappelés ci-dessus peuvent (sous condition CFL assurant le principe du maximum), comme toute transposée de matrice stochastique s'interpréter comme des plans de transports entre mesures discrètes portées par un ensemble

94. Cette condition est vérifiée par tous les schémas consistants usuels, même si la consistance n'implique pas, à strictement parler, la préservation exacte de cette propriété de conservation.

95. Toute matrice réelle qui laisse inchangée la somme des éléments de tout vecteur est la transposée d'une matrice stochastique, il suffit d'écrire la condition sur chaque vecteur de base.

de cardinal J. Le fait que la matrice soit bistochastique dans les cas considérés permet aussi de les voir comme un transport particulier entre la mesure uniforme sur un ensemble X_J à J points vers elle même. Ou, pour rester dans un cadre probabiliste, comme la loi d'une variable aléatoire dans $X_J \times X_J$, dont les projections respectives suivent la loi uniforme.

Exercice 16.2. (Diffusion numérique, point de vue du transport optimal)
On considère le plan de transport associé au schéma explicite décentré amont pour l'équation de transport à vitesse constante. On fixe le pas d'espace Δx. Estimer le coût quadratique de transport associé à ce plan, et préciser son comportement lorsque le pas de temps tend vers 0.

16.6 Extensions, développements

Exercice 16.3. On considère le schéma décentré amont appliqué à l'équation de transport à vitesse constante, en domaine (monodimensionnel) périodique. On considère une condition initiale positive, de masse 1, on peut ainsi voir la collections des valeurs au temps t^n comme la loi d'une variable aléatoire discrète. Montrer que, pour une CFL strictement supérieure à 1, l'entropie est décroissante, i.e.

$$S(u^{n+1}) < S(u^n),$$

dès que u^n n'est pas la loi uniforme. En déduire le comportement du schéma, pour Δx et Δt fixés, lorsque le nombre de pas de temps tend vers l'infini.

Équation des ondes.

S'il est possible d'utiliser des schémas à 3 niveaux pour les équations d'ordre 1 en temps comme celles vues précédemment (cela peut permettre d'augmenter l'ordre de précision en temps), cela devient indispensable pour des équations qui sont nativement d'ordre 2 en temps, comme l'équation des ondes

$$\partial_{tt} u - c^2 \partial_x x u = 0.$$

Un schéma couramment utilisé est le schéma de Crank-Nicholson, i.e.

$$\frac{u_j^{n+1} - 2u_j^n + u_j^{n-1}}{(\Delta t)^2} + \theta c^2 \frac{-u_{j-1}^{n+1} + 2u_j^{n+1} - u_{j+1}^{n+1}}{(\Delta x)^2} + (1-\theta) c^2 \frac{-u_{j-1}^{n} + 2u_j^{n} - u_{j+1}^{n}}{(\Delta x)^2} = 0 \quad (16.9)$$

avec $\theta = 1/2$, qui peut s'écrire matriciellement

$$\left(\mathrm{Id} + \frac{c^2(\Delta t)^2}{2(\Delta x)^2} A \right) u^{n+1} = 2u^n - u^{n-1} - \frac{c^2(\Delta t)^2}{2(\Delta x)^2} A u^n,$$

où A est la matrice du Laplacien discret.

Implémentation effective

Les schémas explicites ne nécessitent en général pas l'assemblage de la matrice. On pourra utiliser avantangeusement les opérateurs de shift à droite S_R et shift à gauche S_R définis, dans un cadre périodique, par

$$S_R(u_1, u_2, \dots u_J) = (u_J, u_1, \dots, u_{J-1}), \; S_L(u_1, u_2, \dots u_J) = (u_2, u_3, \dots, u_J, u_1).$$

En Python, les opérateurs de shift peuvent être implémentées simplement de la façon suivante :

```
uuL = np.roll(uu,-1)
uuR = np.roll(uu,1)
```

Transport. Le schéma décentré amont (la vitesse d'advection est choisie positive) s'écrit ainsi, avec des notations évidentes

$$u^{n+1} = u^n - \frac{V\Delta t}{\Delta x}\left(u^n - S_R u^n\right),$$

et le schéma centré :

$$u^{n+1} = u^n - \frac{V\Delta t}{2\Delta x}\left(S_L u^n - S_R u^n\right).$$

Diffusion.

Le schéma explicite pour l'équation de la chaleur peut être implémenté (cas périodique) en utilisant les opérateurs de shift :

$$u^{n+1} = u^n + \frac{D\Delta t}{(\Delta x)^2}\left(S_R u^n - 2u^n + S_L u^n\right),$$

qui se programme simplement en Python à l'aide de la méthode `np.roll` évoquée précédemment.

Si l'on s'intéresse à des conditions de Dirichlet homogènes, le plus simple est de définir un vecteur de taille $J+1$ (qui contient les valeurs aux extrémités, qui ne sont pas des degrés de libertés), d'initialiser les valeurs extrémales (qui ne seront pas modifiées par le schéma) aux valeurs imposées, et d'incrémenter le sous-vecteur qui correspond effectivement aux degrés de liberté.

Construction des matrices. Pour les schémas implicites, il est naturel[96] d'assembler la matrice intervenant dans le schéma. Il est essentiel de stocker les matrices sous forme creuse, pour limiter le temps de calcul. Le package `scipy` permet de stocker les matrices sous cette forme, et propose des méthodes de résolution optimisées pour ce type de matrices.

```
import scipy.sparse as ssp
import scipy.sparse.linalg as sla
```

96. Cet assemblage n'est pas nécessaire à strictement parler. On peut être amené à utiliser, pour résoudre le système linéaire, des méthodes dites itératives (voir section 18, page 179), basées sur des produits matrice-vecteur successifs. Si l'on programme soi-même l'une de ces méthodes itératives, on peut choisir d'effectuer ces produits matrice-vecteur à la volée, sans préassembler la matrice. Cette approche permet d'éconimiser de l'espace mémoire dans le cas où la matrice contient très peu l'élements différents, ce qui est le cas des matrices résultant de la discrétisation d'opérateurs différentiels invariants par translation, sur un maillage régulier.

La manière la plus simple d'assembler les matrices résultant d'une discrétisation par différences finie est de passer par la commande `ssp.diags`, qui prend en argument des un tableau de vecteurs correspondant aux diagonales non nulles, suivies des indices correspondant aux diagonales (0 pour la diagonale, indices positifs pour la partie triangulaire supérieure, et négatifs de l'autre côté). On pourra par exemple assembler la matrice associée au schéma de transport implicite, i.e.

$$A = \begin{pmatrix} 1 & \beta & 0 & \cdot & \cdot & -\beta \\ -\beta & 1 & \beta & 0 & \cdot & \cdot \\ 0 & -\beta & \cdot & \cdot & & \cdot \\ \cdot & & \cdot & \cdot & \cdot & \cdot \\ \cdot & & & -\beta & 1 & \beta \\ \beta & \cdot & \cdot & 0 & -\beta & 1 \end{pmatrix}$$

avec $\beta = \Delta t V/(2\Delta x)$, de la façon suivante

```
beta = 0.5*V*dt/dx
ones  = np.ones(J)
aux   = [ones,beta*ones[:-1],-beta*ones[:-1],-beta*ones[0],beta*ones[0]]
Adv1d = ssp.diags(aux,[0,1,-1,(J-1),-(J-1)],format='csr')
```

Le calcul du nouveau champ à partir du précédent peut alors se faire à l'aide de la fonction `spsolve` du package `scipy.sparse.linalg` :

```
uu =sla.spsolve(Adv1d,uu)
```

N.B. Le format `csr`[97] spécifié lors de l'assemblage permet une utilisation optimale de `solve`.

Assemblage des matrices du Laplacien en dimension $d \geq 2$.

En dimension 1 la matrice du Laplacien discret avec conditions de Dirichlet (valeur imposée à 0 aux extrémités) s'écrit

$$A_1 = \begin{pmatrix} 2 & -1 & 0 & \cdot & \cdot & 0 \\ -1 & 2 & -1 & 0 & \cdot & \cdot \\ 0 & -1 & \cdot & \cdot & & \cdot \\ \cdot & & \cdot & \cdot & \cdot & \cdot \\ \cdot & & & \cdot & 2 & -1 \\ 0 & \cdot & \cdot & 0 & -1 & 2 \end{pmatrix}$$

97. Voir `http://perso.univ-perp.fr/langlois/images/pdf/mp/scipy.pdf`

En dimension 2 d'espace, le Laplacien discret agit sur les valeurs au point $(i\Delta x, j\Delta x)$ de la discrétisation comme suit

$$4u_{i,j} - u_{i-1,j} - u_{i+1,j} - u_{i,j-1} - u_{i,j+1}.$$

On peut vérifier que la matrice associée peut s'écrire

$$A_2 = A_1 \otimes I_1 + I_1 \otimes A_1,$$

où I_1 est la matrice identité d'ordre le nombre de point dans chaque direction, et $\oplus$ est le produit de *Kronecker* défini de la façon suivante : si $A \in \mathcal{M}_{pq}$ et B_{rs} sont deux matrices, la matrice $C = A \otimes B$ est de taille (pr, qs) a une structure (p, q) par blocs, chaque bloc étant de taille (r, s), égale au produit de a_{ij} par la matrice B. On obtient de façon analogue la matrice du Laplacien 2d pour des conditions aux limites de Neuman, ou des conditions périodiques.

En Python, si `A` et `B` sont des matrices creuses, ce produit de Kronecker s'écrit

```
C = ssp.kron(A,B)
```

Exercice 16.4. Généraliser la construction décrite ci-dessus au cas de la dimension 3.

Exercice 16.5. Proposer une extension de l'approche dans le cas de conditions aux limites panachées, par exemple, sur le carré unité, le cas de conditions de Neuman homogènes le bord $[y = 0]$, et Dirichlet homogène partout ailleurs.

Résolution de grands systèmes linéaires. La résolution de problmème d'évolution par un schéma implicite conduit à la résolution de multiples systèmes linéaires impliquant la même matrice, pour des seconds membres différents (voir remarque 18.7, page 181, dans le cas de la factorisaiton de Cholesky). On peut alors avoir intérêt à pratiquer une pré-factorisation de la matrice, qui va pouvoir ensuite être utilisée pour tous les systèmes.

L'implémentation en Python prend la forme suivante : on convertit tout d'abord la matrice au format approprié, dit `csc`, par `A=A.tocsc()`, puis on factorise la matrice par `fA = sla.factorized(A)`.

La résolution du système s'écrit ensuite comme un simple appel de fonction (comme si `fA` était l'inverse de la matrice A) :

```
uu = fA(rhs)
```

17 Éléments finis

17.1 La méthode

On considère le problème de Poisson dans le domaine le domaine $\Omega =]0,1[\times]0,1[$.

$$\left\{ \begin{array}{rcll} -\Delta u & = & f & \text{dans } \Omega \\ u & = & 0 & \text{sur } \Gamma \end{array} \right. \tag{17.1}$$

Formulation variationnelle. On obtient [98] la formulation variationnelle de ce problème en multipliant la première équation par une fonction test v régulière qui s'annule sur la partie du bord où la température est imposée. On obtient après intégration par parties

$$\int_\Omega \nabla u \cdot \nabla v - \int_\Gamma v \frac{\partial u}{\partial n} = \int fv$$

d'où (les termes de bord s'annulent sur Γ du fait de la nullité de v)

$$\int_\Omega \nabla u \cdot \nabla v = \int fv.$$

Cette démarche d'élaboration de la formulation variationnelle n'est pas à proprement parler mathématique : ni l'espace dans lequel est censé vivre la solution, ni le sens que l'on peut donner à l'équation de départ, n'ont été précisés. C'est cette formulation variationnelle qui va permettre justement de donner un cadre théorique précis au modèle.

Cadre théorique. Ce problème se met donc sous la forme

$$a(u,v) = \langle \varphi \, , v \rangle \quad \forall v \in V,$$

où $a(\,\cdot\,,\,\cdot\,)$ est une forme bilinéaire symétrique sur un espace de Hilbert V, et φ une forme linéaire continue sur ce même espace. L'espace V est l'espace de Sobolev $H_0^1(\Omega)$ (voir section 22) des fonction de L^2 dont les dérivées partielles sont aussi dans L^2, et qui sont nulles [99] sur Γ :

Dans le cas où la forme bilinéaire $a(\,\cdot\,,\,\cdot\,)$ est coercive, c'est à dire (voir définition 20.20) s'il existe $\alpha > 0$ tel que $a(v,v) \geq \alpha \, |v|^2$ pour tout v dans V, le théorème de Lax Milgram (théorème 20.25) assure l'existence et l'unicité d'une solution dans V.

Cette solution peut être caractérisée comme unique minimiseur de la fonctionnelle

$$J(v) = \frac{1}{2} a(v,v) - \langle \varphi \, , v \rangle = \frac{1}{2} \int_\Omega |\nabla v|^2 - \int_\Omega fv.$$

98. Cette démarche en elle-même n'est pas mathématique, elle consiste précisément à faire rentrer le problème dans un cadre mathématique. Pour le mathématicien, non seulement le problème (17.1) n'est pas encore bien posé (il n'est pas sous une forme qui permette l'utilisation directe d'un théorème), mais d'une certaine manière il n'est même pas posé (l'espace dans lequel est supposé vivre l'inconnue n'est pas précisé, ni le sens que peuvent avoir les conditions aux limites). Ces remarques peuvent laisser croire que l'obtention de la formulation variationnelle se fait hors de toute règle. Il faut cependant garder à l'esprit qu'un retour (parfaitement mathématisé celui-là) vers l'équation sera nécessaire pour garantir le lien entre le problème initial et la formulation variationnelle.

99. Le sens que l'on peut donner à l'expression $u_{|\Gamma} = 0$ est précisé dans la section 22.3, page 222.

Le point essentiel pour pouvoir utiliser le théorème de Lax-Milgram est la coercivité de la forme bilinéaire, dont nous verrons qu'elle peut être mise à mal pour des matériaux dégénérés (pour le problème de conduction de la chaleur considéré ici, la dégénérescence se produit lorsque la conductivité tend localement vers 0) . Ici, la coercivité de la forme bilinéaire est assurée d'une part par l'hypothèse $k \geq \eta > 0$, et d'autre part par le fait que l'on peut choisir la quantité $(\int |\nabla u|^2)^{1/2}$ comme norme sur l'espace V, grâce à l'un des corollaires de l'inégalité de Poincaré (voir proposition 22.43, page 22.43).

Retour à l'équation de départ. La formulation variationnelle ayant été construite de façon informelle, il est important de préciser en quel sens le problème mis sous forme variationnelle correspond bien au problème initial. Cette étape peut être très délicate dans certains cas (la difficulté dépendant de la régularité de la frontière du domaine, et des conditions aux limites considérées). Le premier pas consiste à établir à partir de la formulation variationnelle que la solution est en fait plus régulière [100] que la régularité naturelle H^1 (qui intervient dans le cadre de l'utilisation du théorème de Lax-Milgram). La solution u est dite solution faible de

$$-\Delta u = f,$$

avec $f \in L^2(\Omega)$. Dans le cas où k est supposé régulier (C^1), la solution appartient en effet à un espace de fonctions plus régulières, l'espace $H^2(\Omega)$ (voir définition 22.20, et la section 22.7 pour l'énoncé des théorèmes de régularité), de telle sorte que Δu est défini comme fonction de $L^2(\Omega)$, et que l'on peut écrire

$$-\Delta u = f \quad \text{p.p. sur } \Omega.$$

Précisons que l'appartenance à $H^2(\Omega)$ ainsi que l'écriture de l'équation ci-dessus utilisent uniquement la formulation variationnelle pour des fonctions tests à support compact dans Ω (qui sont en particulier nulles au bord).

Les conditions aux limites de Dirichlet sur le bord du domain sont contenues dans l'appartenance de u à l'espace V

Discrétisation en espace. L'approximation de la solution u du problème de départ est basée sur l'introduction d'espaces V_h de fonctions, de dimension finie. Dans le cadre de la méthode des éléments finis dits P^1 (pour polynôme de degré 1), on se donne une suite de triangulations T_h (voir définition 17.14, page 171, pour une définition précise de ce que nous entendons par triangulation), où h est un petit paramètre destiné à tendre vers 0, qui mesure la finesse de la triangulation. On définit alors V_h comme l'espace des fonctions continues, qui vérifient la condition aux limites, et dont la restriction à chaque triangle de T_h est affine :

$$V_h = \left\{ v_h \in V \,,\ v_{h|K} \text{ est affine sur tout } K \in T_h \right\}.$$

Le problème discret s'écrit

$$\left\{ \begin{array}{l} \text{Trouver } u_h \in V_h \text{ tel que} \\ \displaystyle\int_\Omega \nabla u_h \cdot \nabla v_h = \int_\Omega f v_h \quad \forall v_h \in V_h. \end{array} \right. \tag{17.2}$$

100. Précisons que ce résultat de régularité interviendra de façon essentielle dans l'analyse d'erreur de la méthode de discrétisation.

Formulation matricielle. On numérote $i = 1, 2, \ldots, N_h$ les nœuds de la triangulation qui correspondent à des degrés de liberté (c'est à dire les sommets de T_h qui n'appartiennent pas à Γ). La solution recherchée u_h peut s'écrire

$$u_h = \sum_{j=1}^{N_h} u^j w_j,$$

de telle sorte que (17.2) se ramène au système matriciel (on garde la notation u_h pour désigner le vecteur $(u^1, \ldots, u^{N_h})$

$$Au_h = b_h,$$

où A est une matrice carrée d'ordre N_h , et $b_h \in \mathbb{R}^{N_h}$:

$$A = (a_{ij}) = \left(\int_\Omega \nabla w_i \cdot \nabla w_j \right) , \; b_h = \left(\int_\Omega f w_i \right)_i .$$

On peut vérifier que, dans le cas d'un maillage cartésien régulier (cellules carrées coupée en 2 triangles), la matrice obtenue est, à constante multiplicative près, la matrice du Laplacien discret que l'on obtient par une discrétisation dans le cadre de la méthode des différences finies. La mise en œuvre de la présente méthode ne nécessite en revanche aucune hypothèse sur le maillage.

Implantation sur `Freefem++` . Le logiciel `Freefem++` permet de calculer u_h en quelques lignes. Précisons que l'assemblage de la matrice et la résolution des systèmes sont gérés par le logiciel sans que l'utilisateur ait à intervenir (si ce n'est pour préciser éventuellement le choix de telle ou telle méthode de résolution). D'autre part, les conditions de Dirichlet non homogènes (conditions $u = 1$ sur Γ_3) ne nécessitent pas l'introduction explicite d'un relèvement de cette condition au bord.

```
int np=50;
mesh Th=square(np,np);

fespace Vh(Th,P1);
Vh u,tu  ;
func k = 1+0.5*sin(y*4*pi) ;
func f = 1 ;
plot(Th,wait=1);

problem Poisson(u,tu)=
    int2d(Th)(k*(dx(u)*dx(tu)+dy(u)*dy(tu)))
   -int2d(Th)(f*v)
   +on(1,2,3,4,u=0);
Poisson  ;  plot(u, wait=1);
```

Estimation d'erreur. L'estimation d'erreur, détaillée dans la section 17.2, se base sur 2 ingrédients.

1) En premier lieu, il s'agit d'établir une inégalité d'*approximation* du type

$$\inf_{v_h \in V_h} |v_h - u| \leq \varepsilon(h, u),$$

où u est la solution exacte du problème initial, et $\varepsilon(h, u)$ tend vers 0 quand le paramètre de discrétisation h tend lui-même vers 0. Pour le cas des éléments finis d'ordre 1 que nous avons considérés ici, ε est du type $Ch\,\|u\|_{H^2}$, où H^2 désigne l'espace de Sobolev des fonctions de L^2 dont toutes les dérivées secondes sont de carré intégrable. Noter que la régularité de la solution donnée par le théorème d'existence et d'unicité est simplement H^1. Il sera donc nécessaire de montrer que la solution est plus régulière que cela.

2) Le fait que l'estimation d'approximation précédente puisse conduire à une estimation d'erreur sur la solution effectivement calculée (qui a priori n'est pas la meilleure approximation de u par un élément de V_h) se base sur le lemme de Céa (voir section 17.2), qui utilise encore une fois la coercivité de la forme bilinéaire $a(\,\cdot\,,\,\cdot\,)$, et s'exprime ici

$$\|u - u_h\| \leq C \inf_{v_h \in V_h} |v_h - u|\,,$$

où C est une nouvelle constante qui dépend des propriétés de la forme bilinéaire. Nous verrons que dans le cas de matériaux inhomogènes cette constante est susceptible d'être très grande, ce qui suggère une dégradation de la précision numérique. La démonstration de ces propriétés fait l'objet de la section 17.2.

Ces propriétés assurent ici que, si l'on considère (T_h) une famille régulière de triangulations de Ω (voir définition 17.17), V_h l'espace d'approximation associé défini précédemment, alors il existe une constante $C > 0$ telle que

$$|u - u_h|_{\Omega,1} \leq Ch\,|f|_{\Omega,0}\,.$$

C'est une application directe de la proposition 22.55, page 233 (ou plus précisément de la proposition 22.57 qui s'applique au cas d'un polyèdre convexe), du théorème d'approximation 17.18, et du lemme de Céa 17.3.

Remarque 17.1. *On prendra garde au fait que le lemme de Céa est* non local *(l'estimation de l'erreur par l'erreur d'approximation est globale). En particulier, si la solution a la régularité H^2 sauf au voisinage d'un point (par exemple un coin rentrant), on n'a pas forcément approximation d'ordre* 1*, même loin du point problématique : la singularité est susceptible de* polluer *l'ensemble de l'approximation.*

Autres conditions aux limites

Conditions de Neuman. On considère la situation (rencontrée dans les exemples du chapitre I) où la dérivée normale est imposée sur une partie de la frontière. Notons Γ_N cette partie, et Γ_D la composante restante, sur laquelle on choisit d'imposer une condition de Dirichlet homogène. Pour fixer les idées, on considère que Ω est le carré unité, et que Γ_N est le bord inférieur. On se donne une donnée $g \in L^2(\Gamma_N)$ sur le bord [101]. Le problème considéré

101. La question de la régularité de g est un peu délicate. On pourra considérer dans un premier temps $g \in L^2(\Gamma)$, ce qui permet d'obtenir un problème bien posé. En revanche si l'on souhaite démontrer la régularité H^2 de la solution, il est nécessaire de prendre une donnée plus régulière, en l'occurrence $H^{1/2}(\Gamma)$.

est maintenant (avec $k \equiv 1$)

$$\left\{ \begin{array}{rcll} -\Delta u & = & f & \text{in } \Omega \\ u & = & 0 & \text{sur } \Gamma_D \\ \dfrac{\partial u}{\partial n} & = & g & \text{sur } \Gamma_N \end{array} \right. \tag{17.3}$$

On obtient la formulation variationnelle en multipliant par une fonction-test v nulle Γ_D en intégrant par parties, et en remplaçant [102] $\partial u/\partial n$ par g :

$$\int_\Omega \nabla u \cdot \nabla v = \int_\Omega fv + \int_{\Gamma_N} gv.$$

Ce problème se ramène donc à la recherche de $u \in V$ tel que

$$a(u,v) = \langle \varphi \, , v \rangle \quad \forall v \in V,$$

avec [103]

$$V = \left\{ u \in H^1(\Omega) \, , \; u_{|\Gamma_D} = 0 \right\}.$$

L'espace V est sous-espace fermé de $H^1(\Omega)$, c'est donc bien un espace de Hilbert, et $a(\,\cdot\,,\,\cdot\,)$ est une forme bilinéaire continue symétrique. L'intégrale en volume dans le second membre est bien une forme linéaire continue, et

$$\left| \int_{\Gamma_N} gv \right| \; \leq |g|_{L^2(\Gamma_N)} \, |v|_{L^2(\Gamma_N)} \leq C \, |g|_{L^2(\Gamma_N)} \, \|v\|_{H^1(\Omega)} \, ,$$

par continuité de l'application trace, et donc $\varphi \in V'$. Il reste à établir la coercivité de a, ce que permet le corrolaire 22.47, page 230, de l'inégalité de Poincaré généralisée :

$$\int_\Omega |\, \nabla u|^2 \geq \frac{1}{1+C^2} \left(\int_\Omega \, u^2 + \int_\Omega |\, \nabla u|^2 \right).$$

Le problème admet donc une unique solution $u \in V$.

Remarque 17.2. *On peut choisir de munir V d'une autre norme. Ici, l'inégalité de Poincaré généralisée assure que la semi-norme $|u|_1$ est en fait une norme équivalent à la norme H^1 (avec la partie L^2). On peut donc choisir de munir V de cette norme, et par suite la forme est bien sûr coercive, avec une constante de coercivité égale à 1. Dans ce cas l'existence et l'unicité sont directement données par le théorème de Riez-Fréchet.*

Retour à l'équation de départ. Il s'agit de montrer en premier lieu que la solution est H^2, de façon à donner un sens à Δu comme fonction [104]. Cette régularité est assurée sous

102. Il est essentiel de faire disparaître toute trace de $\partial u/\partial n$, car cette quantité n'est pas définie pour des fonctions de H^1. Or la forme bilinéaire impose que l'on se place dans H^1 pour utiliser le théorème de Lax-Milgram.

103. En toute rigueur la condition de Dirichlet sur Γ_D devrait s'écrire en utilisant l'opérateur de trace γ_0. Nous utiserons pourtant dans la suite la notation $u_{|\Gamma_D}$ pour désigner la trace de u sur Γ_D.

104. Il existe une autre manière (que nous ne privilégierons pas ici) de donner un sens à l'équation de Poisson sans l'aide d'aucun théorème de régularité (voir section 22.9, page 235). La formulation variationnelle assure que ∇u admet une divergence faible L^2. On peut donc donner un sens à Δu comme la divergence faible de ∇u, en gardant à l'esprit qu'il s'agit d'une notation globale, et qu'en particulier les dérivées secondes ne sont pas nécessairement définies comme des fonctions de L^2. On peut pousser la démarche jusqu'à donner un sens à $\partial u/\partial n$ comme la trace normale du champ de vecteur $\nabla u \in H(div)$ (voir remarque 22.68), page 235). Cette trace est alors définie dans un sens faible, ce qui interdit par exemple l'écriture $\partial_n u = g$ p.p.

certaines hypothèses, en particulier ici dans le cas de conditions mixtes dans le cas où le raccord entre les différentes composantes se fait à angle droit (voir remarque 22.58, page 233). Nous supposerons ici que la donnée g a été choisie de telle sorte que cette régularité H^2 soit vérifiée.

Cet exemple va nous permettre de faire la distinction entre condition *essentielle* (conditions de Dirichlet), et condition *naturelle* (de Neuman en l'occurence, mais il pourrait s'agir des conditions de Robin). Dans le premier cas, la condition au bord est dans la définition de l'espace sur lequel on travaille : on a $u(x) = 0$ presque partout sur Γ_D par appartenance de u à V. Les conditions de Neuman ont en revanche disparu en tant que telles du problème sous sa forme variationnelle, il est important de vérifier qu'elles sont bien vérifiées dans un certain sens par la solution. On utilise pour cela la régularité H^2 de la solution. On considère alors la formulation variationnelle pour des fonctions-test régulières qui s'annulent sur Γ_D, mais pas forcément sur Γ_N. On utilise alors la formule de Green (voir proposition 22.37), ce qu'autorise la régularité H^2 de la solution u, pour obtenir

$$\int_\Omega (-\Delta u)\, v + \int_{\Gamma_N} \frac{\partial u}{\partial n} v - \int_{\Gamma_N} g v = \int_\Omega f v.$$

Comme l'équation de Poisson est vérifiée presque partout, il reste

$$\int_{\Gamma_N} \left(\frac{\partial u}{\partial n} - g \right) v = 0.$$

La fonction v pouvant être choisie arbitrairement, on en déduit $\partial_n u = g$ presque partout sur Γ_N.

Discrétisation en espace. La discrétisation en espace ne change pas significativement du cas Dirichlet homogène, si ce n'est que les points du maillage situés sur Γ_N correspondent maintenant à des degrés de liberté, et que le second membre contient des termes provenant d'intégrales surfaciques impliquant les fonctions-test associées à ces nouveaux degrés de liberté :

$$b_h = \left(\int_\Omega f v_h + \int_{\Gamma_N} g w_i \right)_i .$$

17.2 Estimation d'erreur pour la méthode des Éléments Finis

Principes abstraits

Soit V un espace de Hilbert, et $a(\,\cdot\,,\,\cdot\,)$ une forme bilinéaire symétrique coercive sur V, de constante de coercivité α et de constante de continuité $\|a\|$, et $f \in V'$. On note u l'élément de V qui minimise la fonctionnelle

$$v \in V \longmapsto J(v) = \frac{1}{2} a(v,v) - \langle \varphi\,,\, v \rangle.$$

Dans le cadre de la discrétisation en espace qui sera présentée dans les sections suivantes, on utilisera la notation V_h pour représenter un espace d'approximation de dimension finie, h étant un paramètre associé au maillage sur lequel cette discrétisation s'effectue. Dans la proposition abstraite qui suit, à la base de la méthode des éléments finis, V_h désigne simplement un sous-espace fermé de V.

Proposition 17.3. *(Lemme de Céa (cas symétrique))*
Soit $a(\cdot\,,\cdot)$ une forme bilinéaire symétrique coercive sur V, de constante de coercivité α et de constante de continuité $\|a\|$, et $\varphi \in V'$. On note u l'élément de V qui minimise la fonctionnelle

$$v \in V \longmapsto J(v) = \frac{1}{2}a(v,v) - \langle \varphi\,, v\rangle.$$

Soit V_h un sous-espace fermé de V. On note u_h l'élément de V_h qui minimise J sur V_h. alors

$$|u_h - u| \leq \sqrt{\frac{\|a\|}{\alpha}} \inf_{v_h \in V_h} |v_h - u|\,.$$

Démonstration. On écrit les formulations variationnelles associées aux problèmes de minimisation sur V et sur V_h, respectivement,

$$a(u,v) = \langle \varphi\,, v\rangle \quad \forall v \in H,$$

$$a(u_h,v_h) = \langle \varphi\,, v_h\rangle \quad \forall v_h \in V_h.$$

On a donc

$$a(u_h - u, v_h) = 0 \quad \forall v_h \in V_h,$$

ce qui exprime que u_h minimise la fonctionnelle $v \mapsto a(v_h - u, v_h - u)$ sur V_h. On a donc, en utilisant la coercivité et la continuité de $a(\cdot\,,\cdot)$,

$$\alpha\,|u_h - u|^2 \leq a(u_h - u, u_h - u) \leq \inf_{v_h \in V_h} a(v_h - u, v_h - u) \leq \|a\| \inf_{v_h \in V_h} |v_h - u|^2\,,$$

d'où l'inégalité annoncée. □

La propriété demeure (avec une constante dégradée) pour une forme non symétrique, comme l'exprime le lemme de Céa général :

Proposition 17.4. *(Lemme de Céa)*
Soit $a(\cdot\,,\cdot)$ une forme bilinéaire (non nécessairement symétrique) coercive sur V, de constante de coercivité α et de constante de continuité $\|a\|$, et $\varphi \in V'$. Soit V_h un sous-espace de V. On note u et u_h les élements de V et V_h, respectivement, qui vérifient

$$a(u,v) = \langle \varphi\,, v\rangle \quad \forall v \in V,$$

$$a(u_h,v_h) = \langle \varphi\,, v_h\rangle \quad \forall v_h \in V_h.$$

Alors

$$|u_h - u| \leq \frac{\|a\|}{\alpha} \inf_{v_h \in V_h} |v_h - u|\,.$$

Démonstration. On utilise comme précédemment

$$a(u_h - u, v_h) = 0 \quad \forall v_h \in V_h,$$

dont on déduit que $a(u_h - u, u_h - u) = a(u_h - u, v_h - u)$, pour tout $v_h \in V_h$, d'où

$$\alpha\,|u_h - u|^2 \leq a(u_h - u, u_h - u) \leq |a(u_h - u, v_h - u)| \leq \|a\|\,|u - u_h| \inf_{v_h \in V_h} |v_h - u|\,,$$

d'où l'on déduit l'inégalité en prenant l'infimum en v_h. □

Approximation sur un simplexe

Dans la suite K désigne un simplexe de $\mathbb{R}^N$ non dégénéré (*i.e.* de volume non nul). On désignera par $\hat{K}$ le simplexe de référence, défini par

$$\hat{K} = \left\{ (x_1, \ldots, x_N) \in \mathbb{R}^N_+ \,,\ x_1 + \cdots + x_N \leq 1 \right\}.$$

On se placera dans ce qui suit en dimension 2 d'espace, où $\hat{K}$ est le triangle de référence

$$\hat{K} = \left\{ (x_1, x_2) \in \mathbb{R}^2_+ \,,\ x_1 + x_2 \leq 1 \right\}.$$

Notation 17.5. Pour toute fonction w définie sur K (ou sur tout autre domaine), on notera (lorsque ces quantités sont définies)

$$|w|_{0,K} = \|w\|_{L^2(K)} \,,\ |w|_{1,K} = \|\nabla w\|_{L^2(K)^2} \,,\ |w|_{2,K} = \left\|D^2 w\right\|_{L^2(K)^{N^2}} = \left(\sum_{i,j} |\partial_{ij} u|^2 \right)^{1/2}.$$

Notation 17.6. On note $P^k(K)$ l'espace des fonctions polynômiales sur K, de degré total inférieur ou égal à k. Ainsi $P^1(K)$ désigne l'espace des fonctions affines sur K, de dimension $N+1$, et $P^0(K)$ la droite des fonctions constantes.

Le cœur théorique de la méthode des éléments finis repose sur une estimation de stabilité sur le simplexe de référence, qui sera étendue à un simplexe quelconque par simple changement de variable affine. On considère ici des polynôme d'ordre 1 (éléments finis dits P^1), on renvoie à la fin de la section pour le cas général.

Lemme 17.7. *Soit I_K un opérateur linéaire continu de $H^2(K)$ dans $H^1(K)$ On suppose que I_K laisse invariant tous les éléments de P^1. Alors il existe une constante C telle que*

$$|v - I_K v|_{1,K} \leq C\, |v|_{2,K} \quad \forall v \in H^2(K).$$

Démonstration. On raisonne par l'absurde, en supposant l'existence d'une suite (v_n) telle que

$$|v_n - I_K v_n|_{1,K} > nC\, |v_n|_{2,K}.$$

On choisit de prendre v_n dans l'orthogonal de P^1 (ce qui est possible, quitte à corriger par un polynôme de degré 1, ce qui ne change aucun des membres), et de norme 1 dans H^2. Cette suite est bornée dans H^2, on peut donc en extraire une sous-suite qui converge faiblement vers $u \in H^2$. Cette sous-suite (toujours notée v_n) converge fortement dans H^1 par injection compacte, et donc fortement en fait dans H^2 car, $|v_n|_{2,K}$ tendant vers 0, elle y est de Cauchy. Elle converge donc fortement vers u. Toutes les dérivées à l'ordre 2 de u sont nulles : il s'agit donc d'un polynôme de degré au plus 1. Comme elle est dans l'orthogonal de P^1, on a donc $u = 0$, ce qui absurde car u est de norme 1 dans H^2. □

Definition 17.8. *(Opérateur d'interpolation)*
On définit l'opérateur d'interpolation I_K comme l'application de $C(K)$ (ensemble des applications continues de K dans $\mathbb{R}$) dans $P^1(K)$ qui à $u \in C(K)$ associe la fonction $I_K u$ affine sur K qui prend la valeur $u(x)$ en chaque sommet x de K. On définit de même I_K^0 l'application de L^1 dans $P^0(K)$ qui à une fonction associe la fonction constante sur K, de même valeur moyenne.

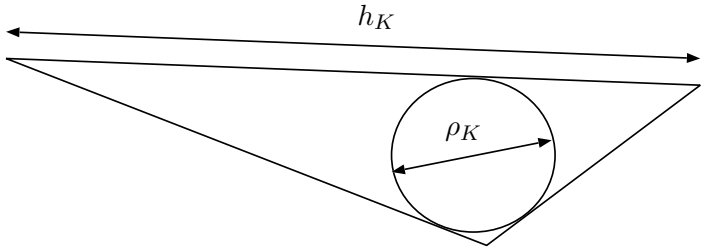


FIGURE 17.1 – Définition de h et ρ pour un triangle

Notation 17.9. On note h_K la longueur de la plus longue arête de K, et ρ_K le diamètre de la plus grande sphère contenue dans K (voir figure 17.1). On a ainsi $h_K/\rho_K \geq 1$. On notera $\tilde{h}$ et $\tilde{\rho}$ les quantités associées au simplexe de référence.

Lemme 17.10. *Soit Φ l'application affine qui envoie $\hat{K}$ dans K (noter que l'on peut choisir Φ linéaire si l'on suppose que 0 est un sommet de chacun des simplexes) :*

$$\hat{x} \longmapsto x = \Phi(\hat{x}) = B\hat{x} + b$$

On a

$$\|\nabla\Phi\| = \left\|{}^t\nabla\Phi\right\| = \|B\| \leq \frac{1}{\hat{\rho}} h_K\,,\ \left\|\nabla\Phi^{-1}\right\| = \left\|{}^t\nabla\Phi^{-1}\right\| = \left\|B^{-1}\right\| \leq \frac{1}{\rho_K}\hat{h}.$$

Démonstration. Soit $\tilde{\xi} \in \mathbb{R}^N$ de norme $\tilde{\rho}$. Il existe $\tilde{x}_1$ et $\tilde{x}_2$ dans $\tilde{K}$ tels que $\tilde{\xi} = \tilde{x}_2 - \tilde{x}_1$. On a donc

$$B\tilde{\xi} = B\tilde{x}_2 - B\tilde{x}_1 = \Phi\tilde{x}_2 - \Phi\tilde{x}_1 = x_2 - x_1,$$

qui est de norme inférieure à h_K par définition. On en déduit la première inégalité. La seconde se montre de la même manière en considérant $\xi = x_2 - x_1$ de norme ρ_K. □

Le cœur des estimations repose sur une formule de changement de variable entre $\hat{K}$ et K, ou plus précisément sur la manière dont le passage de $\hat{K}$ à K (ou l'inverse) est susceptible de modifier les valeurs des dérivées partielle d'une fonction poussée par Φ (ou Φ^{-1}). Pour alléger les notations, on notera simplement h pour h_K, et ρ pour ρ_K, en considérant que ces quantités pour le triangle de références sont des constantes.

Lemme 17.11. *Soit u une fonction régulière définie sur le triangle non dégénéré K (de diamètre h et de diamètre intérieur $\hat{\rho}$, et $\hat{u}$ définie sur $\hat{K}$ par*

$$\hat{u}(\hat{x}) = u \circ \Phi.$$

Soit $\alpha = (\alpha_1, \alpha_2)$ un multi-indice, avec $|\alpha| = \alpha_1 + \alpha_2 = s \in \mathbb{N}$. On a

$$\left|\frac{\partial^s \hat{u}}{\partial^\alpha \hat{x}}\right| \leq Ch^s \sum_{|\alpha|=s} \left|\frac{\partial^s u}{\partial^\alpha x}\right|\,,\ \left|\frac{\partial^s u}{\partial^\alpha x}\right| \leq C\frac{1}{\rho^s} \sum_{|\alpha|=s} \left|\frac{\partial^s \hat{u}}{\partial^\alpha \hat{x}}\right|.$$

Démonstration. Soit u une fonction régulière définie sur K. On a

$$\frac{\partial \hat{u}}{\partial \hat{x}_i} = \nabla u \cdot \frac{\partial \Phi}{\partial \hat{x}_i} = \left((\nabla\Phi)^T \nabla u\right) \cdot \hat{e}_i,$$

de telle sorte que $\nabla\hat{u}(\hat{x}) = (\nabla\Phi)^T \nabla u(x)$. On a donc

$$\left|\frac{\partial \hat{u}}{\partial \hat{x}_i}\right| \leq Ch \sum_{|\alpha|=s} \left|\frac{\partial^s u}{\partial x_i}\right|$$

L'estimation sur les dérivées d'ordre plus élevées, ainsi que les estimations inverses (à partir de $u(x) = \hat{u} \circ \Phi^{-1}$, se démontrent de la même manière. □

Théorème 17.12. *On suppose $N = 1$, 2, ou 3, de telle sorte que $H^2(K)$ s'injecte de façon continue dans $C^0(\overline{K})$. Il existe une constante C universelle telle que, pour tout triangle K du plan, non dégénéré, on a*

$$\begin{aligned}
|I_K u - u|_{1,K} &\leq C\frac{h^2}{\rho}|u|_{2,K} \quad \forall u \in H^2(K) \\
|I_K u - u|_{0,K} &\leq Ch^2|u|_{2,K} \quad \forall u \in H^2(K) \\
\left|I_K^0 u - u\right|_{0,K} &\leq Ch|u|_{1,K} \quad \forall u \in H^1(K)
\end{aligned}$$

Démonstration. Ces estimations se démontrent à partir de l'estimation de stabilité (proposition 17.7) appliquée au simplexe de référence. On transporte $|I_K u - u|^2_{1,K}$ sur le triangle de référence, ce qui fait apparaître $\left|\widehat{I_K u - u}\right|^2_{1,\hat{K}} = |I_{\hat{K}}\hat{u} - \hat{u}|^2_{1,\hat{K}}$ multiplié par le jacobien de Φ, ainsi que par le facteur $1/\rho^2$. On utilise alors l'estimation de stabilité sur $\hat{K}$, qui fait apparaître $|\hat{u}|^2_{2,\hat{K}}$. On fait subir à cet intégrale le sort inverse, en se ramenant sur K, ce qui fait apparaître l'inverse du Jacobien, et le facteur h^4 (à constante multiplicative indépendante de K près). La racine carrée de l'inagalité obtenue donne la première inégalité, les autres se démontrent de la même manière. □

Remarque 17.13. *La démonstration précédente met clairement en évidence la source des puissances de h et ρ dans l'estimation. Le* 1 *du dénominateur ρ vient du* 1 *de la semi norme du membre de gauche, et le* 2 *du numérateur vient de* 2 *de la semi-norme du membre de droite. Une telle estimation sera utilisable dans une optique d'estimation si la puissance du numérateur est strictement supérieur à celle du dénominateur (pour des triangles réguliers, h et ρ sont de même taille). On retrouve un principe extrêmement général en théorie de l'approximation : quand tout se passe bien (i.e.* au mieux*), l'ordre de l'erreur est la différence entre l'ordre de dérivation que l'on contrôle pour la fonction approchée, moins l'ordre de dérivation que l'on cherche à approcher. On retrouvera par exemple ce principe dans un cadre standard pour une fonction de C^m, dont on cherche à approcher la dérivée k-ième par une méthode de type différences finies avec un pas h (il est possible que la convergence soit plus lente que n'importe quelle puissance de h). Pour $k = m$ on a bien convergence ponctuelle, mais sans ordre. Dans le cas $m > k$ l'erreur commise (ici en norme sup) en général sera d'ordre $m - k$.*

Approximation sur un domaine

Definition 17.14. *(Triangulation)*
Soit Ω un domaine polygonal du plan. On appelle triangulation de Ω une famille T_h de

triangles non dégénérés deux à deux disjoints telle que

$$\overline{\Omega} = \bigcup_{K \in T_h} \overline{K},$$

et telle que, pour tous K, K' de T_h, l'intersection $\overline{K} \cup \overline{K}'$ est vide, ou réduite à un sommet commun des triangles, ou réduite à un côté commun des triangles. Les sommets des triangles de T_h sont appelés les nœuds de la triangulation.

Definition 17.15. *(Opérateur d'interpolation)*
Soit Ω un domaine polygonal du plan, et T_h une triangulation de Ω. On définit l'opérateur d'interpolation I_h comme l'application de $C(\overline{\Omega})$ (ensemble des applications continues de $\overline{\Omega}$ dans $\mathbb{R}$) qui à $u \in C(\overline{\Omega})$ associe la fonction u_h affine sur chaque $K \in T_h$ qui prend la valeur $u(x)$ en chaque sommet x de T_h.

Remarque 17.16. *Le paramètre h joue un rôle un peu ambigu dans ce contexte : il désigne à la fois l'indice d'un membre d'une famille de triangulations (c'est donc le* label *d'une triangulation), et ce qu'il est convenu d'appeler le diamètre de la triangulation , c'est à dire le* sup *de h_K pour $K \in T_h$, qui est un nombre réel. C'est évidemment un abus de notation, puisque deux triangulations peuvent avoir le même diamètre sans être identiques. Nous conservons néammoins cet usage, qui permet d'alléger les notations.*

Definition 17.17. *(Famille régulière de triangulations)*
Soit Ω un domaine polygonal. On appelle famille régulière de triangulations une famille (T_h) telle que

(i) il existe une constante σ telle que $\sup_h \sup_{K \in T_h} (h_K/\rho_K) \leq \sigma$,

(ii) le diamètre de T_h tend vers 0, c'est-à-dire que $\sup_{K \in T_h} h_K \longrightarrow 0$.

Théorème 17.18. *Soit Ω un domaine polygonal, et (T_h) une famille régulière de triangulations de Ω. Pour tout $u \in H^2(\Omega)$, on a*

$$|u - I_h u|_{1,\Omega} \leq C\sigma h\, |u|_{2,\Omega}\ ,\ |u - I_u|_{0,\Omega} \leq C h^2\, |u|_{2,\Omega}$$

Démonstration. On a

$$\int_\Omega |u - I_h u|^2 = \sum_{K \in T_h} \int_K |u - I_h u|^2 \leq C^2 h^4 \sum_{K \in T_h} |u|^2_{2,K} \leq C^2 h^2\, |u|^2_{2,\Omega}\,.$$

On raisonne de la même manière pour estimer $|u - I_h u|_{1,\Omega}$. □

Convergence de la méthode pour le problème de Poisson

Proposition 17.19. *Soit Ω un domaine polyédrique convexe, et $(T_h)_h$ une famille régulière de triangulations de Ω. On note V_h l'ensemble des fonctions de $H^1_0(\Omega)$ dont la restriction à chaque triangle de T_h est affine. Pour $f \in L^2(\Omega)$, on note $u \in H^1_0(\Omega)$ la solution faible de*

$$-\triangle u = f,$$

et u_h la solution du problème discrétisé

$$\int_\Omega \nabla u_h \cdot \nabla v_h = \int_\Omega f v_h \quad \forall v_h \in V_h.$$

Il existe une constante $C > 0$ telle que

$$|u - u_h|_{\Omega,1} \leq Ch\,|f|_{\Omega,0}\,.$$

17.3 Estimation de valeurs propres

On s'intéresse ici à l'approximation des valeurs propres d'une forme bilinéaire du type $\int \nabla u \cdot \nabla v$.

Théorème 17.20. *On se place dans le cadre du théorème* **??**, *page* **??**. *On introduit une suite d'espaces d'approximation (V_h) de V, et l'on note (u_h^i, λ_h^i) les solutions du problème aux valeurs propres sur V_h :*

$$a(u_h^i, v) = \lambda_h^i (u_h^i, v),$$

où $(\,\cdot\,,\,\cdot\,)$ est le produit scalaire sur H.

On a alors, pour tout i, convergence de λ_h^i vers λ^i quand h tend vers 0.

Démonstration. On note N_h la dimension de V_h. Notons tout d'abord que le principe du min-max

$$\lambda^i = \min_{W \in E^i} \max_{w \in W \setminus \{0\}} R(w)\,,\ \lambda_h^i = \min_{W \in E_h^i} \max_{w \in W \setminus \{0\}} R(w)$$

où E^i (respectivement E_h^i désigne l'ensemble des sous-espaces vectoriels de V (resp. V_h) de dimension i, implique $\lambda^i \leq \lambda_h^i$ pour tout $i \leq N_h$. s Notons Π_h la projection de V sur V_h pour le produit scalaire associé à $a(\,\cdot\,,\,\cdot\,)$, et W_i l'espace vectoriel engendré par les i premiers vecteurs propres de $a(\,\cdot\,,\,\cdot\,)$. Pour tout $u \in W_i$, on a

$$u = \sum_{k=1}^{i} \beta^k u_k,$$

et ainsi

$$\|\Pi_h u - u\|_V = \left| \sum_{k=1}^{i} \beta^k (\Pi_h u_k - u_k) \right| \leq \left(\sum_{k=1}^{i} \left|\beta^k\right|^2 \right)^{1/2} \left(\sum_{k=1}^{i} \|\Pi_h u_k - u_k\|_V^2 \right)^{1/2}$$

$$= |u| \left(\sum_{k=1}^{i} \|\Pi_h u_k - u_k\|_V^2 \right)^{1/2}.$$

On a donc

$$\lim_{h \to 0} \sup_{u \in W_i} \frac{|\Pi_h u - u|_V}{|u|} = 0$$

Par ailleurs, on a $a(\Pi_h u, \Pi_h u) \leq a(u, u)$, pour tout $u \in V$. Le principe du min-max permet pour finir d'écrire que

$$\lambda_h^i \leq \max_{w \in W_h \setminus \{0\}} R(w),$$

pour tout W_h de dimension i. Prenant $W_h = \Pi_h(W_i)$, il vient

$$\lambda_h^i \leq \max_{u \in W_i \setminus \{0\}} \frac{a(\Pi_h u, \Pi_h u)}{|\Pi_h u|^2} \leq \max_{u \in W_i \setminus \{0\}} \frac{a(u,u)}{|\Pi_h u|^2} \leq \lambda_i \max_{u \in W_i \setminus \{0\}} \frac{|u|^2}{|\Pi_h u|^2}.$$

Mais, d'après ce qui précède, on a

$$|\Pi_h u| = |u| + \mathcal{O}(|\Pi_h u - u|) = |u| + \mathcal{O}(\|\Pi_h u - u\|_V) = |u|\,(1 + o(h))$$

d'où l'on déduit, pour tout i, la convergence de λ_h^i vers λ^i quand h tend vers 0. □

17.4 Extension à des conditions aux limites plus générales

La méthode des éléments finis permet la prise en compte de conditions aux limites non standards de façon naturelle, sous réserve que le problème sous jacent possède une structure variationnelle.

Obstacle de conductivité infinie

On considère comme précédemment un domaine Ω du plan, et ω un sous-domaine fortement inclus dans Ω, c'est-à-dire que $\overline{\omega} \subset \Omega$. Le problème que nous allons considérer maintenant est issu du modèle physique suivant. On considère une plaque conductrice de la chaleur, dont on suppose que les bords sont à température nulle, et l'on suppose qu'une partie de cette plaque (qui correspondra au sous-domaine ω) a une conductivité infinie, de telle sorte que la température y est uniforme. On suppose qu'on chauffe la plaque sur la partie où la température est finie. On cherche ainsi un champ de température solution de l'équation de la chaleur, dans $\overline{\omega} \subset \Omega$, tel que la température est constante sur la frontière de ω, et tel que le flux de chaleur à travers cette frontière est nul.

On se donne donc f une fonction de $L^2(\Omega \setminus \overline{\omega})$, et l'on s'intéresse au problème suivant :

$$\left\{\begin{array}{rcll} -\triangle u & = & f & \text{dans } \Omega \setminus \overline{\omega} \\ u & = & 0 & \text{sur } \partial\Omega \\ u & = & U & \text{sur } \partial\omega \\ \displaystyle\int_{\partial\omega} \frac{\partial u}{\partial n} & = & 0, & \end{array}\right. \tag{17.4}$$

où U est une constante réelle dont la valeur est inconnue.

On introduit l'espace

$$H^1_C(\Omega \setminus \overline{\omega}) = \left\{ u \in H^1(\Omega \setminus \overline{\omega})\,,\ u = 0 \text{ sur } \partial\Omega\,,\ u = \text{cste sur } \partial\omega \right\}.$$

L'approche variationnelle directe est basée sur la fonctionnelle

$$\begin{array}{rcl} H^1_C(\Omega \setminus \overline{\omega}) & \longrightarrow & \mathbb{R} \\ v & \longmapsto & J(v) = \dfrac{1}{2}\displaystyle\int_{\Omega\setminus\overline{\omega}} |\nabla v|^2 - \int_{\Omega\setminus\overline{\omega}} fv, \end{array}$$

Le problème 17.5 consiste donc à minimiser J sur $H^1_C(\Omega \setminus \overline{\omega})$. On notera que la condition de flux nul a disparu. Il s'agit en fait d'une condition dite "naturelle", qui dérive du problème de minimisation, comme le précise la proposition suivante.

Proposition 17.21. *Soit $u \in H^1_C(\Omega \setminus \overline{\omega})$ la fonction qui minimise la fonctionnelle J sur $H^1_C(\Omega \setminus \overline{\omega})$. Alors u est solution du problème (17.4).*

Démonstration. On note U la valeur de u sur la frontière de ω, et l'on construit un relèvement $\tilde{U}$ de U, de régularité C^2, à support compact dans Ω. La fonction $u - \tilde{U}$ est dans $H^1_0(\Omega \setminus \overline{\omega})$, et c'est la solution faible de l'équation

$$-\triangle w = f + \triangle \tilde{U},$$

avec conditions de Dirichlet homogènes. C'est donc un élément de $H^2(\Omega \setminus \overline{\omega})$, et par suite u lui-même a une régularité H^2. On considère maintenant des fonctions-test dans $H^1_0(\Omega \setminus \overline{\omega})$. Par intégration par parties, on obtient $-\triangle u = f$ dans $\Omega \setminus \overline{\omega}$. Pour retrouver la condition de flux nul à travers l'interface, on prend maintenant une fonction test non nulle sur $\partial\omega$, qui prend par exemple la valeur 1. On utilise de nouveau la formule de Green pour obtenir

$$-\int_{\Omega\setminus\overline{\omega}} v\triangle u + \int_{\partial\omega} \frac{\partial u}{\partial n} v = \int fv,$$

d'où

$$\int_{\partial\omega} \frac{\partial u}{\partial n} = 0,$$

ce qui termine la preuve. □

17.5 Méthode des domaines fictifs

On considère un problème (de type Poisson pour fixer les idées) posé sur un domaine de géométrie complexe. L'approche consiste à plonger le problème dans un domaine plus grand, de géométrie plus simple (par exemple un parallélogramme). Cette stratégie permet d'évoter la génération de maillage adapté au domain initial, et de se limiter typiquement à des maillages cartésiens du domaine recouvrant.

Considérons par exemple un domaine du type $\Omega \setminus \overline{\omega}$ de $\mathbb{R}^2$, où Ω est le carré unité, et ω une collection de sous-domaines de Ω. On considérera pour simplifier le cas où ω est un disque fortement inclus dans Ω.

On se donne f dans $L^2(\Omega \setminus \overline{\omega})$, et l'on s'intéresse au problème suivant :

$$\left\{ \begin{array}{rcccl} -\triangle u & = & f & \text{dans} & \Omega \setminus \overline{\omega} \\ u & = & 0 & \text{sur} & \partial\Omega \cup \partial\omega. \end{array} \right. \tag{17.5}$$

Ce problème admet une formulation variationnelle qui rentre dans le cadre du théorème de Lax Milgram sur l'espace de Hilbert $H^1_0(\Omega \setminus \overline{\omega})$, et cette formulation conduit à une discrétisation en espace de type EF basée sur un maillage du domaine d'intérêt $\Omega \setminus \overline{\omega}$.

La présente approche consiste à se placer sur l'espace $V = H^1_0(\Omega)$, et à traiter comme une contrainte le fait d'être nul sur ω. On notera $K \subset V$ le sous-espace des fonctions qui s'annulent presque partout sur ω. On peut écrire le problème sous la forme d'un problème de minimisation, de la fonctionnelle

$$J \ : \ v \longmapsto \frac{1}{2} \int_\Omega |\nabla v|^2 - \int_\Omega fv,$$

sur K.

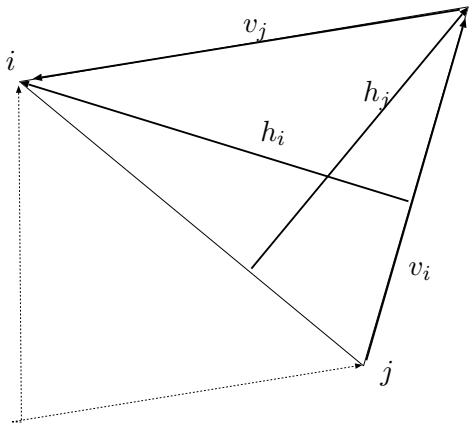


FIGURE 17.2 – Assemblage de la matrice élémentaire

Pénalisation

Cette approche consiste à relaxer la contrainte en considérant le problème de minimisation sur V tout entier, mais en introduisant un terme supplémentaire dans la fonctionnelle, qui *pénalise* le fait de ne pas vérifier la contrainte. On peut par exemple considérer

$$J_\varepsilon \ : \ v \longmapsto \frac{1}{2}\int_\Omega |\nabla v|^2 + \frac{1}{2\varepsilon}\int_\omega v^2 - \int_\Omega fv.$$

17.6 Éléments finis et réseaux résistifs

Soit T_h une triangulation d'un domaine Ω, et A la matrice résultant de la discrétisation par éléments finis P^1 de la forme blinéaire

$$a(u,v) = \int_\Omega \nabla u \cdot \nabla v.$$

Pour i et j voisins, l'intégrale de $\nabla w_i \cdot \nabla w_j$ résulte de deux contributions (les deux triangles qui contiennent i et j). L'une quelconque de ces contributions (voir figure 17.2) s'écrit

$$\int_K \nabla w_i \cdot \nabla w_j = \text{aire}(K)\, h_i \cdot h_j \frac{1}{|h_i|^2\,|h_j|^2}.$$

On note $D = v_i \wedge v_j$. L'aire du triangle vaut $D/2$. Par ailleurs, la hauteur $|h_i|$ du triangle peut s'exprimer

$$|h_i| = v_j \cdot \frac{v_i^\perp}{|v_i|} = \frac{v_i \wedge v_j}{|v_i|}.$$

On a donc

$$\int_K \nabla w_i \cdot \nabla w_j = \text{aire}(K)\, h_i \cdot h_j \frac{1}{|h_i|^2\,|h_j|^2} = \frac{D}{2}\frac{v_i \cdot v_j}{|v_i|\,|v_j|}|h_i|\,|h_j|\frac{1}{|h_i|^2\,|h_j|^2} = \frac{v_i \cdot v_j}{2D}.$$

L'intégrale sur l'ensemble du domaine est ainsi la somme de deux contributions de ce type, correspondant aux deux triangles partageant à la fois i et j. On note c_{ij} l'opposé de cette valeur. En écrivant que la fonction constante égale à 1 est somme des fonctions de base sur l'ensemble du maillage, on obtient

$$0 = \int_\Omega \nabla w_i \cdot \nabla 1 = \int_\Omega |\nabla w_i|^2 - \sum_{j\sim i} c_{ij}.$$

La matrice du Laplacien discrétisé est donc la matrice dont les termes extra-diagonaux sont les $-c_{ij}$, et les éléments diagonaux les $C_i = \sum c_{ij}$. On se trouve donc en présence d'une matrice associée à un réseau résistif (voir section 4), dont les sommets sont les sommets du maillages, les arêtes les côté de ce même maillage, et les résistances sont les inverses des quantités c_{ij} définie ci-dessus. Une solution du problème discret sans second membre peut donc s'interpréter comme un champ de pression sur le réseaux, harmonique sur les points intérieurs.

On prendra cependant garde au fait que les c_{ij} ne sont pas nécessairement positifs. Il ne le sont de façon sûre que si tous les angles de tous les triangles sont *aigus*. Dans le cas contraire, l'analogie doit être considérée avec précaution, certaines résistances du réseau associé pouvant être négatives. L'une des conséquence de cette négativité de certaines résistances est que la méthode ne vérifie plus forcément le principe du maximum discret. En effet, on a pour tout champ harmonique

$$p(i) = \frac{1}{C(i)} \sum_{j\sim j} c_{ij} p(j),$$

mais cette combinaison peut n'être plus barycentrique dans le cas où certains angles sont obtus.

On notera en rechanche que cette invalidation du principe du maximum ne remet pas en cause les propriétés de convergence de la méthode (section 17.2).

Equation de conservation continue associée à la solution discrète

On peut associer à la solution discrète d'un problème de laplace discrétisé par éléments fini une mesure vectorielle vérifiant une équation de conservation stationnaire (au sens des distribution).

Nous considérons pour fixer les idées le cas de conditions aux limites de Dirichlet non homogènes. L e problème consiste à trouver dans l'espace V_h des fonctions continues affines par morceaux une fonction qui prend des valeurs prescrites sur le bord, et qui vérifie la formulation variationnelle discrète (on note p l'inconnue pour expliciter le lien avec la section 4)

$$\int_\Omega \nabla p \cdot \nabla q = 0 \quad \forall v \in V_h^0,$$

où V_h^0 est l'espace des fonctions discrètes qui s'annulent au bord. Pour tout point x de la triangulation situé sur le bord du domaine, on note $\mu(x)$ la mesure atomique associée au flux discret lui même associé au champ de pression défini sur le réseau résistif $\mathcal{N} = (V, E, r, \Gamma)$ correspondant au maillage éléments finis, selon les principes décrit ci-dessus. Plus précisément, on note, pour tout $x \in \Gamma$, on note

$$\mu(x) = \sum_{x\in\Gamma} du(x)\, \delta_x\,,\ du(x) = \sum_{y\sim x} u(y,x) = \sum_{y\sim x} c(x,y)(p(y) - p(x)).$$

On note G la mesure vectorielle associée aux flux discrets sur le maillage, selon la démarche décrite dans la section 4.5. On a alors, au sens des discributions, (voir proposition 4.13, page 53)

$$\nabla \cdot G = \mu.$$

Noter que cette propriété de conservation formelle ne nécessite pas d'hypothèse sur la positivité des résistances. On gardera cependant à l'esprit que, dans le cas où le maillage présente des angles obtus, le réseau résistif associé ne correspond pas forcément à la situation *physique* de résistances positives [105].

105. Un tel réseau serait irréalisable en pratique, qu'il s'agisse d'un circuit électrique, on d'un réseaux de tuyaux au travers duquel s'écoule un fluide visqueux.

18 Résolution des systèmes linéaires

18.1 Conditionnement

La notion de conditionnement d'une matrice (on parle aussi de conditionnement d'un système linéaire) joue un rôle très important dans l'étude de la résolution de systèmes linéaires. Nous verrons plus loin que ce conditionnement intervient notamment de façon essentielle dans le vitesse de convergence de méthodes de résolution itératives.

Le conditionnement d'une matrice apparaît de façon naturelle lorsque l'on cherche à estimer la stabilité de la résolution d'un système linéaire par rapport aux données, indépendamment de la méthode numérique utilisée effectivement pour résoudre le système. Considérons une matrice $A \in \mathcal{M}_n(\mathbb{R})$ inversible, un second membre $b \in \mathbb{R}^n$, et le système linéaire

$$Au = b.$$

Le conditionnement quantifie la confiance que l'on peut avoir dans la solution (exacte) de ce système en fonction de la confiance que l'on a dans les données (en l'occurrence le second membre b), qui sont susceptibles d'être entachées d'erreurs de mesure, d'erreurs liées au stockage sur ordinateur avec une précision finie. Dans ce qui suit nous considérons la norme matricielle $\|A\|_2$, notée simplement $\|A\|$, subordonnée à la norme euclidienne sur $\mathbb{R}^n$. On considère ainsi une perturbation δb du second membre, et l'on cherche à estimer la variation δu induite sur la solution :

$$A(u + \delta u) = b + \delta b.$$

On a donc $\delta u = A^{-1}\delta b$, d'où $|\delta u| \leq \left\|A^{-1}\right\| |\delta b|$. D'autre part $b = Au$ implique $|b| \leq \|A\| \, |\delta b|$, d'où finalement

$$\frac{|\delta u|}{|u|} \leq \left\|A^{-1}\right\| \|A\| \frac{|\delta b|}{|b|}.$$

Definition 18.1. *(Conditionnement)*
Soit A une matrice inversible. On appelle nombre de conditionnement de A le réel

$$\kappa = \left\|A^{-1}\right\| \|A\|.$$

La quantité κ mesure donc le rapport entre l'erreur relative maximale sur la solution et l'erreur relative sur les données. Cette quantité sans dimension est toujours supérieure ou égale à 1 ($1 = \|\mathrm{Id}\| = \left\|AA^{-1}\right\| \leq \kappa$). Pour $\kappa \gg 1$, le problème est très instable par rapport aux données.

Remarque 18.2. *On peut aussi se demander quel est l'effet sur la solution d'une perturbation de la matrice elle-même :*

$$(A + \delta A)(u + \delta u) = b.$$

On obtient au premier ordre (on néglige le terme en $\delta A \delta u$) une formule analogue à la précedente, qui fait intervenir le κ comme un majorant du facteur d'amplification de l'erreur relative :

$$\frac{|\delta u|}{|u|} \leq \left\|A^{-1}\right\| \|A\| \frac{\|\delta A\|}{\|A\|}.$$

Conditionnement des matrices s.d.p. Dans le cas où A est symétrique définie positive, de valeurs propres

$$0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n,$$

le conditionnement s'écrit $\kappa = \lambda_n/\lambda_1$.

Exemple 18.1. Considérons la matrice du Laplacien discret donnée dans la section A.4, dont les valeurs propres sont connues. Le conditionnement de cette matrice est donc

$$\kappa = \lambda_{N-1}/\lambda_1 = \frac{\sin^2\left(\frac{(N-1)\pi}{2N}\right)}{\sin^2\left(\frac{\pi}{2N}\right)} \sim 4N^2 \quad \text{quand } N \to +\infty.$$

Definition 18.3. *Soit $A = (a_{ij})$ une matrice. On dit que A est une matrice-bande s'il existe ℓ tel que $a_{ij} = 0$ dès que $|j - i| > \ell$. Bien sûr cette notion n'a d'intérêt que si ℓ est significativement plus petit que n.*

18.2 Méthodes directes

On s'intéresse dans cette section à la résolution d'un système linéaire $Au = b$ bien posé (matrice A inversible).

Décomposition LU. La décomposition LU est basée sur la méthode du pivot de Gauss. Elle consiste à effectuer une factorisation dite LU de la matrice (L pour *low*, U pour *low* :

$$A = LU$$

, où L (resp. U) est une matrice triangulaire inférieure (resp. supérieure), et L ne contient que des 1 sur la diagonale. Une fois que cette décomposition est réalisée, la solution s'obtient par résolution de 2 systèmes triangulaires.

Il peut être intéressant de choisir le pivot à chaque étape (pour éviter par exemple d'inverser des nombres trop petits). Il s'agit alors de la décomposition avec permutation :

$$A = PLU,$$

où P est une matrice de permutation (les éléments sont des 0 ou des 1, et chaque ligne et chaque colonne contient exactement un 1.

Méthode de Cholesky. La méthode de Cholesky est une forme particulière de décomposition LU tirant partie du caractère symétrique d'une matrice. Cette méthode consiste à décomposer une matrice symétrique définie positive en un produit de 2 matrices triangulaires transposées l'une de l'autre.

Algorithme 18.4. *(Cholesky)*
Soit $A = (a_{ij})$ une matrice symétrique définie positive de $\mathcal{M}_n(\mathbb{R})$. Alors la matrice triangulaire inférieure $L = (b_{ij})_{j \leq i}$ définie par

$$b_{11} = \sqrt{a_{11}}\,,\ b_{21} = a_{21}/b_{11}\,,\ \ldots\,,\ b_{n1} = a_{n1}/b_{11},$$

et, pour $j = 2, \dots, n$,

$$b_{jj} = \sqrt{a_{jj} - \sum_{k=1}^{j-1} b_{jk}^2}\,,\ b_{ij} = \frac{a_{ij} - \sum_{k=1}^{j-1} b_{jk} b_{ik}}{b_{jj}}\,,\ i = j+1, \dots, n,$$

est telle que $A = L\,{}^tL$.

Le système $Au = b$ *est alors résolu par la résolution successive des deux systèmes triangulaires*

$$Lw = b\,,\ {}^tLu = w.$$

Proposition 18.5. *La décomposition d'une matrice* A *s.d.p. de taille* $n \times n$ *par la méthode de Cholesky nécessite* n *extractions de racines, et un équivalent de* $n^3/6$ *divisions ou multiplications.*

La résolution du système linéaire $Au = b$ *par cette méthode nécessite en outre, pour la résolution des deux systèmes triangulaires, l'équivalent de* n^2 *opérations élémentaires (multiplications ou divisions).*

Démonstration: Le nombre d'extraction de racines est bien égal à n. Pour le nombre de multiplications/divisions, on cherche directement un équivalent. La première étape n'est donc pas prise en compte. Le gros du coût est dans le calcul de chacun des éléments extradiagonaux b_{ij}, au nombre de $n-j$ pour j fixé, qui nécessite (on ne garde que l'essentiel) j mutiplications. La complexité est donc en

$$\sum_j (n-j)j,$$

qui est un $\mathcal{O}(n^3)$, avec le coefficient $1/6$ (penser à $\int x(1-x) = 1/6$).

La résolution d'un système triangulaire consiste à effectuer, pour tout $j = 1, \dots, n$, j multiplications et une division. On a donc une complexité en $n^2/2$ pour chacun des systèmes triangulaires. □

Remarque 18.6. *La complexité réelle est en général très inférieure (tout du moins si l'écriture du programme informatique est adaptée à la situation), notamment dans le cas des matrices-bande (voir définition 18.3 ci-dessus), ce qui est souvent le cas des matrices résultants de la discrétisation par éléments finis d'un opérateur elliptique. Dans ce cas, on peut montrer que la matrice* L *associée possède la même structure de matrice bande. En conséquence, pour* j *allant de* 2 *à* n, *le nombre d'éléments extradiagonaux* b_{ij} *chute de* $n-j$ *à* ℓ, *tout comme le nombre d'opérations nécessaire. La complexité descend donc à* $n\ell^2$. *Noter que la résolution des 2 systèmes triangulaires, dont la complexité chute à* $n\ell$, *reste d'un coût négligeable par rapport à la factorisation (au moins dans de la cas d'un seul système, voir à ce sujet la remarque 18.7). Dans le cas du Laplacien discret en dimension* 1, *la largeur de bande est* 2, *d'où une complexité de l'ordre de* n, *le nombre de points (nous ne précisons pas la constante, car la petite largeur de bande rend significatives des opérations dont nous avions négligé le nombre). En dimension* 2, *pour un problème scalaire sur un maillage* $\sqrt{n} \times \sqrt{n}$, *la matrice est de taille* n, *et de largeur de bande* $\sqrt{n}$, *d'où une compléxité en* $n^2/6$.

Remarque 18.7. *Cette méthode peut être particulièrement performante lorsque l'on souhaite résoudre un grand nombre de fois un système*[106] *impliquant une matrice donnée* A *(pour des*

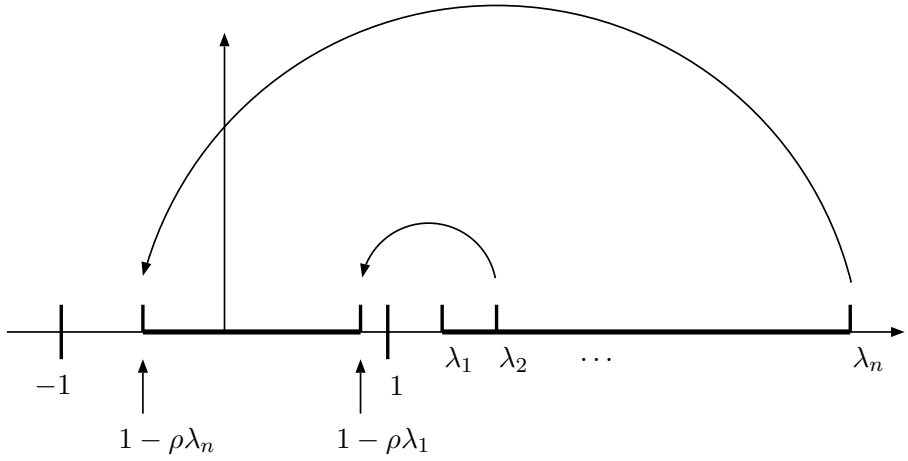


FIGURE 18.1 – Spectre de $\mathrm{Id}-\rho A$

seconds membres distincts). Notons M ce nombre de systèmes à résoudre. la complexité totale est de $n^3/6+Mn^2$, de telle sorte que dans la situation extrême où n devient négligeable devant M, on a une complexité asymptotique de la méthode en n^2 (coût unitaire d'une résolution de système).

18.3 Méthodes itératives

Algorithme 18.8. *Soit A une matrice symétrique définie positive de $\mathcal{M}_n(\mathbb{R})$. L'algorithme du gradient à pas fixe est basé sur la construction suivante : on se donne $\rho > 0$, un vecteur initial $u^0 \in \mathbb{R}^n$, et l'on construit*

$$u^{k+1} = u^k - \rho(Au^k - b).$$

Proposition 18.9. *L'algorithme du gradient à pas fixe converge dès que $\rho \in]0, 2/\lambda_n[$, où λ_n est la plus grande valeur propre de A*

Démonstration: On note $e^k = u^k - u$ l'erreur, qui vérifie $e^{k+1} = (\mathrm{Id}-\rho A)e^k$. Cette erreur converge dès que les valeurs propres de $\mathrm{Id}-\rho A$ sont de module strictement inférieur à 1. L'opération $A \mapsto \mathrm{Id}-\rho A$ renverse le spectre de A comme illustré sur la figure 18.1. Les valeurs propres de la nouvelle matrice sont donc de module strictement inférieur à 1 si et seulement si $1 - \rho\lambda_n > -1$, c'est à dire $0 < \rho < 2/\lambda_n$. □

Remarque 18.10. *Bien que la notion de choix optimal pour ρ soit sujette à caution, on notera que le choix*

$$\rho = 2/(\lambda_1 + \lambda_n)$$

minimise le rayon spectral de $\mathrm{Id}-\rho A$. Pour ce choix, le rapport géométrique de convergence est $1 - 2\lambda_1/(\lambda_1 + \lambda_n)$, donc de l'ordre de $1 - 2\kappa^{-1}$ pour κ grand. La convergence sera donc d'autant plus lente que le conditionnement κ est grand.

106. Cette situation se rencontre par exemple dans le cadre de la discrétisation en temps d'un problème d'évolution par une méthode implicite, qui se ramène à chaque pas de temps à la résolution d'un système pour une même matrice mais des seconds membres différents.

Méthode du gradient à pas optimal

La méthode du gradient à pas optimal est basée sur un calcul explicite du pas ρ de l'algorithme de gradient ci-dessus, de façon à minimiser la valeur de la fonctionnelle J sur la droite $\left\{u^k - \rho(Au^k - b)\,,\ \rho \in \mathbb{R}\right\}$. Un simple calcul permet d'exprimer ce ρ optimal à chaque itération :

Algorithme 18.11. *Soit A une matrice symétrique définie positive de $\mathcal{M}_n(\mathbb{R})$. L'algorithme du gradient à pas optimal est basé sur la construction suivante : on se donne un vecteur initial $u^0 \in \mathbb{R}^n$, et l'on construit*

$$u^{k+1} = u^k - \rho_k(Au^k - b)\,,\ \rho_k = \frac{\left|b - Au^k\right|^2}{|b - Au^k|_A^2}\,, \quad \textit{avec}\ |v|_A^2 = (Av, v)\,.$$

Remarque 18.12. *Noter que ρ_k est minoré et majoré, pour toute matrice s.d.p. A donnée.*

Méthode du gradient conjugué

La méthode du gradient conjugué permet d'approcher numériquement la solution de problèmes du type $Ax = b$, où A est une matrice symétrique définie positive. Nous verrons qu'en fait il s'agit d'une méthode exacte (qui converge en un nombre d'itérations fini égal à la dimension de l'espace), mais elle est dans la pratique utilisée comme un algorithme itératif.

Algorithme 18.13. *Soit A une matrice symétrique définie positive de $\mathcal{M}_n(\mathbb{R})$. L'algorithme du gradient conjugué est basé sur la constrution itérative suivante, à partir d'un vecteur initial $u_0 \in \mathbb{R}^n$. On définit tout d'abord le résidu initial correspondant $r_0 = b - Au_0$, et l'on pose $p_0 = r_0$,*

$$\begin{aligned}
\alpha_k &= \frac{|r_k|^2}{(Ap_k, p_k)} \\
u_{k+1} &= u_k + \alpha_k p_k \\
r_{k+1} &= r_k - \alpha_k A p_k \\
\beta_{k+1} &= |r_{k+1}|^2 / |r_k|^2 \\
p_{k+1} &= r_{k+1} + \beta_{k+1} p_k.
\end{aligned}$$

Proposition 18.14. *Les suites (r_k), (p_k) construites selon l'algorithme du gradient conjugué 18.13 vérifient les propriétés suivantes :*

$$(r_k, p_i) = (r_k, r_i) = 0 \quad \forall i \leq k-1\,,\ (p_k, Ap_i) = 0 \quad \forall i \leq k-1\,,\ |r_{k+1}|_{A^{-1}} \leq |r_k|_{A^{-1}}\,,$$

$$|r_k|_{A^{-1}} = \min_{F_k} |b - Au|_{A^{-1}}\,,\ F_k = u_0 + \mathrm{vect}(p_0, \ldots, p_{k-1}).$$

Démonstration: On démontre ces propriétés par récurrence. On a

$$(r_{k+1}, r_k) = |r_k|^2 - \alpha_k(r_k, Ap_k) = |r_k|^2 - \alpha_k(p_k - \beta_k p_{k-1}, Ap_k) = |r_k|^2 - \alpha_k(p_k -, Ap_k) = 0.$$

Pour tout $i \leq k-1$, on a

$$(r_{k+1}, r_i) = (r_k - \alpha_k Ap_k, r_i) = -\alpha_k(Ap_k, r_i).$$

Comme $r_i = p_i - \beta_i p_{i-1}$, le produit scalaire est nul du fait que les directions p_j sont deux à deux conjuguées pour $j \leq k$ (hypothèse de récurrence).

On a de même $(r_{k+1}, p_i) = 0$ pour tout $i \leq k$, car p_i s'exprime en fonctions des r_j, pour $j \leq i$.

Pour la conjugaison des directions de descente, on a

$$(p_{k+1}, Ap_k) = (r_{k+1} + \beta_{k+1} p_k, (r_k - r_{k+1})/\alpha_k),$$

ce qui donne (on utilise $(r_{k+1}, r_k) = (p_k, r_{k+1}) = 0$)

$$(p_{k+1}, Ap_k) = -\frac{1}{\alpha_k}\left(|r_{k+1}|^2 + \beta_{k+1}(p_k, r_k)\right) = 0$$

car $(p_k, r_k) = |r_k|^2$, et $\beta_{k+1} = |r_{k+1}|^2 / |r_k|^2$.

□

Proposition 18.15. *Soit A une matrice symétrique définie positive, et (u_k) une suite d'itérés produite par l'algorithme du gradient conjugué 18.13. On note $|\cdot|_A$ la norme associée à la matrice A, et $\kappa = \lambda_n/\lambda_1$ le conditionnement de A. On a*

$$|u_k - u|_A \leq 4\,|u_0 - u|_A \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^k.$$

Corollaire 18.16. *La norme de l'erreur vérifie*

$$|u_k - u| \leq 4\kappa\,|u_0 - u| \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^k.$$

Remarque 18.17. *Pour de grands nombres de conditionnement, on a une convergence géométrique de rapport voisin de $1 - 2/\sqrt{\kappa}$. On remarquera que ce taux est bien meilleur que celui trouvé pour la méthode de gradient à pas fixe (égal à $1 - 2/\kappa$, voir remarque 18.10).*

Remarque 18.18. *La convergence étant géométrique de rapport $1-2/\sqrt{\kappa}$, le nombre d'itérations à réaliser pour être sûr d'avoir une précision donnée ε est de l'ordre de $k_\varepsilon = \sqrt{\kappa}\ln(1/\varepsilon)$, contre $\kappa\ln(1/\varepsilon)$ pour le gradient à pas fixe. Le gain potentiel en termes de temps de calcul est donc considérable. Pour la résolution du Laplacien en dimension 1, avec $N = 100$ points, le conditionnement est de l'ordre de 10^4 (voir exemple 18.1, page 180), et le calcul par gradient conjugué va 100 fois plus vite que le calcul par gradient simple.*

Le comportement effectif du gradient conjugué dépend très sensiblement de la matrice bien sûr, mais aussi du second membre considéré. La figure 18.2 représente le logarithme de l'erreur au cours des itérations, pour la matrice du Laplacien discret d'ordre 100, pour un second membre obtenu comme N réalisations indépendantes d'une variable aléatoire de loi uniforme sur $[0, 1]$ (figure de gauche), puis pour un second membre dont tous les éléments sont égaux à 1 (figure de droite). Dans le premier cas, sur la première moitié du parcours, la convergence est géométrique de rapport $1 - 0.014$. Le conditionnement de la matrice est de l'ordre de 10^4, ce qui donne un ordre théorique de $1 - 0.02$, proche de l'ordre effectif. Noter qu'en revanche après l'itération 50 la convergence est beaucoup plus rapide. Ce phénomène est encore plus

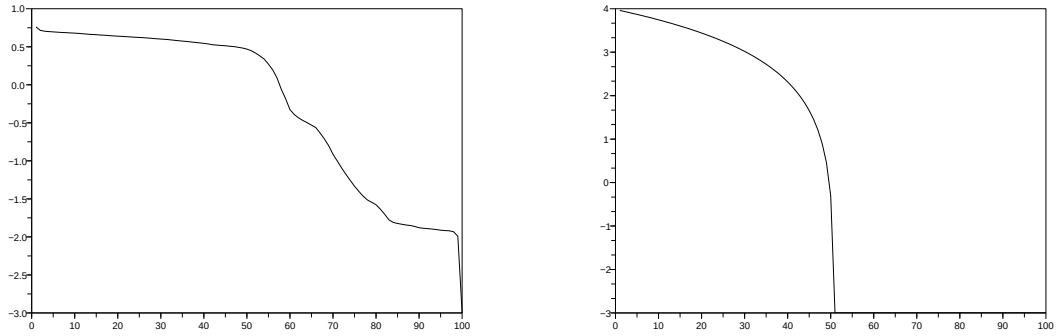


FIGURE 18.2 – Log du résidu au cours des itérations

net pour un second membre "non quelconque", puisqu'on obtient la précision machine après 50 itérations. Par ailleurs, si la pente pour les premières itérations correspond à peu près à la pente théorique, la convergence ne cesse d'accélérer. Ce phénomène reflète l'importance de la régularité du second membre, ou plus précisément le poids respectif des modes propres du Laplacien discret dans le second membre. Un cas simple permet d'appréhender ce phénomène, qui est très peu abordé dans la litérature : considérons un second membre combinaison d'un (petit) nombre p des premiers modes du Laplacien discret. L'algorithme évoluant dans l'espace de Kylov engendré par le second membre, tout se passe exactement comme si l'on résolvait un système de taille p, et la vitesse de convergence peut-être estimée à l'aide de conditionnement de la sous matrice correspondante, qui peut être très inférieur au conditionnement de la matrice globale. On aura par ailleurs une convergence à précision machine en un nobre d'étape au plus égal à la dimension spectrale du second membre. Dans le cas plus général où le second membre est l'interpolée d'une fonction régulière, cette régularité s'exprime dans le faible poids des hautes fréquences dans le représentation modale de la fonction, et la surconvergence peut s'expliquer par la quasi absence de modes de hautes fréquences.

18.4 Méthodes rapides

Le terme de méthode rapide fait référence à des algorithmes particuliers permettant de limiter le nombre d'opérations élémentaires pour réaliser (sans approximation) un calcul donné.

L'exemple le plus simple est le calcul d'une puissance entière d'un nombre réel (ou entier). Calculer x à la puissance 8 requiert a priori 7 multiplications. Mais on peut aussi calculer x^2, multiplier le résultat par lui-même, et encore une fois le résultat par lui-même, pour calculer le même nombre en 3 multiplications.

Dans le même esprit, le calcul de la valeur d'un polynôme

$$a_0 + a_1 X + \cdots + a_n X^n$$

en un point x peut s'écrire

$$(\ldots((a_n x + a_{n-1})x + a_{n-2}) + \cdots + a_1)x + a_0,$$

ce qui permet de limiter le nombre de multiplications à n (algorithme de Horner).

Tranformée de Fourier rapide (dimension 1). Pour ce qui concerne la résolution de problèmes du type de ceux rencontrés, nous nous contentons de donner ici le principe[107] d'une méthode permettant de résoudre rapidement (dans un sens que nous préciserons) des systèmes linéaires du type de ceux résultants de la discrétisation du Laplacien sur un maillage cartésien. Il s'agit de la méthode de transformée de Fourier rapide (*Fast Fourier Transform*). En dimension 1, la discrétisation en espace du problème de Poisson avec condition de Dirichlet homogène

$$-u'' = f\,,\ u(0) = u(1) = 0,$$

conduit à un système linéaire du type

$$Au = b,$$

où A est à une constante multiplicative près ($1/h = N$ en l'occurrence) la matrice du Laplacien discret (voir (A.13), page 257). Cette matrice est symétrique, donc diagonalisable dans une base orthogonale de vecteurs propres. On peut expliciter les éléments propres de cette matrice (voir section A.4), ce qui permet d'écrire

$$A = PDP^t\,,\ D = \operatorname{diag}\left(4\sin^2\left(\frac{k\pi}{2N}\right)\right)_{k=1,\ldots,N-1},$$

et

$$P = \sqrt{\frac{2}{N}}\begin{pmatrix} \sin\left(\frac{\pi}{N}\right) & \sin\left(\frac{2\pi}{N}\right) & \sin\left(\frac{3\pi}{N}\right) & \cdot & \cdot & \sin\left(\frac{(N-1)\pi}{N}\right) \\ \sin\left(\frac{2\pi}{N}\right) & \sin\left(\frac{4\pi}{N}\right) & \sin\left(\frac{6\pi}{N}\right) & \cdot & \cdot & \cdot \\ \sin\left(\frac{3\pi}{N}\right) & \sin\left(\frac{6\pi}{N}\right) & \cdot & \cdot & & \cdot \\ \cdot & & \cdot & \cdot & \cdot & \cdot \\ \cdot & & & \cdot & & \cdot \\ \sin\left(\frac{(N-1)\pi}{N}\right) & \cdot & \cdot & \cdot & \sin\left(\frac{(N-2)(N-1)\pi}{N}\right) & \sin\left(\frac{(N-1)^2\pi}{N}\right) \end{pmatrix}$$

La résolution du problème $Au = b$ se ramène donc (on utilise $P = P^t = P^{-1}$) au calcul de $u = PD^{-1}Pb$. Il s'agit donc de 2 produits matrice-vecteur et de la multiplication par une matrice diagonale. Le cœur de la méthode réside dans la manière d'effectuer le produit Pb (et de la même manière Pc avec $c = D^{-1}Pb$). On introduit le vecteur $\tilde{b} = \mathbb{R}^{2N}$ construit de la façon suivante

$$\tilde{b} = \left(\tilde{b}_0, \ldots, \tilde{b}_{2N-1}\right) = (0, b_1, b_2, \ldots, b_{N-1}, 0, -b_{N-1}, -b_{N-2}, \ldots, -b_1).$$

On a

$$\sqrt{\frac{N}{2}}\,(Pb)_k = \sum_{\ell=1}^{N-1} \sin\left(\frac{kl\pi}{N}\right) b_\ell = \frac{1}{2}\left(\sum_{\ell=1}^{N-1} \sin\left(\frac{2kl\pi}{2N}\right) b_\ell - \sum_{\ell=1}^{N-1} \sin\left(\frac{2k(2N-\ell)\pi}{2N}\right) b_\ell\right)$$

$$= \frac{1}{2}\sum_{\ell=0}^{2N-1} \sin\left(\frac{2kl\pi}{2N}\right)\tilde{b}_\ell = \frac{i}{2}\sum_{\ell=0}^{2N-1} \exp\left(-\frac{2ikl\pi}{2N}\right)\tilde{b}_\ell = \frac{i}{2}\sum_{\ell=0}^{2N-1} \omega_{2N}^{k\ell}\tilde{b}_\ell,$$

107. De nombreuses améliorations sont possibles, qui permettent d'accélérer encore le calcul, mais l'approche basique que nous présentons ici donne l'ordre de grandeur de la complexité, c'est à dire du nombre d'opérations nécessaire à la résolution du problème.

avec

$$\omega_{2N} = \exp\left(-\frac{2i\pi}{2N}\right).$$

Le k-ième coefficient de Pb (au facteur $\sqrt{N/2}$ près) est donc le k-ième coefficient de ce que l'on appelle la transformée de Fourier discrète (d'ordre $2N$, avec indexation de 0 à $2N-1$) du vecteur $\tilde{b}$. On note $\mathcal{F}$ cette transformée de Fourier discrète, de telle sorte que

$$\sqrt{\frac{N}{2}}\,(Pb)_k = \left(\mathcal{F}_{2N}(\tilde{b})\right)_k.$$

La somme ci-dessus peut se décomposer de la façon suivante (on sépare les termes impairs et les termes pairs) :

$$\sum_{\ell=0}^{2N-1} \omega_{2N}^{k\ell}\tilde{b}_\ell = \sum_{\ell=0}^{N-1} \omega_{2N}^{2\ell k}\tilde{b}_{2\ell} + \sum_{\ell=0}^{N-1} \omega_{2N}^{(2\ell+1)k}\tilde{b}_{2\ell+1} = \sum_{\ell=0}^{N-1} \omega_N^{\ell k}\tilde{b}_{2\ell} + \omega_{2N}^{k}\sum_{\ell=0}^{N-1} \omega_N^{\ell k}\tilde{b}_{2\ell+1}$$

$$= \mathcal{F}_N(\tilde{b}^0)_k + \omega_{2N}^{-k}\mathcal{F}_N(\tilde{b}^1)_k.$$

où b^0 (resp. b^1) est le vecteur des termes pairs (resp. impairs) de $\tilde{b}$. Précisons que si k est plus grand que N (c'est a priori inutile ici, mais c'est utile pour la suite), on obtient

$$\mathcal{F}_N(\tilde{b}^0)_{k-N} + \omega_{2N}^{-k}\mathcal{F}_N(\tilde{b}^1)_{k-N}.$$

Supposons que l'on sache calculer tous les termes des deux transformées ci-dessus (vecteurs de taille N). On doit effectuer de l'ordre de N multiplications complexes (on néglige ici les constantes multiplicatives). Si N est une puissance de 2, on peut ainsi récursivement calculer les TFD aux différentes échelles, le coût du passage d'une étape à l'autre étant à chaque fois de l'ordre de $2N$. Le nombre d'étape étant de l'ordre de $\log_2 N$, le coût total est de l'ordre de $N\log_2 N$.

Le lecteur avide de curiosités pourra se reporter à la section **??** pour une présentation de ces principes dans le cadre de la transformée de Fourier sur l'espace $\mathbb{Z}_2$ des entiers dyadiques.

Tranformée de Fourier rapide (dimension 2). On considère maintenant le problème de Poisson en dimension 2 sur un maillage cartésien du carré unité, avec $N+1$ points dans chaque direction, y compris les points au bord, donc au total $(N-1)^2$ degrés de liberté. On note u_{ij} la valeur de la solution approchée au point (ih, jh) (avec $h = 1/N$). Le système résultant de la discrétisation par éléments finis du problème s'écrit

$$Au = b,$$

où $A \in \mathcal{M}_{(N-1)^2}(\mathbb{R})$ peut s'écrire par blocs (avec $B \in \mathcal{M}_{N-1}(\mathbb{R})$)

$$A = \begin{pmatrix} C & -\mathrm{Id} & 0 & \cdot & \cdot & 0 \\ -\mathrm{Id} & C & -\mathrm{Id} & 0 & \cdot & 0 \\ 0 & -\mathrm{Id} & C & -\mathrm{Id} & & \cdot \\ \cdot & & \cdot & \cdot & \cdot & \cdot \\ \cdot & & & \cdot & & -\mathrm{Id} \\ 0 & \cdot & \cdot & 0 & -\mathrm{Id} & C \end{pmatrix}, \; C = \begin{pmatrix} 4 & -1 & 0 & \cdot & \cdot & 0 \\ -1 & 4 & -1 & 0 & \cdot & 0 \\ 0 & -1 & 4 & -1 & & \cdot \\ \cdot & & \cdot & \cdot & \cdot & \cdot \\ \cdot & & & \cdot & & -1 \\ 0 & \cdot & \cdot & 0 & -1 & 4 \end{pmatrix},$$

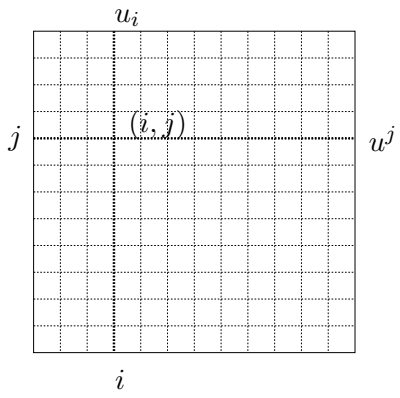


FIGURE 18.3 – Maillage cartésien

et u est le vecteur des inconnues

$$u = (u_{11}, u_{21}, \ldots, u_{N-1,1}, u_{1,2}, \ldots, u_{N-1,N-1})^T$$

On introduit les vecteurs colonne u_i correspondant aux inconnues sur la ligne verticale $x = ih$, et les vecteurs ligne u^j, correspondant aux inconnues sur la ligne horizontale $y = jh$ (voir figure 18.3), ce qui permet d'écrire le vecteur u sous la forme d'une matrice $(u_1, \ldots, u_{N-1}$ (on a une écriture analogue en lignes).

On introduit maintenant la matrice du Laplacien discret (voir (A.13) que l'on note ici Λ. On a (en utilisant une indexation (i,j) pour représenter les vecteurs de $\mathbb{R}^{(N-1)^2}$)

$$(Au)_{i,j} = (\Lambda u_i)_j + \left(\Lambda u^j\right)_i . \tag{18.1}$$

On cherche à réécrire le système de façon plus ramassée en écrivant le vecteur des inconnues sous forme de matrice (deux écriture sont possibles, en colonnes et en lignes)

$$U = (u_1, \ldots, u_{N-1}) = \begin{pmatrix} u^1 \\ u^2 \\ \vdots \\ u^{N-1} \end{pmatrix},$$

on écrit de la même manière le second membre sous la forme d'une matrice B, et l'on remarque

$$\Lambda U = (\Lambda u_1, \ldots, \Lambda u_{N-1}) \ , \ U\Lambda = (\Lambda^T U^T)^T = \begin{pmatrix} \Lambda u^1 \\ \Lambda u^2 \\ \vdots \\ \Lambda u^{N-1} \end{pmatrix}.$$

Le système $(Au)_{i,j} = B_{i,j}$ peut donc s'écrire, d'après (18.1), sous la forme suivante :

$$\Lambda U + U\Lambda = B.$$

Or on a vu que la matrice Λ est diagonalisable (avec une matrice de passage orthogonale et symétrique : $\Lambda = PDP$. On a donc (en multipliant à gauche et à droite par P, et en utilisant $P^2 = \mathrm{Id}$)

$$DPUP + PUPD = PBP.$$

On introduit la matrice $W = PUP$. On s'est finalement ramené au calcul de $B' = PBP$, de la résolution d'un problème du type

$$DW + WD = B' \Longleftrightarrow W_{ij} = \frac{1}{\lambda_i + \lambda_j} B'_{ij},$$

où les λ_i sont connus (voir section A.4), et finalement de $U = PWP$. En dehors de l'étape centrale, pour laquelle on a une formule explicite, il s'agit donc d'effectuer des produits matrice-vecteur du type PX ou XP. Le premier produit consiste en le calcul de la transformée de Fourier discrète (donc potentiellement *rapide*) des vecteurs colonnes de X, et le second $XP = (PX^T)^T$ la TFD des vecteurs lignes de X. Dans les deux cas le calcul par FFT donne une complexité de l'ordre de $N \times N \log_2 N$. On a donc finalement un nombre d'opérations de l'ordre de $m \log_2 m$, où $m = (N-1)^2$ est le nombre d'inconnues.

18.5 Préconditionnement

Les sections précédentes mettent en évidence l'importance du conditionnement dans la rapidité de résolutions des systèmes linéaires, lorsque l'on utilise des méthodes itératives (les plus utilisées dans le cas de grand systèmes linéaires). Il peut être très efficace de remplacer le système $Au = b$ par un système dit préconditionné

$$C^{-1}Au = C^{-1}b.$$

On pourra améliorer très significativement la vitesse de convergence des méthodes si l'on est capable de trouver une matrice C spectralement proche de 1, de telle sorte que le conditionnement de $C^{-1}A$ est très inférieur à celui de A. Pour que cette approche soit efficace, il faut bien sûr que la matrice C soit plus facile à inverser que A.

Un très grand nombre de stratégies sont possibles, parmi lesquelles

1. Préconditionnement diagonal. On prend pour C la matrice diagonale constituée des éléments diagonaux de A. L'inversion de C est alors immédiate, mais l'on vérifie aisément que cette approche est sans intérêt dans certaines situations, par exemple si A est la matrice du Laplacien discrétisé sur maillage cartésien (C est alors proportionnelle à l'identité, de telle sorte que l'on ne change pas le conditionnement de la matrice. En revanche, cette approche peut être féconde dans le cas de maillage très irréguliers, en particuliers lorsque la matrice à inverser est du type $\alpha M + A$, où M est la matrice de masse. Cette approche simpliste peut aussi être efficace dans le cas où la matrice A résulte de la discrétisation par élements finis d'une formulation pénalisée d'un problème sous contrainte.
2. Décomposition incomplète. Dans ce cas, C est construit en effectuant de façon incomplète la décomposition (par exemple de Cholesky) de la matrice A.

Quatrième partie

Aspects théoriques

19 Éléments d'Analyse Fonctionnelle

19.1 Autour du théorème de Hahn-Banach

Théorème 19.1. *(Th. de Hahn-Banach (prolongement))*
Soit E un espace vectoriel normé, G un sous-espace vectoriel de E, et g une forme linéaire sur G, continue. Alors g se prolonge en une forme linéaire continue sur E.

Théorème 19.2. *(Th. de Hahn Banach (séparation))*
Soit E un espace vectoriel normé, X et Y deux convexes de E, non vides, disjoints, avec X fermé et Y compact. Alors il existe un hyperplan fermé qui sépare X et Y au sens strict, i.e. il existe $\varphi \in E'$, $\alpha \in \mathbb{R}$ et $\varepsilon > 0$ tels que

$$\langle \varphi \, , x \rangle \leq \alpha < \alpha + \varepsilon \leq \langle \varphi \, , x \rangle \quad \forall x \in X \, , \; y \in Y.$$

Proposition 19.3. *Soit X un espace vectoriel, et φ, φ_1, …, φ_n des formes linéaires sur X, telles que*

$$\cap \ker \varphi_i \subset \ker \varphi.$$

Alors φ est combinaison linéaire des φ_i.

Démonstration. On considère l'application T qui à $x \in X$ associe $(\varphi(x), \varphi_1(x), \ldots, \varphi_n(x))$ dans $\mathbb{R}^{n+1}$. Par hypothèse, $(1, 0, \ldots, 0)$ n'est pas dans l'image de T, on peut donc séparer ce point de ce convexe fermé par un hyperplan : il existe λ, λ_1, …, λ_n tels que

$$\lambda \leq \alpha < \lambda\varphi(x) + \sum_{i=1}^{n} \lambda_i \varphi_i(x) \quad \forall x \in X.$$

Le membre de droite, linéaire en X et minoré, est nécessairement nul, on a donc

$$\lambda\varphi(x) + \sum_{i=1}^{n} \lambda_i \varphi_i(x) = 0,$$

avec $\lambda < 0$, d'où le résultat.

□

Remarque 19.4. *Le résultat précédent généralise une propriété bien connue sur les matrices. Soit B une matrice réelle $n \times m$, dont les lignes sont les $u_i \in \mathbb{R}^m$, $i = 1, \ldots, n$. Soit u un vecteur orthogonal à tout vecteur orthogonal aux u_i. La proposition précédente (on associe aux vecteurs une forme linéaire basée sur le produit scalaire usuel sur l'espace Euclidien $\mathbb{R}^m$) assure que u est combinaison linéaire des u_i, ce qui exprime*

$$(\ker B)^{\perp} \subset \mathrm{Im} B^T.$$

On a bien sûr égalité entre ces deux espace (l'inclusion inverse est immédiate).

19.2 Autour du théorème de Banach-Steinhaus

Definition 19.5. *On appelle espace de Banach tout espace vectoriel normé complet.*

Definition 19.6. *Soient E et F deux espaces vectoriels normés. On note $\mathcal{L}(E,F)$ l'espace des applications linéaires continues de E dans F. C'est un espace vectoriel normé pour la norme*

$$\|T\|_{\mathcal{L}(E,F)} = \sup_{u\neq 0} \frac{\|Tu\|_F}{\|u\|_E} = \sup_{u\in B_E} \|Tu\|_F.$$

Cet espace est complet dès que F est complet. Lorsque $F = E$, on notera simplement $\mathcal{L}(E)$.

Definition 19.7. *(Adjoint)*
Soient E et F deux espaces vectoriels normés, et $T \in \mathcal{L}(E,\ F)$. On définit l'adjoint de T comme l'opérateur $T^\star$ de F' dans E' qui à $\varphi \in F'$ associe

$$T^\star\varphi \ :\ u \longmapsto \langle T^\star\varphi\,,u\rangle = \langle \varphi\,, Tu\rangle.$$

On vérifie immédiatement que $T^\star \in \mathcal{L}(\ F', E')$, avec $\|T^\star\| = \|T\|$.

Proposition 19.8. *Soit E un espace de Banach, et K un sous-espace vectoriel fermé de E. Pour tout $\tilde{x} \in E/K$, on définit*

$$\|\tilde{x}\|_{E/K} = \inf_{y\in\tilde{x}} \|y\| = \inf_{h\in K} \|x-h\|.$$

L'espace E/K est complet pour la norme $\|\cdot\|_{E/K}$.

Lemme 19.9. *(Baire)*
Soit X un espace métrique complet, et $(X_n)_{n\in\mathbb{N}}$ une suite de fermés de X. On suppose que

$$\mathrm{Int}\,(X_n) = \emptyset \quad \forall n \in \mathbb{N}.$$

On a alors

$$\mathrm{Int}\left(\bigcup_{n=0}^{+\infty} X_n\right) = \emptyset.$$

Théorème 19.10. *(Banach-Steinhaus)*
Soient E et F deux espaces vectoriels normés, avec E complet, et $(T_a)_{a\in A}$ une famille d'opérateurs de $\mathcal{L}(E,F)$. On suppose

$$\sup_{a\in A} \|T_a x\|_F < +\infty \quad \forall x \in E. \tag{19.1}$$

On a alors

$$\sup_{a\in A} \|T_a\|_{\mathcal{L}(E,F)} < +\infty.$$

Exercice 19.1. Montrer qu'un espace de Banach est de dimension soit finie soit non dénombrable.

Corollaire 19.11. *Soient E et F deux espaces de Banach et $(T_n)_{n\in\mathbb{N}}$ une suite d'opérateurs de $\mathcal{L}(E,F)$ telle que, pour tout $x \in E$, $T_n x$ converge vers un élément de F, que l'on note Tx. La suite (T_n) est alors nécessairement bornée dans $\mathcal{L}(E,F)$. De plus, l'opérateur limite T est dans $\mathcal{L}(E,F)$, et sa norme vérifie*

$$\|T\|_{\mathcal{L}(E,F)} \leq \liminf_{n\to+\infty} \|T_n\|_{\mathcal{L}(E,F)}\,.$$

Remarque 19.12. *La dernière inégalité du corollaire précédent peut être stricte. Considérer par exemple $E = \ell^2$ et la suite des formes linéaires*

$$T_k \ : \ x = (x_n)_{n\in\mathbb{N}} \longmapsto x_k \in \mathbb{R}.$$

Cette suite converge ponctuellement vers la forme linéaire nulle. Cet exemple permet d'autre part de vérifier que l'on n'a pas en général convergence de T_k vers T pour la norme d'opérateur.

Remarque 19.13. *On prendra garde au fait que l'hypothèse (19.1) du théorème de Banach-Steinaus, (tout comme l'hypothèse de convergence de $T_n x$ du corollaire ci-dessus), doit être vérifiée pour tout x de E, et non pas seulement sur un sous-ensemble dense.*

Théorème 19.14. *(Application ouverte)*
Soient E et F deux espaces de Banach et soit $T \in \mathcal{L}(E,F)$ surjectif. Alors il existe une constante c telle

$$B_F(0,c) \subset T\left(B_E\right).$$

On en déduit le

Corollaire 19.15. *Soient E et F deux espaces de Banach. et soit $T \in \mathcal{L}(E,F)$ bijectif . Alors T^{-1} est continu de F dans E.*

Dans le cas où T n'est pas surjectif, on peut appliquer ce qui précède à l'application $\tilde{T}$, bijection canoniquement associée à T comme le précise le corollaire ci -dessous.

Corollaire 19.16. *Soient E et F deux espaces de Banach, et $T \in \mathcal{L}(E,\ F)$. On suppose que l'image de T est fermée. L'application $\tilde{T}$ définie de $E/\ker T$ dans $T(V)$ par $\tilde{T}\tilde{x} = Tx$ est une bijection bicontinue. En particulier, il existe une constante α telle que*

$$\|\tilde{u}\|_{E/\ker T} = \inf_{h\in\ker T} \|u-h\| \leq \alpha \|Tu\|.$$

Remarque 19.17. *Dans le cas où E est un espace de Hilbert, l'infimum est atteint pour h égal à la projection de u sur $\ker T$, l'inégalité ci-dessus devient*

$$\left|P_{(\ker T)^\circ} u\right| \leq \alpha \|Tu\|.$$

Proposition 19.18. *Soient E et F deux espaces de Banach, et $T \in \mathcal{L}(E,\ F)$. L'image de T est fermée si et seulement s'il existe $\alpha > 0$ tel que*

$$\forall y \in T(E)\,,\ \exists x \in E\,,\ \|x\| \leq \alpha \|y\|\ ,\ y = Tx. \tag{19.2}$$

Démonstration. La condition nécessaire est une conséquence directe du corollaire précédent. En effet, si l'on note α la constante de continuité de l'aplication $\tilde{T}^{-1}$, on a

$$\forall y \in T(E)\,,\ \left\|\tilde{T}^{-1}y\right\|_{E/\ker T} \leq \alpha \|y\|.$$

Soit z un élément de la classe $\tilde{T}^{-1}y$, on a

$$\left\|\tilde{T}^{-1}y\right\|_{E/\ker T} = \|z - P_{\ker T} z\|,$$

d'où la propriété avec $x = z - P_{\ker T} z$.

Réciproquement, si un tel α existe, alors pour tout suite (x_n) telle que $Tx_n \to y$, on peut construite une suite bornée x'_n avec $Tx_n = Tx'_n$, dont on peut extraire une sous-suite faiblement convergence (toujours notée (x'_n)) vers $x \in E$. La proposition 20.31 assure alors la convergence faible de Tx'_n vers Tx, d'où $y = Tx \in T(E)$. □

Remarque 19.19. *On déduit immédiatement de ce qui précède que l'image d'un sous-espace fermé par une application linéaire injective à image fermée est fermée (comme image réciproque d'un fermé par l'application réciproque, qui est continue).*

Definition 19.20. *(Polaire d'un ensemble)*
Soit E un espace de Banach et K un sous-espace vectoriel de E. On appelle polaire de K l'ensemble

$$K^\circ = \{\varphi \in E' ,\ \langle \varphi , u\rangle = 0 \quad \forall u \in K\} .$$

Les propriétés qui suivent sont essentielles pour établir les résultats afférents à l'existence et l'unicité de point-selle On se reportera à Brezis [2] pour un exposé plus complet des propriétés de l'opérateur adjoint.

Proposition 19.21. *Soient E et F deux espaces de Banach, et $T \in \mathcal{L}(E,\ F)$. On a*

$$\overline{\mathrm{Im} T^\star} \subset (\ker T)^\circ .$$

Dans le cas où E est un espace de Hilbert (et plus généralement dans le cas où E est réflexif), on a l'identité

$$\overline{\mathrm{Im} T^\star} = (\ker T)^\circ .$$

Démonstration: Soit $\varphi \in T^\star(\ F')$, donc de la forme $T^\star\lambda$. On a, pour tout $u \in \ker T$,

$$\langle \varphi , u\rangle = \langle T^\star\lambda , u\rangle = \langle \lambda , Tu\rangle = 0,$$

d'où $T^\star(\ F') \subset (\ker T)^\circ$. Comme $(\ker T)^\circ$ est fermé, cela entraîne $\overline{T^\star(\ F')} \subset (\ker T)^\circ$.

Montrons que cette inclusion ne peut être stricte dans le cas Hilbertien. Supposons qu'elle le soit. Il existe alors $\varphi_0 \in (\ker T)^\circ$ non élément de l'adhérence de $T^\star(\ F')$. Le théorème de Hahn-Banach permet de séparer strictement φ_0 du convexe fermé $\overline{T^\star(\ F')}$: il existe [108] $h \in V$ et $\alpha \in \mathbb{R}$ tels que

$$(T^\star\lambda, h) \leq \alpha < \langle \varphi_0 , h\rangle \quad \forall \lambda \in\ F'.$$

Comme F' est un espace vectoriel, l'ensemble des valeurs prises par $(T^\star\lambda, h)$ est soit $\{0\}$ soit $\mathbb{R}$ tout entier. D'après l'inégalité précédente, c'est nécessairement $\{0\}$. On a donc $\langle \lambda , Th\rangle = 0$ pour tout $\lambda \in\ F'$ d'où $h \in \ker T$, mais alors $\langle \varphi_0 , h\rangle = 0$, ce qui est en contradiction avec l'inégalité ci-dessus. On a donc bien identité entre les deux ensembles. □

Proposition 19.22. *Soient E et F deux espaces de Banach, et $T \in \mathcal{L}(E,\ F)$. Les assertions suivantes sont équivalentes :*

(i) $\mathrm{Im} T$ *est fermée.*

108. C'est ici qu'intervient l'hypothèse de réflexivité de E, dans le fait que la forme linéaire sur E' est de la forme $\varphi \mapsto \langle \varphi , h\rangle$

(ii) $\mathrm{Im}T^\star$ *est fermée.*

(iii) Il existe $C > 0$ *tel que*

$$\forall z \in \mathrm{Im}T\,,\ \exists u \in E\,,\ z = Tu\,,\ \|u\| \leq C\,\|z\|\,,$$

ou, de façon équivalente

$$\|\tilde{u}\|_{E/\ker T} \leq C\,\|Tu\|\,.$$

(iv) Il existe $\beta > 0$ *tel que*

$$\sup_{u\in E} \frac{\langle \lambda\,, Tu\rangle}{\|u\|} \geq \beta\,\|\lambda\|_{\ F'/\ker T^\star}\,.$$

Proposition 19.23. *Soient* E *et* F *deux espaces de Banach, et* $T \in \mathcal{L}(E,\ F)$*. Les assertions suivantes sont équivalentes.*

(i) T *est surjectif.*

(ii) Il existe $\alpha > 0$ *tel que*

$$\|\mu\| \leq \alpha\,\|T^\star\mu\| \quad \forall \mu \in\ F'.$$

(iii) Il existe $\beta > 0$ *tel que*

$$\sup_{u\in E} \frac{|\langle \lambda\,, Tu\rangle|}{|u|\,\|\lambda\|} \geq \beta \quad \forall \lambda \in\ F'.$$

20 Espaces de Hilbert, analyse convexe

20.1 Définitions, principales propriétés

Definition 20.1. *(Produit scalaire)*
Soit H un espace vectoriel sur $\mathbb{R}$. On appelle produit scalaire une forme bilinéaire (u, v) de $H \times H$ dans $\mathbb{R}$, symétrique, définie et positive :

$$(u, v) = (v, u)\,,\ (u, u) \geq 0 \quad \forall u \in H\,, \quad et \ \ (u, u) = 0 \iff u = 0.$$

Un produit scalaire définit sur H une structure d'espace vectoriel normé pour la norme $u \longmapsto |u| = (u, u)^{1/2}$.

Definition 20.2. *(Espace de Hilbert)*
On appelle espace de Hilbert un espace vectoriel muni d'un produit scalaire, et qui est complet pour la norme associée.

Exemple 20.1. Tout espace de dimension finie munie d'un produit scalaire est un espace de Hilbert (espace Euclidien). En dimension infinie, l'exemple le plus simple d'espace de Hilbert de dimension infinie est l'espace ℓ^2 des suites de carré intégrable. On peut définir par extension une infinité de nouveaux espaces dits "à poids" en introduisant, pour $\gamma = (\gamma_n)$ une suite quelconque de réels strictement positifs,

$$\ell^2_\gamma = \left\{(u_n) \in \mathbb{R}^{\mathbb{N}}\,,\ \sum \gamma_n \left|u_n\right|^2 < +\infty\right\}.$$

Proposition 20.3. *(Inégalité de Cauchy-Schwarz)*
Tout produit scalaire vérifie l'inégalité de Cauchy-Schwarz

$$|(u, v)| \leq (u, u)^{1/2}(v, v)^{1/2} \quad \forall u, v \in H.$$

Démonstration: On écrit que $(u + tv, u + tv)$ est positif, pour tout $t \in \mathbb{R}$, notamment pour $t = -(u, v)/\left|v\right|^2$ qui réalise le minimum. □

Proposition 20.4. *(Identité du parallélogramme)*
Toute norme issue d'un produit scalaire vérifie l'identité du parallélogramme

$$\left|\frac{u+v}{2}\right|^2 + \left|\frac{u-v}{2}\right|^2 = \frac{1}{2}(|u|^2 + |v|^2).$$

Proposition 20.5. *Tout sous-espace vectoriel fermé d'un espace de Hilbert est un espace de Hilbert (pour le même produit scalaire).*

Démonstration. La propriété découle simplement du fait que la restriction d'un produit scalaire à un sous-espace est un produit scalaire, et qu'un sous-espace fermé d'un espace complet est complet. □

Notation: Soit H un espace de Hilbert. On appelle boule unité fermée de H l'ensemble

$$B_H = \{u \in H\,,\ |u| \leq 1\}\,.$$

Definition 20.6. *(Séparabilité)*
On dit qu'un espace de Hilbert H est séparable s'il existe un sous-ensemble de H dénombrable et dense dans H.

Théorème 20.7. *(Projection sur un convexe fermé)*
Soit H un espace de Hilbert et K un convexe fermé non vide de H. Pour tout $z \in H$, il existe un unique $u \in K$ (appelée projection de z sur K) tel que

$$|z-u| = \min_{v\in K} |z-v| = \mathrm{dist}(z,K).$$

La projection u est caractérisée par la propriété

$$\begin{cases} u \in K \\ (z-u, v-u) \leq 0 \quad \forall v \in K. \end{cases} \tag{20.1}$$

On notera $u = P_K z$.

Démonstration: On considère une suite minimisante (u_n)

$$u_n \in K\,,\ |z-u_n| \longrightarrow d = \mathrm{dist}(z,K).$$

Pour p, $q \in \mathbb{N}$, on applique l'identité du parallélogramme à $u_p - z$ et $u_q - z$:

$$\left|\frac{u_p+u_q}{2} - z\right|^2 + \left|\frac{u_p-u_q}{2}\right|^2 = \frac{1}{2}(|u_p-z|^2 + |u_q-z|^2).$$

Comme K est convexe $(u_p+u_q)/2 \in K$,

$$\left|\frac{u_p+u_q}{2} - z\right|^2 \geq d^2.$$

On a donc

$$\left|\frac{u_p-u_q}{2}\right|^2 \leq d^2 - d^2 + \varepsilon_p + \varepsilon_q = \varepsilon_p + \varepsilon_q,$$

avec $\varepsilon_n = |u_n - z|^2 - d^2 \longrightarrow 0$. La suite u_n est donc de Cauchy dans H complet, donc converge vers $u \in H$. Comme K est fermé, $u \in K$, et par continuité de la norme, $|u-z| = \mathrm{dist}(z,K)$.

On écrit ensuite simplement que pour tout $v \in K$, l'inégalité $|z-w|^2 \geq |z-u|^2$ est vérifiée pour tout w du segment $[u,v]$ (qu'on écrit $w = u + t(v-u)$, $t \in [0,1]$). □

La démonstration du théorème précédent suggère que toute suite minimisante (u_n) tend nécessairement vers le minimiseur. L'exercice suivant précise cette propriété, en explicitant la vitesse de convergence de la suite des minimiseurs en fonction de la vitesse de convergence de $|u_n - z|$ vers $|u-z|$.

Exercice 20.1. Soit H un espace de Hilbert, K un convexe fermé non vide de H, $z \in H$. On note u la projection de z sur K. Montrer que

$$|v-u| \leq |v-z| \quad \forall v \in K.$$

Exercice 20.2. Soit H un espace de Hilbert, K un convexe fermé non vide de H, $z \in H$. On note u la projection de z sur K. Pour tout $v \in K$, note $d_v = |v-z|$, et $\varepsilon = d_v - d$. Estimer $|v-u|$ en fonction de d_v et ε.

Exercice 20.3. Soit $H = \ell^2$ et K l'ensemble des suites à termes positifs ou nuls. Exprimer la projection d'un élément $z = (z_n)$ sur K.

Remarque 20.8. *Si K est un sous-espace affine fermé de H, alors la caractérisation (20.1) prend la forme*

$$\begin{cases} u \in K \\ (z-u, v-u) = 0 \quad \forall v \in K, \end{cases} \tag{20.2}$$

et si K est un sous-espace vectoriel de H, on a

$$\begin{cases} u \in K \\ (z-u, v) = 0 \quad \forall v \in K. \end{cases} \tag{20.3}$$

Remarque 20.9. *On prendra garde que la projection sur un sous-espace vectoriel n'est en général pas définie, car en dimension infinie les sous-espaces vectoriel peuvent ne pas être fermés (considérer par exemple le sous-espace de ℓ^2 des suites nulles au delà d'un certain rang).*

On peut vérifier que l'application de projection P_K définie par le théorème précédent est 1-lipschitzienne

Proposition 20.10. *Sous les hypothèses du théorème précédent, on a, pour tous f, $g \in H$,*

$$|P_K f - P_K g| \leq |f - g|$$

Démonstration. On utilise la caractérisation de la projection (20.1) :

$$\begin{aligned} (f - P_K f, P_K g - P_K f) &\leq 0, \\ (g - P_K g, P_K f - P_K g) &\leq 0. \end{aligned}$$

En additionnant, il vient,

$$|P_K f - P_K g|^2 \leq (f - g, P_K f - P_K g) \leq |f - g|\,|P_K f - P_K g|\,,$$

d'où l'inégalité annoncée. □

Remarque 20.11. *Ne pas confondre le résultat précédent avec le caractère 1-lipschitzien de la fonction distance à un ensemble quelconque, dans tout espace vectoriel normé.*

La proposition ci-dessus exprime la stabilité de la projection par rapport à l'élément projeté. On peut se demander si cette projection est stable par rapport à l'ensemble sur lequel on projette. C'est l'objet de l'exercice suivant :

Exercice 20.4. Soit H un espace de Hilbert, et z un élément de H fixé. Pour tout couple (K, K') de convexes fermés bornés, on définit leur distance de Hausdorff par

$$d_H(K, K') = \max\left(\sup_{v \in K} d(v, K'),\ \sup_{v' \in K'} d(v', K)\right).$$

On note $u = P_K z$, $u' = P_{K'} z$. Majorer $|u - u'|$ en fonction de $d_H(K, K')$.

Proposition 20.12. *Soit H un espace de Hilbert et K un sous-espace vectoriel fermé de H. Tout u de H s'écrit*

$$u = P_K u + P_{K^\perp} u.$$

Démonstration: On vérifie immédiatement que $u - P_K u$ vérifie les identités qui caractérisent la projection de u sur $K^\perp$. □

Proposition 20.13. *(Caractérisation de la densité)*
Soit H un espace de Hilbert et K un sous-espace de H tel que l'implication suivante soit vérifiée :

$$(h, w) = 0 \quad \forall w \in K \Longrightarrow h = 0.$$

Alors K est dense dans H

Démonstration: Si K n'est pas dense dans H, alors il existe $u \in H$, $u \notin \overline{K}$. On pose $h = u - P_{\overline{K}} u$. On a $(h, w) = 0$ pour tout $w \in K$, et $h \neq 0$ car $u \notin \overline{K}$. □

Théorème 20.14. *(Hahn-Banach)*
Soit H un espace de Hilbert, $K \subset H$ un convexe fermé, et z un point de H qui n'appartient pas à K. Alors il existe un hyperplan fermé qui sépare K et z au sens strict, c'est-à-dire qu'il existe h et x_0 dans H tels que

$$(x - x_0, h) \leq 0 < (z - x_0, h) \quad \forall x \in K.$$

Démonstration: On introduit la projection $u = P_K z$ de z sur K, on définit x_0 comme $(z + u)/2$, et $h = z - u$. Pour tout $x \in K$, on a

$$(x - x_0, h) \;=\; \underbrace{(x - u, z - u)}_{\leq 0} + \underbrace{(u - x_0, h)}_{= -|h|^2/2 \leq 0}$$

et on a par ailleurs $(z - x_0, h) = |h|^2 / 2 > 0$. □

Exercice 20.5. (Lemme des noyaux)
Soient u, u_1, …, u_n, des éléments d'un espace de Hilbert H. Montrer l'équivalence suivante

$$\left(\bigcap u_i^\perp\right) \subset u^\perp \quad \Longleftrightarrow \quad \exists \lambda_1, \ldots, \lambda_n\,,\ u = \sum \lambda_i u_i.$$

Definition 20.15. *(Orthogonal d'un ensemble)*
Soit H un espace de Hilbert et K un sous-ensemble de H. On appelle orthogonal de K l'ensemble

$$K^\perp = \{v \in V\,,\ (v, u) = 0 \quad \forall u \in K\}\,.$$

On vérifie immédiatement que c'est un sous-espace vectoriel fermé.

Proposition 20.16. *Soit H un espace de Hilbert et K un sous-espace vectoriel fermé de H. On a*

$$K^{\perp\perp} = K.$$

Tout espace de Hilbert peut s'identifier à son dual, comme l'exprime le théorème suivant.

Théorème 20.17. *(Riesz-Fréchet)*
Soit $\varphi \in H'$ (dual topologique de H). Il existe $f \in H$ unique tel que

$$\langle \varphi \, , u \rangle = (f, u) \quad \forall u \in H. \tag{20.4}$$

De plus, on a $|f| = \|\varphi\|_{H'}$.

Démonstration: Si φ est la forme nulle, le résultat est immédiat. Dans le cas contraire, on introduit K le noyau de φ. C'est un hyperplan fermé de H. On construit ensuite un $h \in S_H \cap K^\perp$. Pour celà on considère $z \notin K$. D'après la caractérisation (20.3), on a $(z - P_K z, v) = 0$ pour tout $v \in K$. Le vecteur

$$h = \frac{z - P_K z}{|z - P_K z|}$$

convient donc. Pour finir on remarque que tout $v \in H$ peut s'écrire

$$v = \frac{\langle \varphi \, , v \rangle}{\langle \varphi \, , h \rangle} h + \left(v - \frac{\langle \varphi \, , v \rangle}{\langle \varphi \, , h \rangle} h \right) = \lambda h + w,$$

avec $w \in K$. On a donc, pour tout $v \in H$ (on prend le produit scalaire de l'identité précédente avec h),

$$\langle \varphi \, , v \rangle = \langle \varphi \, , h \rangle \, (v, h)$$

d'où l'identité (20.4) avec $f = \langle \varphi \, , h \rangle \, h$. L'unicité d'un tel f est immédiate. □

On prendra garde au fait que cette identification dépend du produit scalaire choisi.

L'identification entre H et son espace dual permet d'étendre immédiatement la caractérisation de la densité 20.13 à un sous-espace du dual :

Proposition 20.18. *(Caractérisation de la densité dans le dual)*
Soit H un espace de Hilbert et K un sous-espace de H' tel que l'implication suivante soit vérifiée :

$$\langle \varphi \, , h \rangle = 0 \quad \forall \varphi \in K \Longrightarrow h = 0.$$

Alors K est dense dans H'.

Proposition 20.19. *(Continuité d'une forme bilinéaire)*
Soit $a : H \times H \longrightarrow \mathbb{R}$ une forme bilinéaire. Alors $a(\,\cdot\,,\,\cdot\,)$ est continue si et seulement s'il existe une constante $\|a\|$ telle que

$$|a(u, v)| \leq \|a\| \, |u| \, |v| \quad \forall u, v \in H.$$

Démonstration. On suppose a continue. La continuité en 0 assure l'existence d'un r tel que $|a(u, v)| \leq 1$ sur $\overline{B(0, r)} \times \overline{B(0, r)}$. On a donc, pour tous u, v, non nuls

$$\left| a\left(r\frac{u}{|u|}, r\frac{v}{|v|} \right) \right| \leq 1 \Longrightarrow |a(u, v)| \leq \frac{1}{r^2} \, |u| \, |v| \, .$$

Réciproquement, le développement

$$a(u + h, v + k) = a(u, v) + a(h, v) + a(u, k) + a(h, k)$$

assure la continuité en tout $(u, v) \in H \times H$. □

Definition 20.20. *(Coercivité d'une forme bilinéaire)*
Soit $a : H \times H \longrightarrow \mathbb{R}$ une forme bilinéaire. On dit que a est coercive s'il existe $\alpha > 0$ tel que

$$a(u,u) \geq \alpha \left|u\right|^2 \quad \forall u \in H.$$

Remarque 20.21. *En dimension finie, et dans le cas où la forme est symétrique ($a(u,v) = a(v,u)$), on retrouve la notion de forme symétrique définie positive. Le plus grand coefficient α est alors la plus petite valeur propre de la matrice associée, et la plus petite constante $\|a\|$ de la continuité sa plus grande valeur propre.*

Exercice 20.6. Soit $\alpha = (\alpha_n)$ une suite bornée de réels, et

$$a \ : \ (u,v) \in \ell^2 \times \ell^2 \longmapsto \sum_{n=0}^{+\infty} \alpha_n u_n v_n.$$

A quelle condition sur α la forme bilinéaire $a(\,\cdot\,,\,\cdot\,)$ est-elle coercive ?

Remarque 20.22. *On verra qu'il existe une définition plus générale de la coercivité (pour des fonctionnelles quelconques, voir théorème 20.44), équivalente à la définition ci-dessus dans le cas particulier des formes bilinéaires.*

Proposition 20.23. *Soit H un espace de Hilbert, et a une forme bilinéaire et continue sur l'espace produit $H \times H$. Pour tout $u \in H$, on note Au l'élément de H qui s'identifie à la forme linéaire $a(u,\,\cdot\,)$:*

$$(Au, v) = a(u,v) \quad \forall v \in H.$$

L'application $u \longmapsto Au$ est linéaire et continue. De plus si $a(\,\cdot\,,\,\cdot\,)$ est coercive, alors l'application A est une bijection.

Démonstration: L'application A est évidemment linéaire, et

$$\left|Au\right| = \sup_{|v|=1} (Au, v) = \sup_{|v|=1} a(u,v) \leq C \left|u\right|,$$

où $\|a\|$ est la constante de continuité de a.

Si a est coercive, on a $(Au,u) = a(u,u) \geq \alpha \left|u\right|^2$, et donc $\left|Au\right| \geq \alpha \left|u\right|$ pour tout u dans H. On vérifie que l'image est fermée en considérant une suite (Au_n) qui converge vers un élément de l'image w. Comme (Au_n) converge, elle est de Cauchy, donc (u_n) est également de Cauchy d'après l'inégalité précédemment démontrée. Elle converge donc vers $u \in H$ qui vérifie $Au = w$ par continuité de A. On a de plus, pour tout $g \in H$,

$$(g, Au) = 0 \quad \forall u \in H \Longrightarrow (g, Ag) = a(g,g) = 0$$

qui entraîne $g = 0$ par coercivité de a. L'image de A est donc fermée et dense dans H : c'est l'espace H lui-même. L'injectivité est une conséquence immédiate de la coercivité. □

Remarque 20.24. *On peut choisir de définir A comme un opérateur de H dans H', en écrivant alors $\langle Au\,, v\rangle = a(u,v)$ pour tout $v \in H$. Les résultats précédents s'étendent bien entendu à cette situation.*

On verra que l'opérateur A est bicontinu (*i.e.* son inverse est lui-même continu), mais cette propriété n'est pas utile pour démontrer le point essentiel de cette section, conséquence directe de la proposition qui précède :

Théorème 20.25. *(Lax-Milgram)*
Soit H un espace de Hilbert, et a une forme bilinéaire continue et coercive sur $H \times H$. Pour tout $\varphi \in H'$, il existe un $u \in H$ unique tel que

$$a(u,v) = \langle \varphi \, , v \rangle \quad \forall v \in H. \tag{20.5}$$

Si a est symétrique, u est l'unique élément de H qui réalise le minimum de la fonctionnelle

$$v \longmapsto J(v) = \frac{1}{2} a(v,v) - \langle \varphi \, , v \rangle.$$

Démonstration. D'après le théorème de représentation de Riesz-Fréchet, il existe un unique $f \in H$ tel que

$$(f,v) = \langle \varphi \, , v \rangle \quad \forall v \in H.$$

On introduit l'opérateur A associé à $a(\,\cdot\,,\,\cdot\,)$, qui est bijectif (voir proposition 20.23). Il existe donc une unique solution u à l'équation $Au = f$.

On suppose maintenant $a(\,\cdot\,,\,\cdot\,)$ symétrique. On note toujours u la solution du problème variationnel (20.6). Pour tout $h \in H$, l'application

$$t \longmapsto \psi(t) = J(u+th) - J(u)$$

est convexe, nulle en 0, de dérivée nulle en 0. Elle est donc positive, et ainsi $J(u+h) \geq J(u)$ pour tout $h \in H$.

De la même manière, si w minimise J, on écrit que la dérivée de la fonction $J(w+th)-J(w)$ est nulle en 0, ce qui est exactement la formulation variationnelle (20.6). □

Corollaire 20.26. *Soit H un espace de Hilbert, $K \subset H$ un sous-espace affine fermé, K^0 l'espace vectoriel sous-jacent. et a une forme bilinéaire continue sur $H \times H$, coercive sur K^0. Pour tout $\varphi \in H'$, il existe un $u \in K$ unique tel que*

$$a(u,v) = \langle \varphi \, , v \rangle \quad \forall v \in K^0. \tag{20.6}$$

Si a est symétrique, u est l'unique élément de K qui réalise le minimum de la fonctionnelle

$$v \longmapsto J(v) = \frac{1}{2} a(v,v) - \langle \varphi \, , v \rangle.$$

Démonstration: On écrit simplement $K = U + K^0$, et l'on cherche la solution sous la forme $u = U + \tilde{u}$, pour se ramener au problème

$$a(\tilde{u},v) = \langle \varphi \, , v \rangle - a(U,v) \quad \forall v \in K^0,$$

qui rentre dans le cadre du théorème de Lax-Milgram. Le principe de minimisation s'en déduit, du fait que

$$\begin{aligned} J(U+h,U+h) &= J(U,U) + \frac{1}{2} a(h,h) + a(U,h) - \langle \varphi \, , U \rangle - \langle \varphi \, , h \rangle \\ &= \frac{1}{2} a(h,h) - (\langle \varphi \, , h \rangle - a(U,h)) + \text{ constante} \end{aligned}$$

□

L'identification établie ci-dessus permet de donner un sens à la notion de différentielle d'une application à valeurs dans $\mathbb{R}$ en tant qu'élément de l'espace de Hilbert :

Definition 20.27. *(Différentiabilité)*
Soit J une application de H dans $\mathbb{R}$, et $u \in H$. On dit que J est différentiable en u s'il existe $\varphi \in H'$ tel que l'on ait, pour h au voisinage de 0,

$$J(u+h) = J(u) + \langle \varphi \, , h \rangle + |h| \, \varepsilon(h),$$

où $\varepsilon : H \longrightarrow H$ est telle que $\varepsilon(h) \longrightarrow 0$ quand $h \longrightarrow 0$. Si un tel φ existe, on peut l'identifier à un élément de H que l'on note $J'(u)$. On dira que J est différentiable si elle admet une différentielle en tout point, et que J est C^1 si l'application $u \longmapsto J'(u)$ est continue.

20.2 Convergence faible

Comme précédemment H désigne un espace de Hilbert réel muni du produit scalaire $(.,.)$ et de la norme $|\ |$.

Definition 20.28. *(Convergence faible)*
Soit (u_n) une suite d'éléments de H. On dit que (u_n) converge faiblement vers u dans H, et on note $u_n \rightharpoonup u$, si

$$(u_n, v) \to (u, v) \quad \forall v \in H,$$

ou de façon équivalente, si

$$< \varphi, u_n > \longrightarrow < \varphi, u > \quad \forall \varphi \in H'.$$

Proposition 20.29. *Soit (u_n) une suite d'un espace de Hilbert H. Si $u_n \rightharpoonup u$, alors (u_n) est bornée et $|u| \leq \liminf |u_n|$.*

Démonstration: C'est une conséquence directe du corollaire 19.11 au théorème de Banach-Steinhaus. □

Proposition 20.30. *Si $u_n \rightharpoonup u$ et $|u_n| \to |u|$, alors la suite u_n converge fortement vers u.*

Démonstration: On écrit

$$|u_n - u|^2 = |u_n|^2 - 2(u_n, u) + |u|^2 .$$

On a $(u_n, u) \to |u|^2$ d'où $|u_n - u|^2 \to 0$. □

Proposition 20.31. *Soient E et F deux espaces de Hilbert, et $T \in \mathcal{L}(E, F)$. Alors*

$$u_n \rightharpoonup u \Longrightarrow Tu_n \rightharpoonup Tu.$$

Démonstration: On écrit simplement que, pour tout $z \in F$,

$$(Tu_n, z) = (u_n, T^\star z) \longrightarrow (u, T^\star z) = (Tu, z),$$

qui exprime la convergence faible de Tu_n vers Tu. □

Le résultat fondamental de cette section est le suivant.

Théorème 20.32. *Soit (u_n) une suite* **bornée** *dans un espace de Hilbert H. Alors on peut extraire une sous-suite convergeant faiblement vers u dans H.*

Démonstration: On raisonne d'abord dans le cas où H est séparable. Il existe donc une famille dénombrable $\{x_k\}_{k\in\mathbb{N}}$ dense dans H. On se propose de suivre le procédé d'extraction diagonale de Cantor.

1. Comme (u_n, x_1) est bornée dans $\mathbb{R}$ on peut extraire une suite $u_{j_1(n)}$ telle que $(u_{j_1(n)}, x_1)$ converge.
2. Comme $(u_{j_1(n)}, x_2)$ est bornée dans $\mathbb{R}$ on peut extraire de $u_{j_1(n)}$ une suite $u_{j_1\circ j_2(n)}$ telle que $(u_{j_1\circ j_2(n)}, x_2)$ converge.
3. Par récurrence, on construit une suite de sous-suites emboitées $u_{j_1\circ j_2\circ\cdots\circ j_k(n)}$ telle que $(u_{j_1\circ j_2\circ\cdots\circ j_k(n)}, x_k)$ converge, pour tout k.
4. On utilise à présent le procédé d'extraction diagonale : on pose $\varphi(k) = j_1\circ j_2\circ\cdots\circ j_k(k)$ (de telle sorte que φ est strictement croissante), et on considère $u_{\varphi(n)}$. Pour tout k, on remarque que $u_{\varphi(n)}$, à partir du rang k, est aussi une suite extraite de $(u_{j_1\circ j_2\circ\cdots\circ j_k(n)})$, de telle sorte que $(u_{\varphi(n)}, x_k)$ converge lorsque $n \to +\infty$.
5. On utilise ensuite la densité des x_k. Pour tout $x \in H$, on montre que $(u_{\varphi(n)}, x)$ est une suite de Cauchy : soit $\varepsilon > 0$, il existe (x_k) tel que $|x - x_k| < \varepsilon$. Comme $(u_{\varphi(n)}, x_k)$ est de Cauchy, il existe un N au-delà duquel $\left|(u_{\varphi(p)}, x_k) - (u_{\varphi(q)}, x_k)\right| < \varepsilon$. Pour tous p, q supérieurs à N, on a donc

$$\begin{aligned}\left|(u_{\varphi(p)}, x) - (u_{\varphi(q)}, x)\right| &\leq \left|(u_{\varphi(p)}, x) - (u_{\varphi(p)}, x_k)\right| + \left|(u_{\varphi(p)}, x_k) - (u_{\varphi(q)}, x_k)\right| \\ &\quad + \left|(u_{\varphi(q)}, x_k) - (u_{\varphi(q)}, x)\right| \\ &\leq M\varepsilon + \varepsilon + M\varepsilon = (1 + 2M)\varepsilon,\end{aligned}$$

   où M est un majorant de $|u_n|$.
   On a donc démontré que, pour tout $x \in H$, $(u_{\varphi(n)}, x)$ converge vers un élément de H que l'on note $h(x)$. L'application $x \mapsto h(x) \in \mathbb{R}$ est linéaire, et on a pour tout $x \in H$

$$|h(x)| = \lim_{n\to\infty} \left|(u_{\varphi(n)}, x)\right| \leq M\,|x|\,,$$

   d'où h continue [109] sur H. D'après le théorème de Riesz-Fréchet, cette forme s'identifie à un élément u de H. On a donc convergence faible de la suite extraite vers u.

Dans le cas où le Hilbert n'est pas séparable, on se place dans l'adhérence de l'espace vectoriel engendré par les termes de la suite, qui est un espace de Hilbert séparable (pour le même produit scalaire) par construction. La convergence faible vers un u de ce sous-espace entraîne la convergence faible dans H.

109. Remarquer qu'il n'est pas nécessaire ici d'utiliser le théorème de Banach–Steinhaus, du fait de l'hypothèse (u_n) bornée.

20.3 Somme Hilbertiennes, bases Hilbertiennes

Definition 20.33. *(Somme Hilbertienne)*
Soit $(E_n)_{n\in\mathbb{N}}$ une suite de sous-espaces fermés d'un espace de Hilbert H. On dit que H est somme Hibertienne des E_n si

(i) Les E_n sont deux à deux orthogonaux, c'est-à-dire

$$(u,v)=0 \quad \forall u\in E_n\,,\ \forall v\in E_m \quad \forall m,n\in\mathbb{N}\,,\ m\neq n.$$

(ii) L'espace vectoriel engendré par les E_n est dense dans H.

Théorème 20.34. *On suppose H somme Hilbertienne des E_n. Pour $u\in H$, on note $u_n = P_{E_n}u$. On a*

$$u=\sum_{i=1}^{\infty} u_n \ \text{ et } \ |u|^2=\sum_{i=1}^{\infty}|u_n|^2\,.$$

Réciproquement, si l'on considère une suite (u_n) avec $u_n\in E_n$ pour tout n, et telle que $\sum|u_n|^2$ converge, alors la série $\sum u_n$ converge, et sa limite $u=\sum u_n$ est telle que $u_n=P_{E_n}u$.

Démonstration. On considère l'opérateur

$$S_k=\sum_{n=1}^{k} P_{E_n}.$$

On a $S_k\in\mathcal{L}(H)$, et S_ku vérifie (les E_n sont orthogonaux deux à deux)

$$|S_ku|^2=\sum_{n=1}^{k}|u_n|^2\,.$$

D'autre part on a, pour tout n

$$(u,u_n)=|u_n|^2\,,$$

d'où, en sommant de 1 à k,

$$(u,S_ku)=|S_ku|^2\,.$$

On a donc $|S_ku|\leq|u|$. On désigne par E l'espace vectoriel engendré par les E_n. Pour tout $\varepsilon>0$, tout u dans H, il existe un $v\in E$ tel que $|v-u|<\varepsilon$. Pour k assez grand, on a $S_kv=v$, et ainsi

$$|S_ku-u|\leq|S_k(u-v)|+|v-u|\leq 2\varepsilon.$$

on a donc bien convergence de S_ku vers u.

D'autre part l'égalité, pour tout k

$$|S_ku|^2=\sum_{n=1}^{k}|u_n|^2\,,$$

entraîne, à la limite,

$$|u|^2=\sum_{n=1}^{+\infty}|u_n|^2\,.$$

Pour la réciproque, on utilise le caractère de Cauchy de la suite $\sum_{n=1}^{k}u_n$, et la continuité des opérateurs de projection. □

Le théorème précédent permet d'introduire la notion de base Hilbertienne :

Definition 20.35. *(Bases hilbertiennes)*
Soit $(e_n)_{n\in\mathbb{N}}$ une famille de vecteurs d'un espace de Hilbert H. On dit que (e_n) est une base Hilbertienne si

(i) $|e_n| = 1$ pour tout $n \in \mathbb{N}$, et $(e_m, e_n) = 0$ pour tous m, n, avec $m \neq n$.

(ii) L'espace vectoriel engendré par les (e_n) est dense dans H.

Théorème 20.36. *Tout espace de Hilbert séparable admet une base Hilbertienne.*

Démonstration. Soit H un espace de Hilbert séparable [110]. On considère $(f_n)_{n\in\mathbb{N}}$ une famille dense dans H. On note F_k l'espace vectoriel engendré par les k premiers vecteurs. L'espace vectoriel engendré par les F_k est dense dans H. On peut construire la base Hilbertienne de la façon suivante : si f_1 est non nul, on prend $f_1/|f_1|$ comme premier vecteur. Une base orthonormale sur F_k étant construite, on complète par une base orthonormale sur F_{k+1} si nécessaire (si $f_{k+1} \notin F_k$). Sinon, on passe au rang suivant. □

20.4 Minimisation de fonctionnelles convexes

Commençons par définir un certain nombre de notions générales afférentes aux applications à valeurs dans $\mathbb{R} \cup \{+\infty\}$.

Definition 20.37. *(Domaine)*
Soit E un ensemble et J une application de E dans $\mathbb{R} \cup \{+\infty\}$. On appelle domaine de J l'ensemble

$$D(J) = \{x \in E\,,\ J(x) < +\infty\}\,.$$

Definition 20.38. *(Semi-continuité inférieure)*
Soit E un espace topologique, et J une application de E dans $\mathbb{R} \cup \{+\infty\}$. On dit que J est semi-continue inférieurement (s.c.i. en abrégé) si, pour tout $\lambda \in \mathbb{R}$, l'ensemble

$$E_\lambda = \{x \in E\,,\ J(x) \leq \lambda\}$$

est fermé.

Definition 20.39. *(Convexité)*
Soit E un espace vectoriel, et J une application de E dans $\mathbb{R} \cup \{+\infty\}$. On dit que J est convexe si

$$J(\theta x + (1-\theta)y) \leq \theta J(x) + (1-\theta)J(y) \quad \forall x, y \in E \quad \forall \theta \in]0,1[,$$

ou, de façon équivalente, si l'ensemble (appelé épigraphe de J)

$$\operatorname{epi} J = \{(x, \lambda) \in E \times \mathbb{R}\,,\ J(x) \leq \lambda\}\,,$$

est convexe.

On dit que J est strictement convexe si

$$J(\theta x + (1-\theta)y) < \theta J(x) + (1-\theta)J(y) \quad \forall x, y \in E \quad \forall \theta \in]0,1[.$$

110. C'est à dire qu'il existe un ensemble dénombrable et dense. C'est le cas pour l'essentiel des espace de Hilbert que l'on rencontre dans la "nature", en particulier pour les espaces fonctionnels de type $L^2(\Omega)$ ou $H^m(\Omega)$.

Definition 20.40. *(Coercivité)*
Soit E un vectoriel normé, et J une application de E dans $\mathbb{R} \cup \{+\infty\}$. On dit que J est coercive si

$$\lim_{\|x\| \to +\infty} J(x) = +\infty.$$

Théorème 20.41. *(Banach-Saks)*
Soit $(x_n)_{n\in\mathbb{N}}$ une suite de H faiblement convergente vers un élément x de H. Alors il existe une suite extraite $y_n = x_{\varphi(n)}$ telle que la suite des moyennes de Césaro

$$\sigma_n = \frac{1}{n} \sum_{k=1}^{n} y_k$$

converge fortement vers x.

Démonstration. Quitte à remplacer la suite x_n par $x_n - x$, on peut supposer sans perte de généralité que $x_n \rightharpoonup 0$. On construit maintenant la suite y_n de la façon suivante :

1. On prend $y_1 = x_1$.
2. Comme x_n converge faiblement vers 0, il existe un indice $\varphi(2)$ tel que
$$\left|(y_1, x_{\varphi(2)})\right| = |(y_1, y_2)| \leq \frac{1}{2}.$$
3. Par récurrence, on construit à partir des termes déjà construits y_1, y_2, …, y_{n-1}, le $n-$ième terme y_n tel que
$$|(y_i, y_n)| \leq \frac{1}{n} \quad \forall i = 1, 2, \ldots, n-1.$$

On pose

$$\sigma_n = \frac{1}{n} \sum_{k=1}^{n} y_k.$$

Montrons que σ_n tend (fortement) vers 0. On développe

$$|\sigma_n|^2 = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} (y_i, y_j),$$

ce qui donne

$$\begin{aligned} |\sigma_n|^2 &\leq \frac{1}{n^2} \left(\sum_{i=1}^{n} |y_i|^2 + 2 \sum_{k=1}^{n} \sum_{\ell=1}^{k-1} |(y_\ell, y_k)| \right) \leq \frac{1}{n^2} \left(nM^2 + 2 \sum_{k=1}^{n} \frac{k-1}{k} \right) \\ &\leq \frac{1}{n^2} \Big(nM^2 + 2n \Big) = \frac{M^2 + 2}{n}, \end{aligned}$$

et donc $\sigma_n \to 0$. □

Ce théorème a plusieurs conséquences importantes, dont la première est le

Théorème 20.42. *Soit $K \subset H$ un ensemble convexe fermé de H. Soit $(x_n)_{n\in\mathbb{N}}$ une suite d'éléments de K qui converge faiblement vers x. Alors $x \in K$. On dit que K est faiblement séquentiellement fermé.*

Démonstration: Le résultat est une conséquence directe du théorème 20.41. □

Exercice 20.7. Montrer que le résultat est faux en général si l'on supprime l'hypothèse de convexité (donner par exemple une suite dans la sphère unité de ℓ^2 qui converge faiblement vers 0).

Une autre conséquence importante du théorème 20.41 est le

Théorème 20.43. *Soit $J : H \longrightarrow \mathbb{R}$ une fonction convexe continue s.c.i, $J \not\equiv +\infty$. Pour toute suite $(x_n)_{n\in\mathbb{N}}$ de H telle que $x_n \rightharpoonup x$, on a*

$$J(x) \leq \liminf J(x_n).$$

(On dit que J est faiblement séquentiellement s.c.i.)

Démonstration: Soit $L := \liminf J(x_n)$ (a priori, $-\infty \leq L \leq +\infty$). Soit y_n une suite extraite telle que l'on ait

$$J(y_n) \longrightarrow L,$$

et telle que

$$\sigma_n = \frac{1}{n}\sum_{i=1}^{n} y_n \longrightarrow x.$$

par semi–continuité inférieure de J, on a $J(x) \leq \liminf J(\sigma_n)$. D'autre part, J étant convexe

$$J(\sigma_n) \leq \frac{1}{n}\sum_{i=1}^{n} J(y_n) \to L.$$

On a donc bien $J(x) \leq L$. □

Ce théorème va nous permettre d'établir le résultat principal de minimisation :

Théorème 20.44. *Soit $J : H \longrightarrow \mathbb{R}$ une fonction convexe s.c.i., $J \not\equiv +\infty$. On suppose que J est coercive, c'est-à-dire que*

$$\lim_{|x|\to+\infty} J(x) = +\infty.$$

Alors il existe $u \in H$ tel que

$$J(u) = \min_{v\in H} J(v).$$

Plus généralement, si $K \subset H$ est un convexe fermé, il existe $u \in K$ tel que

$$J(u) = \min_{v\in K} J(v).$$

Enfin, si J est strictement convexe, alors ces minima sont uniques.

Démonstration: Soit $(x_n)_{n\in\mathbb{N}}$ une suite minimisante : $x_n \in K$ et

$$J(x_n) \to M := \inf_K J.$$

Comme J est coercive, x_n est bornée. Il existe donc une suite extraite y_n telle que $y_n \rightharpoonup x$. Comme K est un convexe fermé, $x \in K$, et

$$J(x) \leq \liminf J(x_n) = M.$$

Mais comme $J(x) > M$ par définition de M, on a $J(x) = M$. □

On remarquera que, pour le résultat concernant K, il suffit que J soit définie sur K. La coercivité signifie que, ou bien K est borné, ou bien

$$\lim_{|x|\to+\infty, x\in K} J(x) = +\infty.$$

Definition 20.45. *(Sous-différentiel)*
Soit H un espace de Hilbert, et Ψ une fonctionnelle convexe de H dans $\mathbb{R}\cup\{+\infty\}$. On définit le sous-différentiel de Ψ en $u \in H$ comme l'ensemble

$$\partial\Psi(u) = \{w \in H\,,\ \Psi(u) + (w,h) \leq \Psi(u+h) \quad \forall h \in H\}\,.$$

20.5 Opérateurs maximaux monotones

Definition 20.46. *(Opérateurs maximaux monotones)*
Soit H un espace de Hilbert, et A une application de H dans 2^H (ensemble des parties de A). On appelle $D(A)$ le domaine de A, i.e. l'ensemble des x tels que $Ax \neq \emptyset$. On dit que A est monotone si

$$\forall x\,,\ x' \in D(A)\,,\ \forall y \in Ax\,,\ y' \in Ax'\,,\ (y'-y, x'-x) \geq 0.$$

On dit que A est maximal *monotone si*

$$A \subset A' \text{ et } A' \text{ monotone } \Longrightarrow A' = A.$$

(par $A \subset A'$) on entend $Ax \subset A'x$ pour tout $x \in H$.

Exercice 20.8. Montrer qu'une fonction f continue croissante de $\mathbb{R}$ dans $\mathbb{R}$ est maximale monotone.

Si f est simplement croissante, construire l'unique fonction maximale monotone qui contient f.

Que se passe-t-il pour une fonction qui tend vers $+\infty$ quand x tend vers a^-, $a \in \mathbb{R}$?

On s'intéresse à des problèmes d'évolution de type

$$\frac{du}{dt} + Au \ni 0\,,\ u(0) = u_0. \tag{20.7}$$

Théorème 20.47. *(Voir [3])*
Soit H un espace de Hilbert et A un opérateur maximal monotone. Pour tout $u_0 \in D(A)$, l'équation (20.7) *admet une solution u de $[0,+\infty[$ dans $D(A)$, au sens suivant*

1. *u est Lipschitzienne ;*
2. *L'équation* (20.7) *est vérifiée presque partout sur $]0,+\infty[$;*
3. *La condition initiale est vérifiée (u étant continue, la condition $u(0) = u_0$ a bien un sens).*

Une telle solution est unique. Elle est de plus dérivable à droite, et l'on a, pour tout $t \in [0,+\infty[$,

$$\frac{du}{dt} = -A^\circ u,$$

où $A^\circ u$ est l'élément de Au de norme minimale.

Ce théorème assure l'existence et l'unicité de solution à des équations d'évolution qui ne rentrent pas dans le cadre du théorème de Cauchy-Lipchitz.

Exemple 20.2. On considère l'opérateur

$$\varphi \,:\, x \in \mathbb{R} \longmapsto \left| \begin{array}{ccc} \{-1\} & \text{si} & x < 0, \\ [-1,1] & \text{si} & x = 0, \\ \{1\} & \text{si} & x > 0, \end{array} \right.$$

Pour toute valeur initiale x_0, la solution unique rejoint 0 à vitesse constante de module 1, puis y stationne.

Noter que si l'on prend l'opposé de cet opérateur, on perd l'unicité : partant de 0, on peut aller vers la droite ou la gauche.

On considère les éléments de Ax comme des vitesses de trajectoires issues de x (noter que, d'après l'équation (20.7), un élément de Ax est effectivement homogène à une vitesse). Le caractère maximal monotone implique que des particules issues de deux points distincts ne se croisent jamais :

Proposition 20.48. *Soit A un opérateur maximal monotone sur H. On a*

$$x_1 \neq x_2\,,\ u_1 \in Ax_1\,,\ u_2 \in Ax_2 \Longrightarrow x_1 + tu_1 \neq x_2 + tu_2 \quad \forall t \geq 0.$$

21 Équations différentielles ordinaires

21.1 Lemme(s) de Gronwall

Proposition 21.1. *Soit φ et g deux fonctions continues sur l'intervalle $[0,T]$, toutes deux positives sur cet intervalle. On suppose qu'il existe une constante $C \geq 0$ telle que*

$$\varphi(t) \leq C + \int_0^t g(s)\varphi(s)\,\mathrm{ds} \quad \forall t \in [0,T].$$

On a alors

$$\varphi(t) \leq C \exp\left(\int_0^t g(s)\,\mathrm{ds}\right) \quad \forall t \in [0,T].$$

Démonstration: On suppose tout d'abord $C > 0$. La fonction $z(t) = C + \int_0^t g(s)\varphi(s)$ est dérivable et de dérivée $z' = g\varphi \leq gz$. On a donc (on sait que z par définition ne s'annule pas)

$$\frac{z'}{z} \leq g \Longrightarrow \varphi \leq z(t) \leq z(0)\exp\left(\int_0^t g(s)\,\mathrm{ds}\right) = C\exp\left(\int_0^t g(s)\,\mathrm{ds}\right).$$

Le cas $C = 0$ est obtenu par passage à la limite. □

On peut affaiblir les hypothèses ci-dessus : pour $\varphi \in L^\infty$ et $g \in L^1$, positives presque partout, la conclusion est la même.

Dans le cas où $g \equiv M =$ constante, on a $\varphi(t) \leq C\exp(Mt)$.

La proposition suivante permet d'obtenir, pour les systèmes dynamiques tels que ceux étudiés au chapitre I, des estimations de meilleure qualité (sans le facteur à croissance exponentielle).

Proposition 21.2. *Soit φ et g deux fonctions continues sur l'intervalle $[0,T]$, toutes deux positives sur cet intervalle. On suppose qu'il existe une constante $C > 0$ telle que*

$$\varphi(t) \leq C + 2\int_0^t g(s)\sqrt{\varphi(s)}\,\mathrm{ds} \quad \forall t \in [0,T].$$

On a alors

$$\varphi(t) \leq \left(\sqrt{C} + \int_0^t g(s)\,\mathrm{ds}\right)^2 \quad \forall t \in [0,T].$$

Démonstration. La démonstration est analogue à la précédente, en considérant maintenant la fonction

$$z(t) = C + 2\int_0^t g(s)\sqrt{\varphi(s)}.$$

□

Théorème 21.3. *(Point fixe de Picard)*
Soit X un espace métrique complet, et T une application de X dans X strictement contractante, c'est à dire telle qu'il existe $k \in]0,1[$ tel que

$$d(T(y),T(x)) \leq kd(y,x).$$

Alors T admet un unique point fixe, c'est à dire qu'il existe $x \in X$ tel que $T(x) = x$.

Il suffit de supposer qu'il existe p tel que $T^p = T \circ T \cdots \circ T$ soit strictement contractante.

Démonstration. On prend $x_0 \in X$ et l'on construit la suite $x_1 = T(x_0)$, $x_2 = T(x_1)$, ...
On a

$$d(x_{n+1}, x_n) \leq kd(x_n, x_{n-1}) \leq \cdots \leq k^n d(x_1, x_0).$$

La suite (x_n) est donc de Cauchy dans X, et donc converge vers $x \in X$, qui vérifie, par passage à la limite dans la relation de récurrence, $x = T(x)$. Ce point fixe est unique, car s'il en existait un autre x' on aurait

$$d(x, x') = d(T(x), T(x')) \leq kd(x, x') < d(x, x'),$$

ce qui est absurde.
Si maintenant on suppose que T^p est strictement contractante, alors T^p admet un point fixe x. Par suite $T(x)$ est aussi point fixe de T^p, il s'identifie donc à x par unicité. On a donc bien $T(x) = x$. □

21.2 Théorème de Cauchy Lipschitz

Soit E un espace de Banach Étant donnés un ouvert U de E, $x_0 \in U$, un intervalle ouvert I de $\mathbb{R}$ contenant 0, une fonction f de $U \times I$ dans E, le problème de Cauchy consiste à trouver $t \in I \longmapsto x(t) \in U$ vérifiant

$$\begin{cases} \dot{x}(t) &= f(x,t), \\ f(t_0) &= x_0. \end{cases} \tag{21.1}$$

Definition 21.4. *(Cylindre de sécurité)*
On appelle cylindre de sécurité pour (x_0, t_0) un ensemble $B_f(x_0, r) \times [t_0 - \eta, t_0 - \eta]$ tel que toute solution $x(t)$ du problème de Cauchy sur $[t_0 - \eta, t_0 + \eta]$ soit contenue dans $B_f(x_0, r)$, et tel que $\|f\|$ est borné par une constante M sur le cylindre, avec $r \leq \eta M$.

Definition 21.5. *(Caractère Lipschitz local)*
On dit que $f : U \times I \mapsto E$ est localement Lipschitzienne par rapport à la première variable si en tout point $(y, t) \in U \times I$, il existe $r > 0$, $\eta > 0$ et une constante $k > 0$ tels que

$$\|f(y_2, s) - f(y_1, s)\| \leq k \|y_2 - y_1\| \qquad \forall y_1\,,\ y_2 \in B_f(y, r)\,,\ s \in [t - \eta, t + \eta].$$

Proposition 21.6. *On suppose que f est continue sur $U \times I$ et localement lipschitzienne par rapport à la première variable. Alors f admet un cylindre de sécurité en tout point $(x_0, t_0) \subset U \times I$.*

Démonstration: Montrons l'existence d'un cylindre de sécurité en $(x_0, 0)$. La fonction f est Lipschitzienne par rapport à la première variable sur un ensemble du type $B_f(x_0, r) \times [-\tau, \tau]$. Elle est donc notamment bornée par $M > 0$. On choisit $\eta = \min(\tau, r/M)$. Toute solution est telle que

$$\|x(t) - x_0\| = \left\| \int_0^t f(x(s), s)\, ds \right\| \leq Mt \leq M\eta \leq r,$$

ce qui assure que $B_f(x_0, r) \times [-\eta, \eta]$ est un cylindre de sécurité. □

Remarque 21.7. *Si E est un espace vectoriel de dimension finie, il suffit de supposer la continuité par rapport au couple (x,t), qui assure l'uniforme continuité (et donc le caractère borné) sur tout compact $B_f(x_0,r)\times[t_0-\tau,t_0-\tau]$, d'où l'existence d'un cylindre de sécurité.*

Definition 21.8. *(Solution maximale)*
On appelle solution maximale du problème de Cauchy (21.1) une fonction $t\mapsto x(t)\in E$ définie sur un intervalle $J\subset I$, solution de (21.1), et qui ne peut pas être prolongée sur un intervalle de temps plus grand, ce que l'on peut exprimer de la manière suivante : si $t\mapsto y(t)\in U$ est solution de (21.1) sur J', et s'identifie à x sur $J\cap J'$, alors nécessairement $J'\subset J$.

Théorème 21.9. *(Cauchy-Lipschitz)*
On considère une donnée de Cauchy $(x_0,t_0)\in U\times I$ (avec U ouvert du Banach E et $I\subset\mathbb{R}$ intervalle ouvert, et on suppose que la fonction f, définie de $U\times I$ dans E, est continue sur $\mathcal{U}\times I$ et localement Lipschitzienne par rapport à la première variable. Alors le problème de Cauchy (21.1) admet une unique solution maximale définie sur $J\subset I$.

Démonstration. La fonction f est Lipschitzienne sur un voisinage de (x_0,t_0), et la proposition 21.6 assure l'existence d'un cylindre de sécurité $B_f(x_0,r)\times[t_0-\eta,t_0+\eta]$ construit dans ce voisinage, de telle sorte que $\eta M\leq r$, où M majore la norme de f sur ce cylindre. On introduit l'espace X des applications continues sur $[\eta,\eta]$ à valeurs dans $B_f(x_0,r)$, muni de la norme de la convergence uniforme, et pour tout $x\in X$, on définit Tx par

$$Tx(t)=x_0+\int_{t_0}^{t}f(x(s),s)\,ds.$$

On a $\|Tx(t)-x_0\|\leq M\eta\leq r$, et ainsi T est une application de X dans lui-même, et une solution du problème de Cauchy définie sur $[\eta,\eta]$ est exactement un point fixe de T.
Montrons qu'il existe $n\in\mathbb{N}$ tel que T^n soit strictement contractante. Soient y, $z\in X$. On note $y_n=T^ny$ (de même pour z). On a

$$\|z_1(t)-y_1(t)\|=\left\|\int_{t_0}^{t}(f(z(s),s)-f(y(s),s)\right\|\leq kt\,\|z-y\|_\infty\,.$$

De même

$$\|z_1(t)-z_2(t)\|=\left\|\int_{t_0}^{t}(f(z_1(s),s)-f(y_1(s),s)\right\|\leq k^2\left|\int_{t_0}^{t}s\,ds\right|\ \|z-y\|_\infty=\frac{k^2t^2}{2}\,\|z-y\|_\infty\,.$$

On montre ainsi par récurrence que

$$\|z_n(t)-z_n(t)\|\leq\frac{k^nt^n}{n!}\,\|z-y\|_\infty\ \text{ d'où }\ \|z_n-z_n\|_\infty\leq\frac{k^n\eta^n}{n!}\,\|z-y\|_\infty$$

de telle sorte que T^n est contractante pour n suffisamment grand. D'après le théorème 21.3, l'application T admet un unique point fixe, et l'on en déduit l'existence d'une solution au problème de Cauchy définie sur $[t_0-\eta,t_0+\eta]$, et unique solution sur cet intervalle.
Soit maintenant J la réunion des intervalles sur lesquels le problème de Cauchy associé à (x_0,t_0) admet une solution. On considère deux solutions x_1 et x_2 du problème de Cauchy, définies sur J_1 et J_2, et l'on introduit l'ensemble

$$K=\{\ t\in J_1\cap J_2\,,\ x_1(t)=x_2(t)\}\,.$$

Il est non vide car $0\in K$, c'est un fermé par continuité de x_1 et x_2 comme fonctions de $J_1\cap J_2$ dans E. Par unicité locale de la solution établie précédemment, c'est également un ouvert. Il s'agit donc de l'intervalle $J_1\cap J_2$ tout entier. On en déduit ainsi l'existence et l'unicité d'une solution maximale. □

21.3 Comportement des solutions

Proposition 21.10. *(Sortie des compacts)*
On se place dans le cadre du théorème 21.9, et l'on note x la solution maximale, définie sur $J =]\tau^-, \tau^+[$. Si J est strictement inclus dans $I =]T^-, T^+[$, par exemple si $\tau^+ < T^+$, alors x sort de tout compact de U lorsque t tend vers τ^+, i.e.

$$\forall K \text{ compact } \subset U\,,\ \exists \eta\,,\ x(t) \notin K \quad \forall t > \tau^+ - \eta,$$

avec un comportement analogue au voisinage de τ^-.

Démonstration: Si la propriété n'est pas vérifiée, il existe un compact $K \subset U$ et une suite (t^n) (croissante) tendant vers τ^+ tels que $x(t^n) \in K$ pour tout n. On peut extraire une sous-suite (que l'on note toujours (t^n)) qui converge vers un éléments x_∞ de K. On peut placer un cylindre de sécurité $B_f(x_\infty, r) \times [\tau^+ - \eta, \tau^+ + \eta]$ sur lequel f est majoré par M, avec $r \leq \eta M$, et sur lequel elle est Lipschitzienne. Pour n assez grand, $x(t^n)$ est dans $B_f(x_\infty, r)$, et $\tau^+ - t^n < \eta/2$. On peut alors reproduire la démonstration de construction d'une solution locale proposée pour le théorème de Cauchy-Lipschitz, qui permet de construire une solution au problème de Cauchy associé aux données $(x(t^n), t^n)$ et définie sur $[t^n, t^n + \eta]$. Cette solution s'identifie à x jusqu'à τ^+, mais la prolonge strictement au delà de τ^+, ce qui est absurde. □

21.4 Dépendance par rapport aux conditions initiales

Proposition 21.11. *Soit U un ouvert de l'espace de Banach E, I un intervalle de $\mathbb{R}$, et f une fonction continue de $U \times I$ dans $\mathbb{R}$, Lispschitzienne par rapport à la première variable. Pour x_0, y_0 donnés dans U, on note x et y les solutions au problèmes de Cauchy associées à ces conditions initiales au temps $t_0 \in I$. Alors sur leur intervalle de définition, on a*

$$\|y(t) - x(t)\| \leq e^{k(t-t_0)} \|y_0 - x_0\|\,.$$

Démonstration: On a

$$\|y(t) - x(t)\| = \left\| y_0 - x_0 + \int_{t_0}^{t} (f(y(s), s) - f(x(s), s)) \right\| \leq \|y_0 - x_0\| + k \int_{t_0}^{t} \|y(s) - x(s)\|$$

Le lemme de Gronwall 21.1 assure l'inégalité anoncée. □

On se place ici dans l'espace euclidien $\mathbb{R}^N$.

Proposition 21.12. *Soit $f : \mathbb{R}^N \times I \longrightarrow \mathbb{R}$ vérifiant les hypothèses du théorème de Cauchy Lipschitz. On suppose qu'il existe deux constantes A et B telles que*

$$|f(x,t)| \leq A\,|x| + B \qquad \text{sur } \mathbb{R}^N \times I.$$

Alors toute solution au problème de Cauchy est définie sur I tout entier.

Démonstration: D'après la proposition 21.10, les solutions maximales ne sont définies sur un sous-intervalle strict que si $|x|$ tend vers $+\infty$. Or (on considère ici $t > t_0$ pour simplifier)

$$\|x(t)\| \leq \|x_0\| + B(t - t_0) + A \int_{t_0}^{t} \|x(s)\|$$

D'après le lemme de Gronwall 21.1 appliqué à $\varphi(t) = \|x(t_0 + t)\|$, on ne peut donc avoir divergence de $|x|$ vers $+\infty$ en temps fini. □

21.5 Points fixes, stabilité

Definition 21.13. *(Stabilité, stabilité asymptotique)*
Soit $t \mapsto x(t)$ une solution du problème de Cauchy (21.1) associé à (x_0, t_0), que l'on suppose définie sur $[t_0, +\infty[$. On dit que la solution x est

(i) stable *si pour tout $\varepsilon > 0$, il existe $\eta > 0$ tel que, pour tout y_0 tel que $\|y_0 - x_0\| < \eta$, la trajectoire $t \mapsto y(t)$ associée à la condition initiale y_0 reste à distance de $x(t)$ inférieure à ε ;*

(ii) asymptotiquement stable *si (i) est vérifié, et que de plus $\|y(t) - x(t)\|$ tend vers 0 quand t tend vers $+\infty$.*

Remarque 21.14. *On s'intéressera souvent au cas de systèmes autonomes, i.e. tels que f ne dépend pas du temps, et pour des trajectoires stationnaires correspondant à des x_0 qui annulent f. Dans ce cas on parle de point d'équilibre stable (ou asymptotiquement stable) selon la la terminologie introduite ci-dessus, avec une trajectoire stationnaire $x(t) \equiv x_0$.*

Le théorème suivant donne une condition suffisante de stabilité asymptotique, ainsi qu'une condition suffisante de non stabilité, pour un point d'équilibre dans le cas autonome dans $\mathbb{R}^N$.

Théorème 21.15. *On se place dans $\mathbb{R}^N$. Soit x_0 un point fixe de l'équation $\dot{x} = f(x)$. On suppose f continûment différentiable dans un voisinage de x_0, et l'on introduit le gradient*

$$\nabla f = \left(\frac{\partial f_i}{\partial x_j} \right)_{1 \le i,j \le N}$$

1. *Si toutes les valeurs propres de ∇f sont de parties réelles strictement négatives, alors le point x_0 est asymptotiquement stable.*
2. *Si l'une (au moins) des valeurs propres a une partie réelle strictement positive, alors x_0 n'est pas stable.*

Exemple 21.1. Dans le cas où les parties réelles des valeurs propres sont nulles, tous les cas peuvent se produire, comme l'illustre la situation suivante. On considère le flot dans $\mathbb{R}^2$ associé à

$$f(x) = \begin{pmatrix} -x_2 + \alpha |x|^2 x_1 \\ x_1 + \alpha |x|^2 x_2 \end{pmatrix}$$

Notons en premier lieu que pour tout α réel, le gradient de f a des valeurs propres imaginaires pures (i et $-i$). Dans le cas $\alpha = 0$, le point fixe $x_0 = 0$ est stable (mais non asymptotiquement stable). Pour $\alpha > 0$, le point est instable, et pour $\alpha < 0$, le point est asymptotiquement stable.

Proposition 21.16. *Soit φ une fonction C^1 de $\mathbb{R}^N$ dans $\mathbb{R}$. On note $W = \{x\,,\ \varphi(x) \le 0\}$, et l'on considère une fonction f définie sur $U \times \mathbb{R}$, qui vérifie les hypothèses du théorème de Cauchy Lipschitz, avec $W \subset U$. Si*

$$\nabla\varphi \cdot f(x,t) < 0 \quad \forall t\,,\ x \in \varphi^{-1}(0),$$

alors les trajectoires à droite (vers les temps positifs) du problème de Cauchy-Lipschitz associées aux données (x_0, t_0) avec $x_0 \in W$ sont dans W.

Corollaire 21.17. *Dans les hypothèses de la proposition précédentes, si l'on suppose de plus W compact, la solution est définie sur tout $[t_0, +\infty[$.*

Definition 21.18. *(Fonction de Lyapunov)*
On considère un point d'équilibre de l'équation autonome $\dot{x} = f(x)$ dans $\mathbb{R}^N$, c'est-à-dire un point x_0 tel que $f(x_0) = 0$. On appelle fonction de Lyapunov pour x_0 une fonction φ continue sur un voisinage V de x_0, continûment différentiable sur $V \setminus \{x_0\}$, et telle que

1. *x_0 est un minimum strict de φ sur V,*
2. *$\nabla\varphi(x) \cdot f(x) \leq 0$ pour tout $x \in V \setminus \{x_0\}$,*

Proposition 21.19. *Si le point fixe x_0 admet une fonctionnelle de Lyapunov, alors il est stable. Si la fonctionnelle peut être choisie de telle sorte que l'inégalité (ii) est stricte (pour $x \neq x_0$), alors x_0 est asymptotiquement stable.*

Démonstration: Soit $\varepsilon > 0$, suffisamment petit pour que $\overline{B}(x_0, \varepsilon)$ soit dans V. Le minimum de φ sur la sphère est atteint, il est donc strictement plus grand que la valeur en x_0. On choisit β compris strictement entre ces deux valeurs, et l'on introduit

$$W = \varphi^{-1}(]-\infty, \beta[) \cap B(x_0, \varepsilon).$$

C'est un ouvert qui contient x_0, il contient donc une boule $B(x_0, \eta)$. Pour toute condition initiale dans cette boule, la trajectoire reste dans $B(x_0, \varepsilon)$, car $\varphi(x(t))$ est décroissant, donc reste inférieur à β, donc ne peut s'approcher de la frontière de $B(x_0, \varepsilon)$.

On suppose maintenant l'inégalité est stricte. On considère une trajectoire $t \mapsto y(t)$ issue de $y(0) \in B(x_0, \eta)$. Comme $\varphi(y(t))$ est décroissante, elle converge vers une limite ℓ. Si ℓ est le minimum de φ sur V, alors toute valeur d'adhérence x de la trajectoire vérifie $\varphi(x) = \ell$, d'où $x = x_0$, et on a convergence de la trajectoire (qui est incluse dans le compact $\overline{B}(x_0, \varepsilon)$) vers x_0. Si la limite est strictement supérieure à ce minimum, on considère l'ensemble

$$A = \varphi^{-1}([\beta, +\infty[) \cap \overline{B}(x_0, \varepsilon).$$

Cet ensemble est compact car fermé borné. La fonction

$$x \longmapsto \nabla\varphi(x) \cdot f(x)$$

y atteint donc son maximum, qui est strictement négatif d'après l'hypothèse :

$$\nabla\varphi(x) \cdot f(x) \leq \alpha < 0 \quad \forall x \in A.$$

La trajectoire considérée étant incluse dans A, on a

$$\frac{d}{dt}\varphi(y(t)) = \nabla\varphi(y(t)) \cdot f(y(t)) \leq \alpha < 0,$$

d'où l'on déduit que $\varphi(y(t))$ tend vers $-\infty$, ce qui est absurde. □

21.6 Compléments

Definition 21.20. *(Flot d'une équation différentielle)*
On considère l'équation différentielle (21.1), sous les hypothèses du théorème (21.9). On appelle flot de l'équation différentielle l'application Φ qui au triplet $(x_0, t_0; t)$ associe la solution au temps t du problème de Cauchy pour la donnée (x_0, t_0). Cette application vérifie donc

$$\left\{ \begin{array}{rcl} \dfrac{\partial\Phi}{\partial t}(x_0, t_0; t) & = & f(\Phi(x_0, t_0; t), t), \\ \Phi(x_0, t_0; t_0) & = & x_0. \end{array} \right. \tag{21.2}$$

Cette application est définie sur

$$\bigcup_{(x_0,t_0)\in U\times I} \{(x_0,t_0)\} \times I_{(x_0,t_0)}$$

où $I_{(x_0,t_0)}$ est l'intervalle de définition de la solution maximale associée à la donnée de Cauchy (x_0,t_0).

Proposition 21.21. *On se place dans le cadre de la définition précédente, en supposant de plus que la fonction f est globalement Lipschitizienne par rapport à la première variable sur $U\times I$, de constante de Lipschitz k. Alors*

$$\|\Phi(y_0,t_0;t)-\Phi(x_0,t_0;t)\| \leq e^{k(t-t_0)}\,\|y_0-x_0\|\,.$$

Démonstration. C'est une application directe de la proposition (21.11). □

22 Espaces de Sobolev

22.1 Rappels sur l'espace $L^2(\Omega)$

On désigne par Ω un ouvert de $\mathbb{R}^N$ muni de la mesure de Lebesgue dx.

Definition 22.1. *On définit l'espace $L^2(\Omega)$ comme*

$$L^2(\Omega) = \left\{ f : \Omega \to \mathbb{R}\,,\ f \text{ mesurable},\ \int_\Omega |f(x)|^2\, dx < +\infty \right\}.$$

On le munit de la norme $\|f\|_2 = \left(\int_\Omega |f|^2\right)^{1/2}$. On notera $L^2(\Omega)^N$ l'espace des champs de vecteurs dont chaque composante appartient à $L^2(\Omega)$.

Proposition 22.2. *L'espace $L^2(\Omega)$ est un espace de Hilbert pour le produit scalaire*

$$(u, v) = \int_\Omega u(x)v(x)\, dx,$$

comme pour tout produit du type

$$(u, v)_k = \int_\Omega k(x)u(x)v(x)\, dx,$$

où k est une fonction mesurable telle que $0 < m \leq k(x) \leq M$ presque partout.

Démonstration: Le fait que cette forme bilinéaire soit bien définie sur $L^2 \times L^2$ est conséquence directe de l'inégalité de Cauchy-Schwarz. Il s'agit alors de montrer que L^2 est bien complet pour la norme associée. Pour cela on considère une suite de Cauchy, on montre par un argument de convergence monotone que la suite converge presque partout vers une limite, que la limite appartient bien à L^2, et que l'on a bien convergence pour la norme L^2 vers cette limite. On trouvera une démonstration détaillée dans [2], page 57.

Definition 22.3. *(Suite régularisante)*
On appelle suite régularisante une suite (ρ_n) de fontions C^∞ de $\mathbb{R}^N$ dans $\mathbb{R}$ telle que, pour tout $n \in \mathbb{N}$,

$$\mathrm{supp}(\rho_n) \subset B(0, 1/n)\,,\ \int_{\mathbb{R}} \rho_n = 1\,,\ \rho_n(x) \geq 0 \quad \forall x \in \mathbb{R}^N.$$

Proposition 22.4. *Soit $f \in L^2(\mathbb{R}^N)$. On définit la fonction $\rho_n \star f$ par*

$$(\rho_n \star f)(x) = \int_{\mathbb{R}^N} \rho_n(x - y)f(y)\, dy.$$

Alors la fonction $\rho_n \star f$ est dans $C^\infty(\mathbb{R}^N) \cap L^2(\mathbb{R}^N)$. On a

$$\rho_n \star f \longrightarrow f \text{ dans } L^2(\mathbb{R}^N).$$

Remarque 22.5. *Toute fonction f de $L^2(\Omega)$ peut être prolongée par 0 à $\mathbb{R}^N$ tout entier. On peut donc appliquer ce qui précède. Les propriétés de convergence énoncées ci-dessus s'appliquent ainsi à la restriction de $\rho_n \star f$ à Ω.*

Definition 22.6. *On note $\mathcal{D}(\Omega)$ l'espace des fonctions $\mathcal{C}^\infty$ à support compact dans Ω. On vérifie que cet espace est non vide en considérant une boule ouverte $B(a,r)$ dont l'adhérence est dans Ω, et la fonction*

$$\varphi(x) = \exp\left(\frac{1}{|x-a|^2 - r^2}\right) \text{ si } x \in B(a,r)\,,\ \varphi(x) = 0 \text{ si } x \notin B(a,r).$$

Proposition 22.7. *L'espace $\mathcal{D}(\Omega)$ est dense dans $L^2(\Omega)$.*

Remarque 22.8. *L'appartenance à L^2 n'exige aucune régularité en espace (aucune "corrélation spatiale" n'est exigée). En particulier, si l'on considère une partition de Ω sous la forme $\overline{\Omega} = \overline{\Omega}_1 \cup \overline{\Omega}_2$, $\Omega_1 \cap \Omega_2 = \emptyset$, où les Ω_i sont des ouverts tels que $\partial\Omega_1 \cap \partial\Omega_2$ est de mesure nulle, pour toutes fonctions $f_i \in L^2(\Omega_i)$, la fonction f dont la restriction à Ω_i est f_i est dans $L^2(\Omega)$. Nous verrons qu'une telle construction par morceaux d'une fonction est en général impossible pour les espaces de Sobolev.*

22.2 Définitions, propriétés générales

Definition 22.9. *(Gradient)*
Soit φ une fonction C^1 de Ω dans $\mathbb{R}$. On appelle gradient de φ la fonction de Ω dans $\mathbb{R}^N$ définie par

$$\nabla\varphi = \begin{pmatrix} \dfrac{\partial\varphi}{\partial x_1} \\ \vdots \\ \dfrac{\partial\varphi}{\partial x_N} \end{pmatrix}.$$

Definition 22.10. *On définit l'espace de Sobolev $H^1(\Omega)$ comme l'ensemble des fonctions u dans $L^2(\Omega)$ telles qu'il existe $v = (v_1, \ldots, v_N) \in (L^2(\Omega))^N$ vérifiant*

$$\int_\Omega u \frac{\partial\varphi}{\partial x_i} = -\int_\Omega \varphi\, v_i \quad \forall\varphi \in \mathcal{D}(\Omega)\,,\ \forall i = 1, \ldots, N.$$

On notera alors $v = \nabla u$.

La fonction ∇u de $\mathbb{R}$ dans $\mathbb{R}^N$ est ainsi définie comme l'unique fonction vectorielle à composantes dans $L^2(\Omega)$ telle que l'identité entre vecteurs de $\mathbb{R}^N$

$$\int_\Omega u\nabla\varphi = -\int_\Omega \varphi\nabla u$$

soit vérifiée pour tout $\varphi \in \mathcal{D}(\Omega)$.

On notera $H^1(\Omega)^N$ l'espace des champs de vecteurs dont chaque composante appartient à $H^1(\Omega)$. Le gradient ∇u est alors une matrice dont la ligne i est le gradient de la i-ème composante de u.

Proposition 22.11. *L'espace $H^1(\Omega)$ muni de la norme $\|\cdot\|$ définie par*

$$\|v\|^2 = \int_\Omega u^2 + \int_\Omega |\nabla u|^2$$

est un espace de Hilbert séparable[111]*. .*

111. Il contient un sous-ensemble dénombrable et dense

Démonstration: On construit pour cela une isométrie entre $H^1(\Omega)$ et un sous-espace fermé de $L^2(\Omega) \times L^2(\Omega)^N$. Voir [2, Prop. IX.1]. □

Notation: On désignera par $|u|_{0,\Omega}$ la norme L^2 de u sur Ω (nous omettrons Ω quand il n'y a pas d'ambigüité), et par $|u|_{1,\Omega}$ la semi norme H^1 :

$$|u|_{1,\Omega}^2 = \int_\Omega |\nabla u|^2 \,,$$

de telle sorte que

$$\|u\|_{H^1}^2 = |u|_{0,\Omega}^2 + |u|_{1,\Omega}^2 \,.$$

Proposition 22.12. *Si $u \in C^1(\Omega) \cap L^2(\Omega)$ et $\nabla u \in (L^2(\Omega))^N$, alors $u \in H^1(\Omega)$, et le gradient de u au sens classique (définition 22.9) s'identifie au gradient au sens de Sobolev (définition 22.10).*

Proposition 22.13. *Soit $u \in H^1(\Omega)$ telle que $\nabla u = 0$ presque partout sur Ω. Alors u est constante sur chaque composante connexe de Ω.*

En dimension 1, une fonction peut s'écrire comme intégrale de sa dérivée, comme le précise la proposition suivante.

Proposition 22.14. *Soit I un intervalle de $\mathbb{R}$. Toute fonction $u \in H^1(I)$ admet un représentant continu $\tilde{u}$, qui vérifie*

$$\tilde{u}(x) = u(x) \quad \text{p.p. sur } I \ , \ \tilde{u}(y) - \tilde{u}(x) = \int_x^y u'(t)\, dt.$$

Cette fonction continue sur I est prolongeable par continuité aux extrémités de I.

Démonstration: Voir Brezis [2, Th. VIII.2]. □

Proposition 22.15. *Soit u une fonction de $L^2(\Omega)$. Les assertions suivantes sont équivalentes :*

(i) $u \in H^1(\Omega)$.

(ii) Il existe une constante C telle que

$$\left| \int_\Omega u \nabla \varphi \right| \leq C \, \|\varphi\|_{L^2} \quad \forall \varphi \in \mathcal{D}(\Omega).$$

(iii) Il existe une constante C telle que, pour tout $\omega \subset\subset \Omega$, pour tout h tel que $|h| < \operatorname{dist}(\omega, \Omega^c)$,

$$\|\tau_h u - u\|_{L^2(\omega)} \leq C \, |h| \,.$$

Démonstration: $(i) \Longrightarrow (ii)$ est une conséquence immédiate de la définition.

$(ii) \Longrightarrow (i)$ Pour i entre 1 et N, on considère la forme linéaire définie sur $C_c^\infty \subset L^2(\Omega)$

$$\varphi \longmapsto \int_\Omega v \partial_{x_i} \varphi.$$

Cette forme linéaire est continue pour la norme L^2 par hypothèse. Elle se prolonge donc par densité de $C_c^\infty(\Omega)$ en une forme linéaire continue sur $L^2(\Omega)$. Le théorème de représentation de Riesz-Fréchet assure donc l'existence de $w_i \in L^2(\Omega)$ tel que

$$\int_\Omega v\partial_{x_i}\varphi = -\int_\Omega w_i\varphi,$$

d'où $u \in H^1$ avec $\nabla u = (w_1, \dots, w_N)$. □

$(i) \Longrightarrow (iii)$ Soit $\omega \subset\subset \Omega$, et $h < \operatorname{dist}(\omega, \Omega^c)$. On considère dans un premier temps une fonction u régulière ($u \in \mathcal{D}(\Omega)$). On a

$$u(x+h) = u(x) + \int_0^1 \nabla u(x+th)\cdot h\, dt,$$

d'où

$$|u(x+h) - u(x)|^2 \le |h|^2 \int_0^1 |\nabla u(x+th)|^2,$$

et donc

$$\int_\omega |\tau_h u - u(x)|^2 \le |h|^2 \int_\omega \int_0^1 |\nabla u(x+th)|^2 \le |h|^2 \int_0^1 \int_\omega |\nabla u(x+th)|^2.$$

On choisit maintenant ω' fortement inclus dans Ω, qui contient tous les translatés de ω par th, pour $t \in [0,1]$. On a

$$\|\tau_h u - u\|_{L^2} \le |h| \int_{\omega'} |\nabla u|^2.$$

On conclut en utilisant la propriété de densité 22.17.

$(iii) \Longrightarrow (ii)$ Soit $\varphi \in C_c^\infty(\Omega)$, et $\omega \subset\subset \Omega$ qui contient le support de φ. Pour tout h tel que $h < \operatorname{dist}(\omega, \Omega^c)$, on a

$$\left|\int_\omega (\tau_h u - u)\varphi\right| \le C\,\|\varphi\|_{L^2(\omega)}\,|h| \le C\,\|\varphi\|_{L^2(\Omega)}\,|h|.$$

D'autre part,

$$\int_\omega (u(x+h) - u(x))\varphi(x) = \int_\Omega (u(x+h) - u(x))\varphi(x) = \int_\Omega u(y)(\varphi(y-h) - \varphi(y)).$$

La majoration (iii) implique donc

$$\left|\int_\Omega u(y)\frac{\varphi(y-h) - \varphi(y)}{|h|}\right| \le C\,\|\varphi\|_{L^2}.$$

On conclut en prenant h de la forme $t\vec{e}_i$ et en faisant tendre t vers 0. □

Proposition 22.16. *L'espace $\mathcal{D}(\mathbb{R}^N)$ est dense dans $H^1(\mathbb{R}^N)$.*

Notation: On dit que ω est fortement inclus dans Ω si $\overline{\omega}$ est compact et inclus dans Ω. On note $\omega \subset\subset \Omega$.

Proposition 22.17. *Pour tout $\omega \subset\subset \Omega$, tout $u \in H^1(\Omega)$, il existe une suite (u_n) dans $\mathcal{D}(\Omega)$ telle que*

$$u_n \longrightarrow u \text{ dans } L^2(\Omega)\,,\ \nabla u_n \longrightarrow \nabla u \text{ dans } L^2(\omega)^N.$$

Corollaire 22.18. *Soit (ω_n) une suite de domaines fortements inclus dans Ω, et $u \in H^1(\Omega)$. Il existe une suite (u_n) dans $\mathcal{D}(\Omega)$ telle que*

$$\|u_n - u\|_{L^2(\Omega)} \longrightarrow 0\,,\ \|\nabla u_n - \nabla u\|_{L^2(\omega_n)^N} \longrightarrow 0.$$

Definition 22.19. *On définit $H_0^1(\Omega)$ comme l'adhérence de $\mathcal{D}(\Omega)$ dans $H^1(\Omega)$.*

Noter que, d'après la proposition 22.17, on a $H_0^1(\mathbb{R}^N) = H^1(\mathbb{R}^N)$

Par rapport à H_0^1, l'espace H^1 peut se décrire comme l'ensemble des fonctions L^2 de gradient L^2 qui peuvent "prendre des valeurs non nulles sur le bord". Cette expression ne pourra se voir donner un cadre mathématique précis qu'après que l'on aura défini la notion de régularité du bord (voir, section 22.3, la définition de l'opérateur trace sur le bord γ_0). On peut néammoins dès maintenant donner un sens abstrait à la notion de valeur au bord, sans faire aucune hypothèse sur la géométrie de Ω. Par analogie avec l'espace des traces des fonctions de H^1 dans le cas d'un bord régulier (voir définition 22.31), nous noterons $\tilde{H}^{1/2}$ l'espace abstrait correspondant.

Definition 22.20. *On définit l'espace $H^2(\Omega)$ comme l'ensemble des fonctions de $H^1(\Omega)$ dont toutes les dérivées partielles par rapport à l'une des composantes sont elles-mêmes dans $H^1(\Omega)$. C'est un espace de Hilbert muni de la norme*

$$\|u\|^2_{H^2(\Omega)} = |u|_0^2 + \sum_i \left|\frac{\partial u}{\partial x_i}\right|_0^2 + \sum_{i,j} \left|\frac{\partial^2 u}{\partial x_i \partial x_j}\right|_0^2 = |u|^2_{0,\Omega} + |u|^2_{1,\Omega} + |u|^2_{2,\Omega}\,.$$

On peut définir de façon analogue les espaces $H^m(\Omega)$ pour $m = 3,\ 4,\ \ldots$, mais nous n'utiliserons ici que $m \leq 2$.

Definition 22.21. *(Espace H^m_{loc})*
Soit m un entier positif (on utilisera le cas $m = 2$ dans la suite). On définit l'espace $H^m_{loc}(\Omega)$ comme l'espace vectoriel des (classes de) fonctions de Ω dans $\mathbb{R}$ dont la restriction à ω est dans $H^m(\omega)$, pour tout ω fortement inclus dans Ω. De façon équivalente, c'est l'ensemble des fonctions u de Ω dans $\mathbb{R}$ telles que θu est dans $H^m(\Omega)$ pour tout θ dans $\mathcal{D}(\Omega)$.

Noter que l'appartenance d'une fonction à H^m_{loc} permet de parler de ses dérivées m-ièmes comme de fonctions (mesurables) définies sur Ω. On donne ainsi un sens à des expressions du type $\partial^m u/\partial x_i^m = g$ presque partout dans Ω, où g est une fonction de L^2_{loc}.

22.3 Traces

En élasticité, le problème le plus couramment rencontré consiste à trouver le champ de déplacement d'un solide déformable soumis à certaines sollicitations sur son bord (déplacement imposé). Ces sollicitations au bord ne peuvent avoir un sens que si l'on est capable de parler d'un champ de déplacement sur le bord du domaine. Lorsque l'on considère des fonctions régulières (au moins continues sur $\overline{\Omega}$), on peut parler simplement de la restriction de la

fonction à $\partial\Omega$. Dans le contexte présent, nous avons vu que les fonctions de $H^1(\Omega)$ ne sont pas nécessairement continues, et ne sont définies a priori que comme des classes de fonctions (à un ensemble de mesure nulle près). La frontière d'un ouvert régulier étant de mesure nulle, la notion de restriction n'a pas de sens. Nous allons montrer ici qu'il est possible de donner un sens précis à cette notion de trace, dès que les fonctions que l'on considère ont une régularité suffisante en espace.

Definition 22.22. *(Espace des traces abstrait)*
On définit l'espace $\tilde{H}^{1/2}$ comme l'espace quotient $H^1(\Omega)/H_0^1(\Omega)$. C'est un espace vectoriel normé pour la norme quotient

$$\|\tilde{u}\|_{H^1/H_0^1} = \inf_{v\in\tilde{u}} \|v\|_{H^1} = \inf_{h\in H_0^1} \|u-h\|_{H^1}\,.$$

Noter que, d'après la définition de H_0^1, on a aussi $\|\tilde{u}\|_{H^1/H_0^1} = \inf_{h\in\mathcal{D}(\Omega)} \|u-h\|_{H^1}$.

Remarque 22.23. *On a $H_0^1(\mathbb{R}^N) = H^1(\mathbb{R}^N)$ (d'après la proposition 22.16), et l'on peut avoir $H_0^1(\Omega) = H^1(\Omega)$ même si Ω est strictement inclus dans $\mathbb{R}^N$ (de telle sorte que $\mathcal{D}(\Omega)$ soit strictement inclus dans $\mathcal{D}(\mathbb{R}^N)$). L'espace quotient défini précédemment est alors l'espace trivial $\{0\}$. C'est le cas par exemple de $\mathbb{R}^2$ privé d'un point, ou de $\mathbb{R}^3$ privé d'un point ou d'une droite (voir l'exercice 22.1 ci-après sur la notion de* capacité*).*

Exercice 22.1. (Impossibilité de définir la valeur ponctuelle d'un champ)
Soient Ω et ω deux domaines réguliers, avec $\omega\subset\Omega$. On définit la capacité de ω vis-à-vis de Ω (on dira simplement capacité s'il n'y a pas d'ambigüité) la quantité

$$C_\omega = \inf\left\{\int_\Omega |\nabla u|^2\ ,\ v_{|\omega}\equiv 1 \text{ sur } \omega\,,\ v\in D(\Omega)\right\}.$$

1) Calculer la capacité C_r^R d'une boule de rayon r vis-à-vis d'une boule de rayon R, dans $\mathbb{R}^n$ pour $n=1$, $n=2$, et $n=3$.

2) Préciser la limite de cette capacité lorsque le rayon intérieur r tend vers 0, à $R>0$ fixé.

3) En déduire qu'en dimension 2 ou 3 la notion de valeur ponctuelle d'un champ de $H^1(\Omega)$ n'a pas de signification. On pourra montrer par exemple que le sous-espace des fonctions régulières qui prennent la valeur 1 en un point intérieur à Ω est dense dans $H^1(\Omega)$.

Proposition 22.24. *Soit $u\in H_0^1(\Omega)$. On définit $\tilde{u}$ comme la fonction qui vaut $u(x)$ pour tout $x\in\Omega$, et qui prend la valeur 0 à l'extérieur de Ω. Alors $\tilde{u}\in H^1(\mathbb{R}^N)$.*

Démonstration: Tout d'abord remarquons que $\tilde{u}$ est dans $L^2(\mathbb{R}^N)$. Par définition de H_0^1, u est limite d'une suite (u_n) de fonctions C^∞ à support compact dans Ω. Pour tout $\varphi\in\mathcal{D}(\mathbb{R}^N)$, on a

$$\begin{aligned}\int_{\mathbb{R}^N}\tilde{u}\nabla\varphi &= \int_\Omega u\nabla\varphi = \lim_{n\to+\infty}\int_\Omega u_n\nabla\varphi\\ &= -\lim_{n\to+\infty}\int_\Omega\varphi\nabla u_n = -\int_\Omega\varphi\nabla u = -\int_{\mathbb{R}^N}\varphi v.\end{aligned}$$

où v est le champ de vecteurs qui vaut ∇u dans Ω, et 0 à l'extérieur de Ω. □

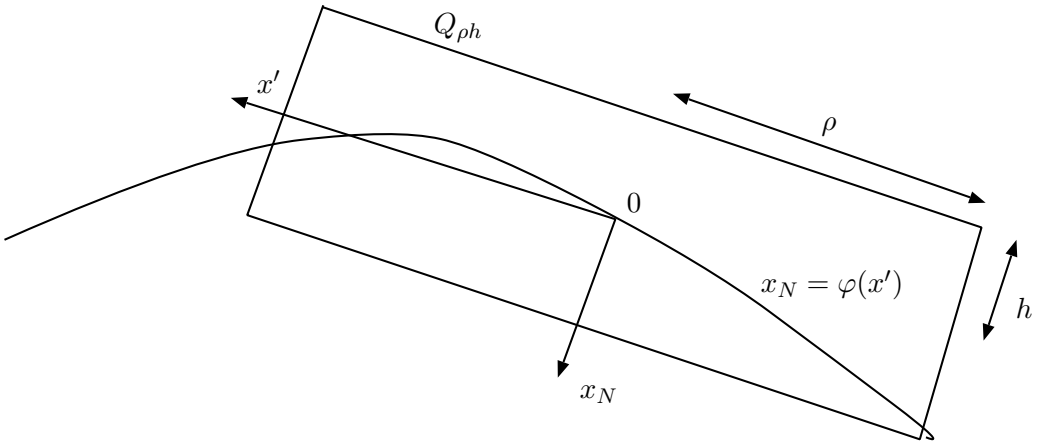


FIGURE 22.1 – Régularité de la frontière

Dans cette section nous précisons les propriétés qui vont nous permettre de définir des valeurs au bord pour des fonctions appartenant aux espaces de Sobolev introduits précédemment. On se reportera à [6] ou [2] pour les démonstrations détaillées.

On définit le cylindre $Q_{\rho h}$ de $\mathbb{R}^N$ par

$$Q_{\rho h} = \left\{ x \in \mathbb{R}^N \,,\ x = (x', x_N) = (x_1, \ldots, x_N) \,,\ |x'| < \rho \,,\ -h < x_N < h \right\}.$$

Dans la définition qui suit, "X" représente une régularité fonctionnelle du type C^o, Lipschitz, C^k, etc...

Definition 22.25. *Soit Ω un ouvert de $\mathbb{R}^N$. On dit que la frontière de Ω est de classe X si en tout point $a \in \partial\Omega$, il existe un système de coordonnées et ρ, $h > 0$, tels qu'il existe une application*

$$\varphi \ :\ \left\{ x' \in \mathbb{R}^{N-1} \,,\ |x'| < \rho \right\} \longrightarrow \mathbb{R}$$

de classe X telle que

(i) $\forall x'$, $|x'| < \rho \Rightarrow |\varphi(x')| < h$,

(ii) $\varphi(0) = 0$,

(iii) $Q_{\rho h} \cap \partial\Omega$ *coïncide avec le graphe de* φ,

(iv) $U \cap \Omega = \{(x', x_N) \,,\ |x'| \leq \rho \,,\ \varphi(x') < x_N < h\}$.

Definition 22.26. *(vecteur normal)*
Soit Ω un ouvert de classe $\mathcal{C}^1$, a un point de $\Gamma = \partial\Omega$. On note φ l'application définie ci-dessus. On appelle vecteur normal à Γ au point a le vecteur

$$n = \frac{(\nabla\varphi, -1)}{|(\nabla\varphi, -1)|}.$$

Noter que l'on peut définir presque partout un tel vecteur sur une frontière supposée seulement Lipschitzienne.

On note $\mathcal{D}(\overline{\Omega})$ l'ensemble des restrictions des fonctions de $\mathcal{D}(\mathbb{R}^N)$ à $\overline{\Omega}$.

Proposition 22.27. *Soit Ω un ouvert de frontière Γ Lipschitzienne et bornée. Il existe un opérateur de prolongement*

$$P \ : \ H^1(\Omega) \longrightarrow H^1(\mathbb{R}^N),$$

linéaire continu, tel que, pour tout $u \in H^1(\Omega)$, la restriction de Pu à Ω s'identifie à u.

Démonstration: Voir Brezis [2, Th. IX.7] dans le cas d'un ouvert C^1. L'ingrédient principal de la démonstration est le prolongement par réflexion dont nous indiquons ici le principe dans le cas $N = 1$. On considère $u \in H^1(]0,1[)$, et l'on construit $\tilde{u}$ comme la fonction qui s'identifie à u sur $]0,1[$, et telle que $\tilde{u}(x) = u(-x)$ sur $]-1,0[$. La fonction $\tilde{u}$ est dans $L^2(]-1,1[)$, et sa dérivée $\tilde{u}'$ est définie presque partout sur $]-1,1[$ (avec $\tilde{u}'(-x) = -u'(x)$ pour $x > 0$. Nous allons montrer que cette fonction $\tilde{u}'$ est bien la dérivée de u au sens de Sobolev sur $]-1,1[$. Pour toute fonction-test $\varphi \in \mathcal{D}(]-1,1[)$, si l'on note $\tilde{\varphi}(x) = \varphi(-x)$, on a

$$\int_{-1}^{1} u\varphi' = \int_{-1}^{0} u\varphi' + \int_{0}^{1} u\varphi' = -\int_{0}^{1} u\tilde{\varphi}' + \int_{0}^{1} u\varphi' = \int_{0}^{1} u(\varphi - \tilde{\varphi})'.$$

Notons $\psi = \varphi - \tilde{\varphi}$. On ne peut pas utiliser l'appartenance de u à $H^1(]0,1[)$ car ψ n'est pas à support compact dans $]0,1[$. On se ramène à une fonction à support compact en introduisant, pour $\varepsilon > 0$, la fonction $x \longmapsto \eta_\varepsilon(x) = \eta(x/\varepsilon)$, où η est une fonction C^∞ sur $\mathbb{R}^+$, nulle sur $[0,1/2]$ et sur $[1,+\infty[$. La fonction $\psi_\varepsilon = \eta_\varepsilon\psi$ est dans $\mathcal{D}(]0,1[)$. On a d'une part

$$\int_0^1 u\psi_\varepsilon' = -\int_0^1 \psi_\varepsilon u' \longrightarrow -\int_0^1 \psi u' = -\int_{-1}^1 \varphi\tilde{u}',$$

et d'autre part

$$\int_0^1 u\psi_\varepsilon' = \int_0^1 \eta_\varepsilon\psi' u + \int_0^1 \eta_\varepsilon'\psi u.$$

Le second terme se majore (en utilisant $\psi(x) = \mathcal{O}(x)$ et $|\eta_\varepsilon'| \leq C/\varepsilon$),

$$\left|\int_0^1 \eta_\varepsilon'\psi u\right| = \left|\int_0^\varepsilon \eta_\varepsilon'\psi u\right| \leq C\varepsilon\frac{1}{\varepsilon}\int_0^\varepsilon |u| \leq C\sqrt{\varepsilon}.$$

d'où $\int_0^1 u\psi_\varepsilon' \longrightarrow \int_0^1 \psi' u$,

On a donc $\tilde{u} \in H^1(]-1,1[)$. □

Proposition 22.28. *Soit Ω un ouvert de frontière Γ Lipschitzienne. Alors $\mathcal{D}(\overline{\Omega})$ est dense dans $H^1(\Omega)$.*

Proposition 22.29. *Soit Ω un ouvert de frontière Γ Lipschitzienne et bornée. L'application*

$$\gamma_0 : \varphi \in \mathcal{D}(\overline{\Omega}) \longmapsto \varphi_{|\Gamma},$$

se prolonge par continuité en une application linéaire de $H^1(\Omega)$ dans $L^2(\Gamma)$.

Démonstration: On se limite ici à une démonstration dans le cas du demi espace $\mathbb{R}^{N-1} \times \mathbb{R}^+$ (pour lequel le résultat est vrai malgré le caractère non borné), et l'on se reportera à [2] pour

une démonstration plus complète. On peut se limiter à des fonctions régulières nulles pour $x_N \geq 1$. Pour une telle fonction, on a

$$\varphi(x', 0) = \int_1^0 \partial_N \varphi,$$

d'où

$$\int_{\mathbb{R}^{N-1}} \varphi(x', 0)^2 = \int_{\mathbb{R}^N} \left(\int_1^0 \partial_N \varphi \right)^2 \leq \int_{\mathbb{R}^N} |\partial_N \varphi|^2 \leq \int_{\mathbb{R}^N} |\nabla u|^2 .$$

□

Remarque 22.30. *On notera que seul le contrôle sur la dérivée dans la direction verticale (normale à la frontière) a été utilisé dans la démonstration précédente. La rigidité transverse (selon $\mathbb{R}^{N-1}$ dans le cas précédent) va conditionner la régularité de la trace (dont on peut montrer qu'elle est strictement plus régulière que L^2).*

Definition 22.31. *(Espace $H^{1/2}(\Gamma)$)*
On note $H^{1/2}(\Gamma) \subset L^2(\Gamma)$ l'image de l'application $\gamma_0 : H^1(\Omega) \longmapsto L^2(\Gamma)$ définie ci-dessus. C'est un espace de Banach pour la norme

$$\|g\|_{H^{1/2}(\Gamma)} = \inf_{\gamma_0 v = g} \|v\|_{H^1(\Omega)} .$$

Remarque 22.32. *L'espace $H^{1/2}$ peut se définir sur l'espace entier par la transformée de Fourier (voir définition **??**), puis par cartes locales sur une variété régulière. Il est essentiel de garder à l'esprit que l'inclusion de $H^{1/2}$ est stricte. En particulier, l'appartenance à $H^{1/2}$ exclut les discontinuités franches (voir remarque 22.32, page 226).*

Proposition 22.33. *L'espace $H_0^1(\Omega)$ est constitué des fonctions de $H^1(\Omega)$ dont la trace sur $\partial\Omega$ est nulle.*

Démonstration: Voir Raviart [6]. □

Definition 22.34. *(Dérivée normale)*
Soit Ω un domaine de frontière Lipschitzienne. On note n le vecteur normal à Γ dirigé vers l'extérieur de Ω. Ce vecteur est défini presque partout. Pour toute fonction $\varphi \in \mathcal{D}(\overline{\Omega})$, on appelle dérivée normale de φ en un point de Γ la quantité

$$\frac{\partial \varphi}{\partial n} = \nabla \varphi \cdot n.$$

Definition 22.35. *Soit Ω un ouvert borné de frontière Γ lipschitzienne. On définit γ_1 comme l'application de $H^2(\Omega)$ dans $L^2(\Gamma)$ qui à $u \in H^2(\Omega)$ associe $\nabla u \cdot n$, où la trace de chaque composante de ∇u est définie comme précédemment. On notera*

$$\gamma_1 u = \frac{\partial u}{\partial n}.$$

Noter que l'on n'utilise pas ici la densité de $\mathcal{D}(\overline{\Omega})$ dans $H^2(\Omega)$ (qui, de fait, n'est pas exigée).

Proposition 22.36. *(Première formule de Green)*
Soit Ω un ouvert borné de frontière Γ Lipschitzienne. Pour tous u et v dans $H^1(\Omega)$, on a

$$\int_\Omega v\nabla u = -\int_\Omega u\nabla v + \int_\Gamma uvn.$$

Proposition 22.37. *(Deuxième formule de Green)*
Soit Ω un ouvert borné de frontière Γ lipschitzienne. Pour tout u dans $H^2(\Omega)$ et tout v dans $H^1(\Omega)$, on a

$$-\int_\Omega v\triangle u = \int_\Omega \nabla u\cdot\nabla v - \int_\Gamma \frac{\partial u}{\partial n}v.$$

Proposition 22.38. *Soit Ω un ouvert borné de frontière Γ lipschitzienne. On suppose que Ω se décompose de la façon suivante*

$$\overline{\Omega} = \bigcup_{i=1,\dots,p} \overline{\Omega}_i,$$

où les Ω_i sont des ouverts de frontière lipschitzienne, inclus dans Ω, deux à deux disjoints. On note $\Gamma_{ij} = \overline{\Omega}_i \cap \overline{\Omega}_j$. Soit u une fonction définie sur Ω, dont la restriction u_i à Ω_i est dans $H^1(\Omega_i)$ pour tout $i = 1,\dots,p$. On suppose que pour tous i, j tels que $\Gamma_{ij} \neq \emptyset$ les traces de u_i et u_j sur Γ_{ij} s'identifient. Alors u est dans $H^1(\Omega)$.

Démonstration: On note v la fonction de $L^2(\Omega)$ qui s'identifie à ∇u sur chacun des Ω_r. Pour tout $\varphi \in \mathcal{D}(\mathbb{R}^N)$, on a (en utilisant la proposition 22.36 sur chacun des Ω_r),

$$\begin{aligned}\int_\Omega v\varphi &= \sum_{i=1}^p \int_{\Omega_i} v\varphi \\ &= -\sum_{i=1}^p \int_{\Omega_i} u\nabla\varphi + \sum_{i,j}\int_{\Gamma_{ij}} u\varphi(n_i+n_j),\end{aligned}$$

où n_i (resp. n_j) est la normale à Γ_{ij} sortante au domaine Ω_i (resp. Ω_j), de telle sorte que $n_i + n_j = 0$. On a donc bien $u \in H^1(\Omega)$ avec $\nabla u = v$. □

Remarque 22.39. *On prendra garde au fait que (on reprend les notation du théorème précédent), même si u est dans $H^2(\Omega_i)$ pour tout i, le raccord des traces sur les interfaces ne suffit pas pour assurer l'appartenance de u à $H^2(\Omega)$. Cette remarque est à la base des difficultés que l'on peut avoir à approcher une fonction sur un maillage qui ne respecte pas la géométrie.*

Proposition 22.40. *On se replace dans le cadre des notations de la proposition précédente. Soit u une fonction définie sur Ω, dont la restriction u_i à Ω_i est dans $H^2(\Omega_i)$ pour tout $i = 1,\dots,R$. On suppose que pour tous i, j tels que $\Gamma_{ij} \neq \emptyset$ les traces de u_i et u_j sur Γ_{ij} s'identifient. On suppose d'autre part le raccord des dérivées normales : $\partial u_i/\partial n = \partial u_j/\partial n$ sur Γ_{ij}. Alors u est dans $H^2(\Omega)$.*

22.4 Injections

Théorème 22.41. *Soit Ω un domaine borné de frontière Lipschitzienne. Alors, pour tout entier $m > N/2$, $H^m(\Omega)$ s'injecte de façon continue dans $C^0(\overline{\Omega})$. En particulier les fonctions de $H^2(\Omega)$ sont continues pour les dimensions physiques $N = 1$, 2, ou 3.*

On retrouve notamment le fait déjà énoncé que les fonctions de $H^1(I)$, où I est un intervalle réel, sont continues. En revanche, le théorème ne s'applique pas à $H^1(\Omega)$ en dimension 2. Il existe effectivement des fonctions de $H^1(\mathbb{R}^2)$ qui ne sont pas continues.

On notera également qu'une fonction de $H^2(\Omega)$ est continue sur Ω, sans hypothèse de régularité, car tout $x \in \Omega$ est dans une boule incluse dans Ω. En l'absence de régularité du bord, il est en revanche possible que l'on n'ait pas $\|u\|_\infty \leq C \|u\|_{H^2}$.

Théorème 22.42. *(Rellich)*
Soit Ω un domaine borné de frontière Lipschitzienne. Alors l'injection de $H^1(\Omega)$ dans $L^2(\Omega)$ est compacte. L'injection de $H_0^1(\Omega)$ dans $L^2(\Omega)$ est compacte pour tout Ω borné (sans hypothèse de régularité). De même, l'injection de $H^{m+1}(\Omega)$ dans $H^m(\Omega)$ est compacte.

Démonstration: On se reportera à la section consacrée à la transformée de Fourier (voir théorème 22.64) pour une démonstration de ce théorème. On peut également démontrer la compacité de l'injection en utilisant le point (*iii*) de la caractérisation 22.15 de $H^1(\Omega)$, et le théorème de Riesz-Fréchet-Kolmogorov qui donne un critère suffisant de relative compacité pour des familles de fonctions de $L^2(\Omega)$ (voir Brezis [2, Th. IV.25 & Cor. IV.26]). □

Exercice 22.2. Montrer que l'injection de $H^1(\Omega)$ dans $L^2(\Omega)$ n'est jamais compacte quand Ω n'est pas borné.

22.5 Inégalités de Poincaré

Proposition 22.43. *(Inégalité de Poincaré)*
Soit Ω un domaine de $\mathbb{R}^N$ borné dans une direction, c'est-à-dire tel que

$$\Omega \subset \left\{ x \in \mathbb{R}^N \,,\ \xi \cdot x \in]a, b[\right\}.$$

Alors il existe une constante $C > 0$ telle que

$$\left(\int_\Omega |u|^2 \right)^{1/2} \leq C \left(\int_\Omega |\nabla u|^2 \right)^{1/2} \quad \forall u \in H_0^1(\Omega).$$

Démonstration: On note toujours u le prolongement par 0 de u sur $\mathbb{R}^N$ tout entier. Quitte à effectuer une translation et une rotation du système de coordonnées, on suppose que la bande qui contient Ω se met sous la forme

$$\left\{ x = (x_1, \ldots, x_N) = (x', x_N) \in \mathbb{R}^N \,,\ x_N \in]0, L[\right\}.$$

On suppose dans un premier temps u régulière. Pour tout $x = (x', x_N) \in \Omega$, on a

$$u(x', x_N) = u(x', 0) + \int_0^{x_N} \partial_N u = \int_0^{x_N} \partial_N u,$$

d'où, d'aprés l'inégalité de Cauchy-Schwarz,

$$u(x', x_N)^2 \leq L \int_0^L |\nabla u|^2 .$$

On a donc

$$\begin{aligned}\int_\Omega u^2 &\leq L\int_{\mathbb{R}^{N-1}}\int_0^L\int_0^L |\nabla u|^2 \\ &\leq L^2\int_{\mathbb{R}^{N-1}}\int_0^L |\nabla u|^2 = \int_\Omega |\nabla u|^2 .\end{aligned}$$

On conclut en utilisant la densité des fonctions régulières. □

Remarque 22.44. *On appelle constante de Poincaré du domaine Ω le plus petit réel C_Ω tel que l'inégalité ci-dessus est vérifiée. On a*

$$\frac{1}{C_\Omega^2} = \inf_{u\neq 0} \frac{\displaystyle\int_\Omega |\nabla u|^2}{\displaystyle\int_\Omega |u|^2}.$$

On peut ainsi montrer $1/C_\Omega^2 = \lambda_1$, où λ_1 est la plus petite valeur propre du Laplacien avec conditions de Dirichlet, c'est-à-dire le plus petit réel tel qu'il existe $u \in H_0^1(\Omega)$ non nul vérifiant[112]

$$-\triangle u = \lambda u.$$

La proposition précédente assure $\lambda_1 \geq 1/L^2$, pour tout domaine Ω inclus dans une bande d'épaisseur L.

Corollaire 22.45. *Soit Ω un domaine de $\mathbb{R}^N$ borné dans une direction. Alors la forme bilinéaire*

$$(u,v) \longmapsto \int_\Omega \nabla u\cdot\nabla v$$

est un produit scalaire sur $H_0^1(\Omega)$, qui induit une norme équivalente à la norme de départ.

L'inégalité de Poincaré énoncée ci-dessus est un cas particulier d'une inégalité plus générale :

Proposition 22.46. *(Inégalité de Poincaré généralisée)*
Soit Ω un domaine régulier, borné, et connexe, et T une application linéaire continue de $H^1(\Omega)$ dans un espace de Hilbert M. On suppose que l'image par T d'une fonction constante non nulle est elle-même non nulle. Alors il existe une constante C telle que

$$|u|_0 \leq C\left(|Tu|_M + |\nabla u|_0\right) \quad \forall u \in H^1(\Omega).$$

Démonstration: On raisonne par l'absurde. Si la propriété est fausse, alors pour tout n on peut construire $u_n \in H^1(\Omega)$ tel que

$$\|u_n\|_{L^2} > n\left(|Tu_n|_M + |\nabla u_n|_0\right) \quad \forall u \in H^1(\Omega).$$

112. L'opérateur de Laplace $-\triangle$, qui fait intervenir des dérivées secondes, n'est *a priori* défini pour des fonctions de H^1 qu'au sens des distributions. On verra par la suite que ces dérivées secondes du minimiseur u peuvent en fait être définies dans le cadre de ce chapitre, c'est-à-dire en tant que fonctions de $L^2(\Omega)$ (ou tout du moins L^2_{loc} sans hypothèse sur le domaine), de telle sorte que l'on pourra écrire $-\triangle u = \lambda u$ presque partout.

On peut choisir u_n tel que $\|u_n\| = 1$. La suite u_n étant bornée dans H^1, on peut en extraire une sous-suite (que nous noterons toujours (u_n)) qui converge fortement dans $L^2(\Omega)$ (l'injection de $H^1(\Omega)$ dans $L^2(\Omega)$ étant compacte), vers $u \in L^2(\Omega)$). Comme la suite (∇u_n) tend vers 0 dans L^2, elle est de Cauchy, et par suite (u_n) est de Cauchy dans H^1. Elle converge donc dans H^1 vers une limite, qui est nécessairement la limite u dans L^2. Comme Tu_n tend vers 0, on a nécessairement $Tu = 0$. D'autre part, comme $(\nabla u_n) \to 0$, on a $\nabla u = 0$, et ainsi u est constante sur Ω (voir proposition 22.13, page 220). Comme $Tu = 0$, cette constante est nulle, ce qui est absurde car $\|u\| = \lim \|u_n\| = 1$ □

La démonstration ci-dessus permet d'établir directement la propriété suivante :

Corollaire 22.47. *Soit Ω un domaine régulier, borné, et connexe, et V un sous-espace fermé de $H^1(\Omega)$ qui ne contient aucune fonction constante autre que 0. Alors il existe $C > 0$ tel que*

$$|u|_0 \leq C\,|\nabla u|_0 \quad \forall u \in V.$$

Remarque 22.48. *Ce corollaire s'appliquera notamment au cas où V est un espace de fonctions qui s'annulent sur une partie de la frontière de mesure non nulle. Sur un tel espace, $|u|_1$ est une norme équivalent à la norme H^1.*

22.6 Problèmes aux limites elliptiques

Nous présentons dans cette section des résultats classiques d'existence et d'unicité de solutions pour le problème de Poisson.

Conditions aux limites de Dirichlet.

On s'intéresse ici à des problèmes du type

$$\left\{ \begin{array}{rcll} -\triangle u & = & f & \text{dans} \quad \Omega \\ u & = & 0 & \text{sur} \quad \partial\Omega, \end{array} \right. \tag{22.1}$$

où f est une fonction de $L^2(\Omega)$ donnée. On parlera du problème de Poisson dans le domaine Ω.

Definition 22.49. *(Solution faible)*
On appellera solution faible de (22.1) une fonction de $H^1_0(\Omega)$ telle que

$$\int_\Omega \nabla u \cdot \nabla v = \int_\Omega f v \quad \forall v \ H^1_0(\Omega). \tag{22.2}$$

Proposition 22.50. *(Principe de Dirichlet)*
On suppose Ω borné dans une direction. Soit $f \in L^2(\Omega)$. Alors le problème 22.1 admet une unique solution faible : il existe un unique $u \in H^1_0(\Omega)$ solution de la formulation variationnelle (22.2). C'est l'unique élément de $H^1_0(\Omega)$ qui minimise la fonctionnelle

$$v \longmapsto \frac{1}{2}\int_\Omega |\nabla v|^2 - \int f v.$$

Démonstration: C'est une application directe du théorème de Lax-Milgram, avec

$$a(u,v) = \int_\Omega \nabla u \cdot \nabla v\,,\ \langle \varphi\,, v\rangle = \int_\Omega fv.$$

Noter que la forme bilinéaire $a(\,\cdot\,,\,\cdot\,)$ est bien coercive grâce à l'inégalité de Poincaré (proposition 22.43, page 228). □

Conditions aux limites de Neumann.

On considère maintenant des conditions au bord de type Neumann. Comme ces conditions ne font intervenir que les dérivées, comme l'opérateur de Laplacien lui-même, le problème de Poisson avec de telles conditions est évidemment mal posé (si l'on ajoute une fonction constante, qui est bien dans $H^1(\Omega)$ dès que Ω est borné, à n'importe quelle solution, on obtient bien une autre solution). On verra à la fin de cette section que ce problème est pourtant bien posé dans un certain espace, sous réserve que f vérifie une certaine condition. Dans un premier temps, nous utilisons un moyen élémentaire de contourner ce problème, qui consiste à rajouter au Laplacien un terme d'ordre 0. On s'intéressera donc au problème suivant

$$\left\{\begin{array}{rcll} u - \triangle u &=& f \quad \text{dans} & \Omega \\ \dfrac{\partial u}{\partial n} &=& 0 \quad \text{sur} & \partial\Omega, \end{array}\right. \tag{22.3}$$

où f est donnée.

Definition 22.51. *On appellera solution classique (dans le cas où f est au moins continue) une fonction de $C^2(\overline{\Omega})$ qui vérifie le système ci-dessus, et solution faible une fonction de $H^1(\Omega)$ telle que*

$$\int_\Omega uv + \int_\Omega \nabla u \cdot \nabla v = \int_\Omega fv \quad \forall v\ H^1(\Omega). \tag{22.4}$$

L'existence et l'unicité d'une solution faible est immédiate sans qu'il soit nécessaire de faire des hypothèses sur le domaine, comme le précise la proposition ci-dessous. Il est en revanche délicat de préciser en quel sens une solution faible est solution de (22.3), car la dérivée normale n'est en général pas définie sur le bord.

Proposition 22.52. *Soit $f \in L^2(\Omega)$. Alors le problème 22.3 admet une unique solution faible. Cette solution faible est l'élément de $H^1_0(\Omega)$ qui minimise la fonctionnelle*

$$v \longmapsto \frac{1}{2}\int_\Omega |v|^2 + \frac{1}{2}\int_\Omega |\nabla v|^2 - \int fv.$$

Démonstration: C'est de nouveau une application directe du théorème de Lax-Milgram dans $H = H^1(\Omega)$. □

22.7 Régularité des solutions faibles

Nous abordons maintenant le problème de régularité des solutions faibles construites précédemment. Il s'agit notamment de déterminer si l'équation de départ est vérifiée comme

identité entre fonctions mesurables (auquel cas il est licite de préciser *presque partout*), ou dans un sens plus faible. On considère ainsi des équations aux dérivées partielles du type

$$-\triangle u = f\,,\ u - \triangle u = f \text{ ou } -\nabla k \cdot \nabla u = f,$$

où $\triangle$ est le Laplacien $\triangle = \sum \partial^2/\partial x_i^2$, k est un champ scalaire régulier tel que $0 < m \leq k(x) \leq M < +\infty$.

Proposition 22.53. *Soit Ω un domaine de $\mathbb{R}^N$ et $u \in H^1(\Omega)$. On suppose qu'il existe $f \in L^2(\Omega)$ tel que*

$$\int_\Omega \nabla u \cdot \nabla \varphi = \int_\Omega f\varphi \quad \forall \varphi \in \mathcal{D}(\Omega).$$

Alors u est dans $H^2_{loc}(\Omega)$ et vérifie

$$-\triangle u = f \quad p.p.$$

Démonstration: On suppose dans un premier temps que Ω est l'espace $\mathbb{R}^N$ tout entier. Comme $\mathcal{D}(\Omega)$ est alors dense dans $H^1(\Omega)$, la formulation variationnelle est vérifiée pour toute fonction test de $H^1(\Omega)$, en particulier les fonctions-test particulières que nous allons construire à partir de u. Pour $h \in \mathbb{R}^N$, on introduit

$$D_h u = \frac{1}{|h|}\left(\tau_h u - u\right),$$

et l'on écrit la formulation variationnelle avec $v = D_{-h}D_h u$. Il vient

$$\int_{\mathbb{R}^N} \nabla u \cdot \nabla v = \frac{1}{|h|^2}\int_{\mathbb{R}^N} \nabla u \cdot \left(\tau_h \nabla u - 2\nabla u + \tau_{-h}\nabla u\right).$$

On peut écrire

$$\int_{\mathbb{R}^N} \nabla u \cdot \left(-\nabla u + \tau_{-h}\nabla u\right) = \int_{\mathbb{R}^N} \tau_h \nabla u \cdot \left(-\tau_h \nabla u + \nabla u\right),$$

d'où finalement

$$\int_{\mathbb{R}^N} |D_h \nabla u|^2 \leq \|f\|_{L^2} \|D_{-h}D_h u\|_{L^2} \leq \|f\|_{L^2} \|\nabla D_h u\|_{L^2} = \|f\|_{L^2} \|D_h \nabla u\|_{L^2},$$

d'après la proposition 22.15 ($(i) \Rightarrow (iii)$). On a donc

$$\|D_h \nabla u\|_{L^2} \leq \|f\|_{L^2}$$

pour tout $h \in \mathbb{R}^N$. On a donc $\|D_h \partial_i u\|_{L^2}$ uniformément borné, et donc, toujours d'après la proposition 22.15, $\partial_i u \in H^1(\mathbb{R}^N)$ pour tout $i = 1, \ldots, N$.

Dans le cas général on considère une fonction $\theta \in \mathcal{D}(\Omega)$. On a

$$\nabla(\theta u) \cdot \nabla \varphi = \nabla u \cdot \nabla(\theta\varphi) + \nabla\theta \cdot \nabla(u\varphi) - 2\varphi \nabla u \cdot \nabla \varphi,$$

et ainsi la fonction $\theta u \in H^1(\mathbb{R}^N)$ vérifie

$$\int_{\mathbb{R}^N} \nabla\left(\theta u\right) \cdot \nabla\varphi = \int_{\mathbb{R}^N} \theta f \varphi - 2\int_{\mathbb{R}^N} \varphi \nabla u \cdot \nabla\theta - \int_{\mathbb{R}^N} \varphi u \triangle\theta = \int_{\mathbb{R}^N} g\varphi \quad \forall \varphi \in \mathcal{D}(\Omega).$$

avec $g \in L^2(\mathbb{R}^N)$. La fonction θu est donc dans $H^2(\mathbb{R}^N)$ d'après ce qui précède. On a donc bien $u \in H^2_{\text{loc}}(\Omega)$. □

Proposition 22.54. *On suppose Ω borné dans une direction. Soit f un élément de $L^2(\Omega)$. La solution faible $u \in H_0^1(\Omega)$ de (22.2)) avec conditions de Dirichlet homogènes est dans $H^2_{loc}(\Omega)$ et vérifie*

$$-\triangle u = f \quad p.p.$$

Démonstration: C'est une application directe de la proposition 22.53. □

Le passage de la régularité H^2_{loc} à l'appartenance à $H^2(\Omega)$ est loin d'être immédiat. Nous nous bornerons ici à énoncer des résultats de régularité dans un certain nombre de situations.

Proposition 22.55. *Soit Ω un domaine de classe C^2, borné dans une direction, et de frontière Γ bornée. Pour tout f dans $L^2(\Omega)$, la solution faible de $-\triangle u = f$ avec conditions aux limites de Dirichlet homogènes appartient à H^2, et il existe une constante C (qui dépend du domaine Ω) telle que*

$$\|u\|_{H^2} \leq C \|f\|_{L^2}.$$

Démonstration: L'appartenance à $H^2_{\text{loc}}(\Omega)$ est assurée par la proposition 22.53. On se reportera à Brezis [2, Th. IX.25] pour une étude détaillée de la régularité près du bord. La démonstration, très technique, utilise des changements de variables permettant de se ramener au cas d'une frontière hyperplane. Pour ce dernier cas, la régularité jusqu'au bord est démontrée selon une méthode de translation analogue à celle utilisée dans la proposition 22.53, les translations étant effectuées parallèlement au bord considéré. □

Proposition 22.56. *Les conclusions du théorème ci-dessus sont valides si l'on suppose le domaine polyédrique et convexe.*

Proposition 22.57. *Les conclusions du théorème ci-dessus s'appliquent à l'équation*

$$-\nabla \cdot k \nabla u = f,$$

où k est une fonction C^1 de la variable d'espace sur $\overline{\Omega}$, minorée par une constante

Remarque 22.58. *Le cas de conditions aux limites panachées (Dirichlet sur une partie du bord, Neumann sur une autre) et très délicat. Nous admettrons que le passage d'un type de condition à l'autre ne pose pas de problème lorsque les deux composantes de la frontière se rencontrent à angle droit. On trouvera dans Costabel*[113] *une analyse détaillée de la régularité dans ce type de situation, en fonction de l'angle du raccord entre les composantes.*

Remarque 22.59. *Si l'on considère le problème*

$$u - \triangle u = f,$$

avec conditions aux limites de Dirichlet, tout ce qui a été dit précédemment reste valable, sans que l'on ait besoin de l'hypothèse que Ω soit borné dans une direction pour assurer l'existence et l'unicité d'une solution faible.

Proposition 22.60. *Soit Ω un domaine de frontière C^2 et bornée, et f un élément de $L^2(\Omega)$. La solution de (22.4) appartient à H^2, et sa dérivée normale est nulle sur $\Gamma = \partial\Omega$.*

113. M. Costabel, M. Dauge, Edge singularities for elliptic boundary value problems, Journées équations aux dérivées partielles, 1992, pp. 1–12.
`http://www.math.sciences.univ-nantes.fr/~sjm/CDROM/data/pdf/1992/A4.pdf`

22.8 Espaces de Sobolev et transformation de Fourier

On peut définir les espaces de Sobolev l'aide de la transformée de Fourier. Cette approche est particulièrement adaptée aux problèmes posés sur l'espace tout entier, ou en géométrie périodique, ce qui la place un peu en marge de cet ouvrage dont l'un des objectifs est précisément la prise en compte de géométries complexes en domaines bornés. Nous indiquons néammoins ici certains éléments de cette approche, qui permet notamment de bien comprendre le théorème de Rellich, qui est à la base de l'analyse de la méthode des éléments finis.

Definition 22.61. *Soit $u \in L^2(\mathbb{R}^N)$. On définit sa transformée de Fourier comme la fonction définie par*

$$\tilde{u}(\xi) = \frac{1}{(2\pi)^{-n/2}} \int_{\mathbb{R}^N} e^{-i\xi \cdot x} u(x)\, dx.$$

Théorème 22.62. *L'application $u \longmapsto \tilde{u}$ est une isométrie de $L^2(\mathbb{R}^N)$ sur lui-même.*

On peut définir l'espace $H^1(\mathbb{R}^N)$ à l'aide de la transformée de Fourier, ce que nous présentons ici comme un thérorème si l'on prend la définition 22.10, page 219 comme référence.

Théorème 22.63. *L'espace $H^1(\mathbb{R}^N)$ est l'ensemble des fonctions u de $L^2(\mathbb{R}^N)$ telles que*

$$\left(1 + |\xi|^2\right)^{1/2} \tilde{u} \in L^2(\mathbb{R}^N).$$

Nous démontrons à présent le théorème de Rellich 22.42 déjà énoncé à la page 228.

Théorème 22.64. *Soit Ω un domaine borné de frontière lipschitzienne. L'injection de $H^1(\Omega)$ dans $L^2(\Omega)$ est compacte.*

Démonstration: On considère une suite (u_n) bornée dans $H^1(\Omega)$. On note P l'opérateur de prolongement de la proposition 22.27, page 225. On choisit P de telle sorte que Pv soit nul à l'extérieur d'un borné K, pour tout $v \in H^1(\Omega)$. On conserve la notation (u_n) pour désigner l'image par P de la suite initiale. D'après le théorème 20.32, page 204, on peut en extraire une sous-suite qui converge faiblement dans $H^1(\mathbb{R}^N)$. On notera toujours (u_n) cette sous-suite. Quitte à translater la suite, on suppose que la limite faible est 0. On écrit à présent, pour tout $M \geq 0$

$$\|u_n\|_{L^2}^2 = \|\tilde{u}_n\|_{L^2}^2 = \int_{|\xi|<M} |\tilde{u}_n|^2 + \int_{|\xi|>M} |\tilde{u}_n|^2 \leq \int_{|\xi|<M} |\tilde{u}_n|^2 + \frac{1}{1+M^2} \int_{|\xi|>M} \left(1 + |\xi|^2\right) |\tilde{u}_n|^2 .$$

Le second terme tend vers 0 quand M tend vers $+\infty$. Il suffit donc de montrer que, pour M fixé, le premier terme tend vers 0. On a, pour tout ξ,

$$\tilde{u}_n(\xi) = \frac{1}{(2\pi)^{-n/2}} \int_{\mathbb{R}^N} e^{-i\xi \cdot x} u_n(x)\, dx = \frac{1}{(2\pi)^{-n/2}} \int_{\mathbb{R}^N} \chi_K e^{-i\xi \cdot x} u_n(x)\, dx,$$

où χ_K est la fonction caractéristique de K (de telle sorte que $\chi_K e^{-i\xi \cdot x}$ est dans $L^2(\mathbb{R})$), Cette quantité tend donc vers 0 quand n tend vers $+\infty$ d'après la convergence faible de u_n vers 0 dans L^2. Comme par ailleurs $|\tilde{u}_n(\xi)|^2$ est majoré par une constante, le théorème de convergence dominée assure donc la convergence de $|\tilde{u}_n(\xi)|^2$ vers 0 dans $L^1(B(0,M))$. On a donc bien convergence vers 0 de $\|u_n\|_{L^2}$. □

22.9 Approche H_{div}

Nous décrivons ici une approche qui permet de donner un sens aux équations de type problème de Poisson comme identité entre fonctions de L^2 sans passer par la régularité H^2.

Proposition 22.65. *Soit Ω un domaine quelconque, et $v \in L^2(\Omega)^N$. On a l'équivalence suivante :*

$$\exists C\,,\ \left|\int_\Omega v\cdot\nabla\varphi\right| \le C\,\|\varphi\|_{L^2(\Omega)} \quad \forall\varphi\in\mathcal{D}(\Omega) \Longleftrightarrow \exists \mathrm{q}\in L^2(\Omega) \text{ tel que } \int_\Omega v\cdot\nabla\varphi = -\int_\Omega \mathrm{q}\varphi.$$

On dit alors que v admet une divergence faible dans $L^2(\Omega)$, et l'on écrit $\nabla\cdot v = \mathrm{q}$.

Démonstration: La condition suffisante est conséquence immédiate de l'inégalité de Cauchy-Schwarz. Pour la condition nécessaire, on considère la forme linéaire

$$\varphi \longmapsto \int_\Omega v\cdot\nabla\varphi$$

définie sur $\mathcal{D}(\Omega)$. Comme elle est continue pour la norme $L^2(\Omega)$ d'après l'hypothèse, cette forme se prolonge par densité en une forme linéaire continue sur $L^2(\Omega)$. Comme il s'agit d'un espace de Hilbert, cette forme admet un représentant $\mathrm{q}\in L^2(\Omega)$. □

Definition 22.66. *(Espace H_{div})*
On notera H_{div} l'ensemble des champs de vecteurs $u\in L^2(\Omega)^N$ qui admettent une divergence faible L^2 au sens de la proposition précédente.

Proposition 22.67. *L'espace H_{div} est un espace de Hilbert pour le produit scalaire*

$$(u,v)_{H_{div}} = \int_\Omega u\cdot v + \int_\Omega (\nabla\cdot u)\,(\nabla\cdot v)\,.$$

Démonstration: On considère une suite de Cauchy (u_n) dans H_{div}. On a $u_n\to u\in L^2$, et $\nabla\cdot u_n\to \mathrm{q}\in L^2$. On a

$$\int_\Omega u\cdot\nabla\varphi = \lim\int_\Omega u_n\cdot\nabla\varphi = -\lim\int_\Omega \varphi\,\nabla\cdot u_n = -\int_\Omega \varphi\,\mathrm{q},$$

d'où l'on déduit que u est dans H_{div}, avec $\nabla\cdot u = \mathrm{q}$. On vérifie immédiatement la convergence de u_n vers u pour la norme de H_{div}. □

Remarque 22.68. *On peut identifier la trace normale d'un champ de H_{div} à un élément du dual topologique de $H^{1/2}$. On considère Ω un ouvert de frontière Γ Lipschitzienne et bornée. L'application qui à $u\in\mathcal{D}(\overline{\Omega})$ associe la restriction à Γ de la quantité $\nabla u\cdot n$ peut être identifiée à un élément du dual de $H^1(\Omega)$ grâce au fait que, pour toute fonction $\varphi\in\mathcal{D}(\overline{\Omega})$,*

$$\int_\Gamma \varphi u\cdot n = \int_\Omega \varphi\nabla\cdot u + \int_\Omega u\cdot\nabla\varphi.$$

L'application $\varphi\mapsto\int_\Gamma\varphi u\cdot n$ se prolonge donc par continuité en une forme linéaire continue sur $H^1(\Omega)$, que nous noterons ψ_u. Vérifions que $<\psi_u, v>$ ne dépend que de la valeur de v sur le bord. Il suffit pour cela de vérifier que H^1_0 est dans le noyau de Ψ_u. Considérons

donc $v \in H^1_0(\Omega)$. D'après la proposition 22.33, v s'écrit comme limite de fonctions v_n dans $\mathcal{D}(\Omega)$. On note ω_n le support de v_n. En admettant que la propriété de densité 22.18, page 222, s'étend à H_{div} c'est-à-dire qu'il existe $u_n \in \mathcal{D}(\Omega)^N$ tel que

$$\|u_n - u\|_{L^2(\Omega)} \to 0\,,\ \|\nabla \cdot u_n - \nabla \cdot u\|_{L^2(\omega_n)} \to 0,$$

on obtient $\langle \Psi_u\,, v\rangle = 0$. La forme linéaire s'annule donc sur H^1_0, et par suite elle peut être vue comme une forme linéaire sur l'espace quotient H^1/H^1_0 que nous avons défini comme $\tilde{H}^{1/2}$. Comme $\tilde{H}^{1/2}$ s'identifie à $H^{1/2}$ dans le cas d'une frontière Lipschitz (par l'isométrie $\tilde{v} \in \tilde{H}^{1/2} \mapsto \gamma v$), on a bien donné un sens à $u \cdot n$ sur Γ en tant qu'élement du dual de $H^{1/2}(\Gamma)$. On écrira ainsi

$$u \cdot n|_\Gamma \in H^{-1/2}(\Gamma),$$

en prenant bien garde au fait qu'il s'agit d'une identification faite selon le procédé ci-dessus. Il est en particulier illicite d'écrire "presque partout" à côté d'une égalité identifiant deux élements de cet espace.

Considérons maintenant la formulation variationnelle

$$\int_\Omega \nabla u \cdot \nabla v = \int_\Omega f v \quad \forall v \in \mathcal{D}(\Omega).$$

Cela implique que ∇v possède une divergence faible L^2. Si l'on décide de désigner par $\triangle$ l'opérateur $\nabla \cdot \nabla$, à valeurs dans $L^2(\Omega)$, défini sur l'ensemble des champs de $H^1(\Omega)$ dont le gradient admet une divergence L^2, alors on peut écrire

$$-\triangle u = f \quad \text{p.p.}$$

D'après la remarque qui précède, on peut aussi donner un sens à la trace normale du gradient $\partial u/\partial n$, non pas en tant que fonction, mais en temps que forme linéaire sur l'espace $H^{1/2}(\Gamma)$ des traces des fonctions de H^1.

22.10 Exercices

Exercice 22.3. On définit Ω et ω comme les boules de $\mathbb{R}^d$, centrées en 0, de rayons respectifs R et $r < R$.

On définit la capacité de ω (sous-entendu : vis-à-vis de Ω), comme

$$C_\omega = \inf\left\{\int_\Omega |\nabla v|^2\ ,\ v \in H^1_0(\Omega)\,,\ v = 1 \text{ p.p. sur } \omega\right\}$$

1) Montrer que l'infimum est atteint en un point unique, et que la fonction u qui réalise le minimum est solution (sur $\Omega \setminus \overline{\omega}$) du problème aux limites

$$\begin{aligned} -\Delta u &= 0 \text{ dans } \Omega \setminus \overline{\omega}, \\ u &= 0 \text{ sur } \partial\Omega, \\ u &= 1 \text{ sur } \partial\omega. \end{aligned}$$

2) Montrer que la fonction qui réalise l'infimum ne dépend que du rayon ρ (distance à l'origine).

3) On rappelle que le Laplacien d'une fonction radiale en dimension d'espace $d \geq 1$ s'écrit

$$\Delta v(\rho) = \frac{\partial^2 v}{\partial \rho^2} + \frac{(d-1)}{\rho} \frac{\partial v}{\partial \rho}.$$

Expliciter le minimiseur (solution du problème de Dirichlet ci-dessus) pour les dimensions d'espace $d = 1$, 2 et 3, et en déduire dans chacun de ces cas la valeur de la capacité comme fonction de R et r.

4) Dans quel sens peut on dire qu'un point est de capacité nulle pour les dimensions 2 et 3 ?

5) (*Cette dernière question vise à préciser le fait qu'il est impossible de donner un sens à la valeur ponctuelle d'une fonction de* $H^1(\mathbb{R}^d)$ *dès que* $d \geq 2$.)
Montrer que, pour $d = 2$ et $d = 3$, l'ensemble des fonctions C^∞ à support compact dans $\mathbb{R}^d$ privé d'un point est dense dans $H^1(\mathbb{R}^d)$.

23 Optimisation sous contrainte

23.1 Conditions nécessaires d'optimalité

Definition 23.1. *(Différentielle d'une fonctionnelle)*
Soit E un espace vectoriel normé, et J une fonctionnelle continue d'un ouvert U de E dans $\mathbb{R}$. On dit que J est différentiable en $x \in U$ s'il existe $DJ_x \in E'$ telle que

$$J(x+h) = J(x) + \langle DJ_x\,, h\rangle + o(h).$$

On appelle DJ_x la différentielle de J en x.

Definition 23.2. *(Gradient d'une fonctionnelle)*
Soit H un espace de Hilbert, et J une fonctionnelle continue d'un ouvert U de H dans $\mathbb{R}$, différentiable en $u \in U$. On appelle gradient de J en u le vecteur de H qui s'identifie à DF_u par le théorème de Riez-Fréchet. Ce gradient, noté $\nabla J(u)$, est défini par

$$J(u+h) = J(u) + (\nabla J, h) + o(h).$$

Proposition 23.3. *Soit U un ouvert d'un espace de Hilbert H, et J une fonctionnelle différentiable. Si u est un minimum local de J sur U, alors $\nabla J(u) = 0$.*

Démonstration. Pour tout $h \in H$, $u + \varepsilon h$ est dans U pour ε suffisamment petit, on a donc

$$J(u+\varepsilon h) \geq J(u),$$

pour ε petit, d'où $\nabla J(u) \cdot h = 0$. □

Proposition 23.4. *Soit U un ouvert convexe d'un espace de Hilbert, et J une fonctionnelle différentiable et convexe. Si $\nabla J(u) = 0$, alors u est un minimum global de J sur U. Si la fonctionnelle est strictement convexe, ce minimiseur est unique.*

Démonstration. Pour tout $v \in U$, on a

$$J(v) \geq J(u) + (\nabla J(u), v-u) = J(u).$$

□

L'essentiel de ce qui suit est consacré à la notion de multiplicateur de Lagrange, variable auxilliaire permettant de prendre en compte une contrainte dans un problème de minimisation. Le cœur de l'approche repose sur l'utilisation de variations autour d'un minimiseur. Dans le cas sans contrainte vu prédédemment, toutes les directions étaient permises, ce qui a permis de conclure à l'annulation de la différentielle. Dans le cas contraint, seules les variations qui ne font pas sortir de l'ensemble sont autorisées.

Proposition 23.5. *Soit J une fonctionnelle C^1 sur un ouvert U de $V = \mathbb{R}^d$. On suppose que J admet un minimum local sur $U \cap K$ en u, avec*

$$K = u_0 + \ker B\,,\ B \in \mathcal{M}_{Nd}(\mathbb{R}).$$

Il existe alors $\lambda \in \mathbb{R}^N$ tel que

$$\nabla J(u) + B^\star \lambda = 0.$$

Démonstration. Pour tout $v \in \ker B$ de norme ≤ 1, tout ε assez petit, on a

$$J(u + \varepsilon v) \geq J(u).$$

Pour v fixé, on a donc

$$J(u) + \varepsilon \nabla J(u) \cdot v + o(\varepsilon) \geq J(u),$$

d'où l'on déduit que $\nabla J(u) \cdot v = 0$. On a donc $\nabla J(u) \in K^\perp = (\ker B)^\perp = \operatorname{im} B^\star$, d'où le résultat □

Remarque 23.6. *La proposition précédente s'applique immédiatement au cas où V est un espace de Hilbert, qui peut être de dimension infinie, il suffit de remplacer la matrice B exprimant les contraintes (qui se trouverait avoir une infinité de colonnes) par une application qui envoie V dans $\mathbb{R}^N$:*

$$B \ : \ v \longmapsto \left(\langle \varphi_i \, , v \rangle \right)_i ,$$

où les φ_i sont éléments de V'. L'image de B étant fermée, on a $(\ker B)^\perp = \operatorname{im} B^\star$, d'où l'existence du vecteur λ de multiplicateurs de Lagrange.

Si maintenant B, envoie V linéairement et continûment dans Λ, espace de Hilbert de dimension infinie, alors on a seulement (voir proposition 19.21, page 194)

$$(\ker B)^\perp = \overline{\operatorname{im} B^\star}.$$

Si l'image de B est fermée (ce qui est équivalent au fait que l'image de $B^\star$ soit fermée d'après la proposition 19.22, page 194), on aura bien existence d'un $\lambda \in \Lambda$ comme dans la proposition ci-dessus (on identifie Λ à son dual) :

Proposition 23.7. *Soit J une fonctionnelle C^1 sur un ouvert U d'un espace de Hilbert V. On considère*

$$K = u_0 + \ker B,$$

avec $B \in \mathcal{L}(V, \Lambda)$ à image fermée. Si u est un minimiseur local de J sur $U \cap K$, alors il existe $\lambda \in \Lambda$ tel que

$$\begin{aligned} \nabla J(u) + B^\star \lambda &= 0 \\ Bu &= Bu_0. \end{aligned}$$

Remarque 23.8. *Dans le cas où l'image de B n'est pas fermée, il est possible qu'un tel λ n'existe pas. On pourra en revanche toujours trouver une suite (λ_ε) telle que*

$$\nabla J(u) + B^\star \lambda_\varepsilon = o(1).$$

23.2 Contraintes non linéaires d'égalité

On s'intéresse à la minimisation d'une fonctionnelle J sur un ouvert U de $\mathbb{R}^d$, sur un sous-ensemble défini par N contraintes :

$$K = \Big\{ \ v \in \mathbb{R}^d \, , \ \varphi_i(v) = 0 \, , \ i = 1, \ldots, N \Big\} .$$

Proposition 23.9. *(Muliplicateurs de Lagrange, contraintes d'égalité)*
Soit $J : U \subset \mathbb{R}^d \longrightarrow \mathbb{R}$ une fonctionnelle C^1 sur l'ouvert U. Soit u un point de $U \cap K$ en lequel J réalise un minimum local de J sur $U \cap K$. On suppose que les gradients des fonctionnelles φ_i forment une famille libre. Il existe alors $\lambda_1, \ldots, \lambda_N$, tels que

$$\nabla J(u) + \sum_{i=1}^{N} \lambda_i \nabla \varphi_i(u) = 0.$$

Démonstration. Le point clé consiste à montrer que tout vecteur h orthogonal à tous les $\nabla\varphi_i(u)$, est une *direction admissible* en u, c'est à dire qu'il existe $\eta(t)$ défini dans un voisinage de 0, avec $\eta(0) = 0$, tel que $u + \eta(t) \in K$, et que la tangente en 0 soit h, c'est à dire que $\dot\eta(0) = h$. Si cette propriété est vraie, alors on peut écrire pour tout h orthogonal aux $\nabla\varphi_i(u)$, et η une trajectoire associée selon les considérations précédentes,

$$J(u + \eta(t)) \geq J(u)$$

pour tout t dans un voisinage de 0, d'où

$$\nabla J \cdot \dot\eta(0) = \nabla J \cdot h = 0.$$

Le gradient de J est ainsi orthogonal à l'orthogonal de $\mathrm{vect}(\nabla\varphi_i(u))_i$, ce qui termine la preuve.

Montrons maintenant que tout vecteur h orthogonal à tous les $\nabla\varphi_i(u)$, est une *direction admissible* en u.

On note $g_i = \nabla\varphi_i(u)$, et

$$V = \mathrm{vect}(g_1, \ldots, g_N)^\perp.$$

Comme les vecteurs g_i forment une famille libre, V est de dimension $d - N$. On considère une base $(h_1, \ldots, h_{d-N})$ de V, on note

$$x = (x_1, \ldots, x_{d-N}) \in \mathbb{R}^{d-N}\ ,\ y = (y_1, \ldots, y_N) \in \mathbb{R}^N$$

et l'on définit γ l'application

$$\gamma\ :\ (x, y) \in \mathbb{R}^d \longmapsto \gamma(x, y) = u + x_1 h_1 + \ldots x_{d-N} h_{d-N} + y_1 g_1 + \ldots y_N g_N.$$

On notera γ_k l'application qui ne dépend que de x_k et des y_i, les autres x_j étant fixés à 0. Pour construire une courbe dans K qui passe par u, dont la tangente en u est h_k, on considère l'application

$$(x_k, y_1, y_2, \ldots, y_N) \longmapsto \varphi \circ \gamma_k(x_k, y_1, \ldots, y_N),$$

où l'on note $\varphi(v)$ le vecteur de dimension N dont les composantes sont les $\varphi_i(v)$. Comme $u \in K$, l'application $\varphi \circ \gamma_k$ est nulle en 0. Montrons que l'on peut utiliser le théorème des fonctions implicites pour construire une courbe $(y_1, \ldots, y_N) = y = y(x_k)$ au voisinage de $(x_k, y) = 0$ qui annule $\varphi \circ \gamma_k$, ce qui assurera l'appartenance de $\gamma_k(x_k, y)$ à K . La différentielle de la $i^{\text{ième}}$ composante de $\varphi \circ \gamma_k$ par rapport à y_j est

$$\frac{\partial(\varphi_i \circ \gamma_k)}{\partial y_j} = \nabla\varphi_i(x_k, y) \cdot g_j = \nabla\varphi_i(x_k, y) \cdot \nabla\varphi_j(0, 0).$$

Notons G la matrice dont les colonnes sont les gradients des φ_j en $\gamma_k(0,0) = u$. Le gradient de l'application $\varphi \circ \gamma_k$ est ainsi $G^T G$, qui est inversible puisque les g_i forment une famille libre.

On a par ailleurs

$$\frac{\partial(\varphi_i \circ \gamma_k)}{\partial x_k} = \nabla\varphi_i(x_k, y) \cdot h_k\,, \quad \text{d'où} \quad \frac{\partial(\varphi \circ \gamma_k)}{\partial x_k}|_{(0,0)} = G^T\, h_k.$$

On peut donc construire une courbe $y = y(t)$ dans un voisinage de 0 telle que

$$\varphi \circ \gamma_k(t, y(t)) = 0$$

c'est à dire que la courbe est dans K. La dérivée de y en 0 s'écrit, d'après le théorème des fonctions implicites,

$$\dot{y}(0) = -\left(\nabla(\varphi \circ \gamma_k)\right)^{-1} \frac{\partial(\varphi \circ \gamma_k)}{\partial x_k} = \left(G^T G\right)^{-1} \left(G^T h_k\right)$$

qui est nul car h_k est orthogonal à tous les g_i. On a donc

$$\frac{d}{dt}\gamma_k(t, y(y))_{|t=0} = h_k + \dot{y}_1(0)g_1 + \cdots + \dot{y}_N(0)g_N = h_k,$$

ce qui termine la démonstration. □

Remarque 23.10. *La condition d'indépendance des gradients est essentielle dans la proposition précédente. On pourra par exemple considérer, dans* $\mathbb{R}^2$, $\varphi_1(x,y) = y$ *et* $\varphi_1(x,y) = y - x^2$. *L'ensemble* K *est réduit au point* $(0,0)$, *et n'importe quelle fonctionnelle dont le gradient en* $(0,0)$ *n'est pas colinéaire à* $(0,1)$ *invalide la proposition.*

23.3 Contraintes unilatérales (ou d'inégalité)

H désigne dans la suite un espace de Hilbert.

Definition 23.11. *(Cône)*
On appelle cône de sommet $s \in H$ *une partie* C *de* H *telle que*

$$u - s \in C \Longrightarrow \lambda(u - s) \in C \quad \forall \lambda > 0,$$

Lorsque le sommet est l'origine 0, on omettra de le préciser. Un cône convexe fermé C (de sommet 0), est donc un ensemble convexe fermé tel que $\mathbb{R}^+ C \subset C$.

Definition 23.12. *(Polaire d'un ensemble)*
Soit K *une partie de* H, *on définit le polaire de* K *comme*

$$K^\circ = \{v \in H\,,\ (v,u) \leq 0 \quad \forall u \in K\}\,.$$

Noter que dans le cas où K est un sous-espace vectoriel de H, l'ensemble K° est simplement l'orthogonal de K. Cette définition est donc une généralisation de la définition 19.20, page 194.

Proposition 23.13. *Pour tout $K \subset H$, K° est un cône convexe fermé.*

Definition 23.14. *(Enveloppe conique)*
Soit $K \subset H$. On appelle enveloppe convexe conique (on dira simplement enveloppe conique) de K le plus petit cône convexe qui contient K. On la note $co(K)$ C'est l'intersection des cônes convexes qui contiennent K. On appelle enveloppe conique fermée le plus petit cône convexe fermé qui contient K. On notera cet ensemble $\overline{co}(K)$.

Proposition 23.15. *Soit $K \subset H$ une partie de H. On a*

$$K^\circ = (co(K))^\circ = (\overline{co}(K))^\circ .$$

Proposition 23.16. *Soit $K \in H$ une partie quelconque de H, K° son polaire, et $K^{\circ\circ} = (K^\circ)^\circ$ son bipolaire. Alors $K^{\circ\circ}$ est l'enveloppe convexe fermée conique de K. En particulier, si K est un cône convexe fermé (de sommet 0), alors $K^{\circ\circ} = K$.*

Démonstration. L'inclusion $K \subset K^{\circ\circ}$ est immédiate : tout v dans K a un produit scalaire négatif contre tout élément de K°, il est donc dans $K^{\circ\circ}$. Comme $K^{\circ\circ}$ est un cône convexe fermé, l'inclusion demeure par passage à l'enveloppe convexe fermé conique.

On appelle C l'enveloppe convexe fermée conique de K. Si l'inclusion est stricte, il existe $z \in K^{\circ\circ}$ qui n'appartient pas à C. On peut alors, d'après [114] le théorème de Hahn-Banach 19.2, page 191, séparer le convexe fermé C de $\{z\}$: il existe h tel que

$$(h, v) \leq \alpha < (h, z) \quad \forall v \in C.$$

Comme v décrit un cône de sommet 0, (h, v) est forcément négatif ou nul pour tout v (s'il prenait une valeur strictement positive, le sup serait $+\infty$, ce qui est exclut par la majoration ci-dessus). On a donc $h \in C^\circ$. Par ailleurs le maximum de (h, v) est 0, et donc $\alpha \geq 0$, d'où $(h, z) > 0$ ce qui est absurde car $h \in C^\circ$ et $z \in C^{\circ\circ}$. □

On s'intéressera en particulier à des ensembles de la forme

$$C = \left\{ \sum_{i=1}^{n} \lambda_i g_i \,,\ \lambda_i \geq 0 \quad \forall i = 1, \ldots, n \right\}, \tag{23.1}$$

où les g_i sont des points d'un espace de Hilbert H. L'ensemble défini précédemment est de façon évidente un cône convexe. S'il est immédiat que l'espace vectoriel engendré par une famille finie de vecteurs est fermée, il est un peu plus délicat de démontrer une telle propriété de fermeture pour le cône (convexe) engendré par une telle famille. C'est l'objet de la proposition suivante :

Proposition 23.17. *Le cône convexe C défini par* (23.1) *est fermé.*

114. Pour le lecteur qui s'inquièterait légitimement du fait que l'on doive utiliser l'axiome du choix (au cœur du "grand" théorème de Hahn-Banach) pour donner un sens par exemple à la pression ressentie par les passagers du métro aux heures de pointe, précisons que nous n'avons en fait besoin ici que d'un propriété de séparation d'un convexe fermé et d'un point, dans un espace de Hilbert. Une telle propriété se montre immédiatement à l'aide de la projection du point sur le convexe fermé.

Démonstration. Supposons dans un premier temps que les g_i forment une famille libre. On se place dans l'espace vectoriel W engendré par les g_i, et l'on note G l'application (linéaire continue) qui à un vecteur de cet espace associe le vecteur des coefficients dans la base des g_i. On considère une suite $v^k = \sum \lambda_i^k g_i$ qui converge vers $v \in W$. Alors Gv^k converge vers Gv, i.e. le vecteur λ^k converge vers un vecteur λ de $\mathbb{R}$, dont toutes les composantes sont positives ou nulle par continuité, on a donc bien $v \in C$.

Si maintenant la famille est liée, on raisonne par récurrence sur le nombre de vecteurs g_i. Supposons que tout cône convexe engendré par n vecteurs est fermé, et considérons une famille de $n+1$ vecteurs. Il existe μ_1, ..., μ_{n+1}, non tous nuls, tels que

$$\sum_{i=1}^{n+1} \mu_i g_i = 0. \tag{23.2}$$

On considère une suite dans K qui converge vers $v \in H$:

$$\sum_{i=1}^{n+1} \lambda_i^k g_i \longrightarrow v.$$

Si l'une des suites $(\lambda_i^k)_k$ est bornée, par exemple $(\lambda_1^k)_k$, on peut en extraire une sous-suite qui converge vers $\lambda_1 \in \mathbb{R}^+$, et par suite

$$v = \lim \sum_{i=1}^{n+1} \lambda_i^k g_i = \lambda_1 g_1 + \lim \sum_{i=2}^{n+1} \lambda_i^k g_i.$$

D'après l'hypothèse de récurrence, la limite ci-dessus est dans le cône convexe engendré par les $(g_i)_{2 \leq i \leq n+1}$, et par suite v est dans le cône convexe engendré par les $(g_i)_{1 \leq i \leq n+1}$. Si l'une des suites est bornée, on montre ainsi que la limite est dans K. Il reste a étudier le cas où toutes les suites sont non bornées. Quitte à extraire une sous-suite, on peut supposer que toutes ces suites (de termes positifs ou nuls) tendent vers $+\infty$.

On reprend maintenant la combinaison non triviale (23.2), en supposant (quitte à prendre la combinaison opposée), que l'un des coefficients est strictement négatif. On considère alors, pour tout k, le plus grand $\beta^k > 0$ tel que $\lambda_i^k + \beta^k \mu_i \geq 0$ pour tout $1 \leq i \leq n+1$. L'inégalité est en fait une égalité pour au moins l'un des indices. Au moins l'un des indices i_0 réalise l'égalité une infinité de fois, on extrait la sous-suite correspondante (sans changer les indices pour alléger les notations). La limite v s'écrit donc comme

$$v = \lim \sum_{i \neq i_0} (\lambda_i^k + \beta^k \mu_i) g_i$$

qui est dans le cône convexe engendré par les n vecteurs $(g_i)_{i \neq i_0}$ (d'après l'hypothèse de récurrence), donc dans C. □

On déduit de ce caractère fermé une propriété essentielle

Proposition 23.18. *(Lemme de Farkas)*
Soient $(g_i)_I$ une famille finie de vecteurs d'un espace de Hilbert H, et

$$K = \{h \in H\,,\ g_i \cdot h \leq 0 \quad \forall i \in I.\}$$

L'ensemble des vecteurs qui ont un produit scalaire négatif avec tous les éléments de K est

$$K^\circ = \left\{ \sum_{i\in I} \lambda_i g_i \,,\ \lambda_i \geq 0 \quad \forall i \right\}.$$

Démonstration. L'ensemble K est de façon évidente le cône polaire de

$$C = \left\{ \sum_{i\in I} \lambda_i g_i \,,\ \lambda_i \geq 0 \quad \forall i \right\},$$

qui, comme cône convexe fermé (d'après la proposition 23.17), s'identifie à son bipolaire (proposition 23.16). On a donc

$$K^\circ = C^{\circ\circ} = C.$$

□

Remarque 23.19. *On peut voir ce lemme de Farkas comme une version unilatérale de la proposition 19.3, page 191, qui est elle-même une généralisation de la propriété* $(\ker B)^\perp = \mathrm{Im} B^\star$ *pour les matrices. Cette proposition assure que si un vecteur g est orthogonal à tout vecteur h lui-même orthogonal à des vecteurs g_1, …, g_n, alors g est combinaison linéaire des g_i. Le présent lemme de Farkas est en fait une stricte généralisation (dans le contexte Hilbertien) de cette proposition, puisqu'il suffit de dédoubler la famille des g_i (en rajoutant $-g_i$) pour que C soit en fait le sous-espace orthogonal à* $\mathrm{vect}(g_i)$.

Exercice 23.1. Énoncer et démontrer une version non hilbertienne du lemme de Farkas. On pourra considérer un e.v.n. E, g_1, …, g_n des éléments de E, et définir K comme l'ensemble des $f \in E'$ négatives contre tout g_i.

Contraintes d'inégalité.

On s'intéresse ici à la minimisation de fonctionnelles sur des ensembles du type

$$K = \{\ v \in H\,,\ \varphi_i(v) \leq 0\,,\ i = 1, \ldots, N\} \tag{23.3}$$

Definition 23.20. *(Contraintes actives)*
On dit que la contrainte i est active en $u \in H$ dès que $\varphi(u) = 0$. On note I_u l'ensemble des i tels que la contrainte i est active en u.

Definition 23.21. *(Qualification des contraintes)*
Soit $u \in H$, et I_u l'ensemble des contraintes actives en u. On dit que les contraintes $[\varphi_i \leq 0]$ sont qualifiées en $u \in H$ s'il existe un vecteur $h \in H$ tel que

$$\nabla\varphi_i(u) \cdot h < 0$$

ou simplement $\nabla\varphi_i(u) \cdot h \leq 0$ si φ_i est affine, pour tout $i \in I_u$.

Proposition 23.22. *Soit J une fonctionnelle C^1 sur H, et u un minimiseur local de J sur K (défini par* (23.3)*). On suppose que les contraintes sont qualifiées en u. Il existe alors λ_1, λ_2, …, $\lambda_N \geq 0$ tels que*

$$\nabla J(u) + \sum_{i=1}^{N} \lambda_i \nabla\varphi_i = 0\,,$$

avec $\varphi(u) \cdot \lambda = 0$ (ce qui implique que $\lambda_i = 0$ dès que la contrainte i n'est pas saturée).

Démonstration. Soit h vérifiant $\nabla\varphi_i(u)\cdot h < 0$ pour toute contrainte i active en u (avec éventuellement égalité pour une contrainte affine). Pour $t > 0$ suffisamment petit, on a $u+th \in K$, et donc

$$J(u+th) \geq J(u) \quad \forall t \in [0, t^\star[,$$

d'où

$$J(u) + t\nabla J(u)\cdot h + o(t) \geq \nabla J(u),$$

et donc nécessairement

$$\nabla J(u)\cdot h \geq 0.$$

Pour tout h tel que l'on ait simplement l'inégalité au sens large $\nabla\varphi_i(u)\cdot h \leq 0$, on a la même propriété. En effet, considérons un $h^\star$ pour lequel on a les inégalités strictes, on préserve les inégalités strictes pour $(1-\varepsilon)h + \varepsilon h^\star$, d'où

$$\nabla J(u)\cdot\ ((1-\varepsilon)h + \varepsilon h^\star) \geq 0,$$

et donc $\nabla J(u)\cdot h \geq 0$ par passage à la limite $\varepsilon \to 0$.

Le vecteur $-\nabla J$ est donc dans $C^{\circ\circ}$, polaire de

$$C^\circ = \{h \in H\,,\ \nabla\varphi_i \cdot h \leq 0\}$$

qui s'identifie à

$$C = \left\{ \sum_{i\in I_u} \lambda_i \nabla\varphi_i(u)\,,\ \lambda_i \geq 0 \right\}$$

d'après le lemme de Farkas (proposition 23.18). Il existe donc des λ_i positifs ou nuls tels que

$$\nabla J(u) + \sum_{i\in I_u} \lambda_i \nabla\varphi_i(u) = 0.$$

On obtient une somme sur tous les i en complétant par des multiplicateurs de Lagrange nuls sur les contraintes non actives.

□

23.4 Point-selle, théorème de Kuhn et Tucker

Lemme 23.23. *Soient V et Λ deux* ensembles*, et $L(\,\cdot\,,\,\cdot\,)$ une application de $V\times\Lambda$ dans $\mathbb{R}$. On définit*

$$G(q) = \inf_{v\in V} L(v,q) \in [-\infty, +\infty[\,,\ F(v) = \sup_{q\in\Lambda} L(v,q) \in]-\infty, +\infty]. \tag{23.4}$$

On a alors

$$G(q) \leq F(v) \quad \forall q \in \Lambda\,,\ v \in V.$$

Démonstration. On écrit simplement, pour tout $q \in \Lambda$, tout $v \in V$,

$$G(q) \leq L(v,q) \leq F(v).$$

□

Definition 23.24. *Dans le contexte, et avec les notations, du lemme précédent, on appellera*

- *problème* primal *le problème de minimisation de F sur V, et*

- *problème* dual *le problème de maximisation de G sur Λ.*

Definition 23.25. *(Point-selle)*
Soient V et Λ deux ensembles, *et $L(\,\cdot\,,\,\cdot\,)$ une application de $V \times \Lambda$ dans $\mathbb{R}$. On dit que (u, p) est un point selle de L (sur $V \times \Lambda$) si*

$$L(u, q) \leq L(u, p) \leq L(v, p) \quad \forall q \in \Lambda\,,\ v \in V.$$

Proposition 23.26. *Soient V et Λ deux* ensembles, *$L(\,\cdot\,,\,\cdot\,)$ une application de $V \times \Lambda$ dans $\mathbb{R}$, et G et F définies par* (23.4)*. Les deux assertions suivantes sont équivalentes :*

(i) $L(\,\cdot\,,\,\cdot\,)$ admet un point-selle (u, p) (Def. 23.25)

(ii) Le sup de G est atteint en un point $p \in \Lambda$, l'inf de F est atteint en un point $u \in V$, et ces deux quantités sont égales.

Démonstration. $(i) \Longrightarrow (ii)$ On note $m = L(u, p)$ la valeur de L au point-selle. On a $F(u) \geq L(u, q)$ pour tout q, en particulier $F(u) \geq L(u, p) = m$. Par ailleurs $F(u) = \sup_q L(u, q) \leq L(u, p) = m$. Donc $F(u) = m$. On a maintenant $F(v) \geq L(v, q)$ pour tout q, en particulier $F(v) \geq L(v, p) \geq L(u, p) = m$ d'après la seconde inégalité du point-selle. L'infimum de F est donc bien atteint, en u, avec $F(u) = m$. On montre de façon symétrique que le supremum de G est atteint, en p, avec $G(p) = m$.

$(ii) \Longrightarrow (i)$ On suppose maintenant

$$\sup G = G(p) = m = F(u) = \inf F.$$

On a $m = G(p) \leq L(u, p) \leq F(u) = m$, d'où $L(u, p) = m$. On a par ailleurs, pour tout $q \in \Lambda$, $L(u, q) \leq F(u) = m$, et pour tout $v \in V$, $L(v, p) \geq G(p) = m$.

□

Le lien entre les problèmes de minimisation sous contraintes et la notion de point-selle passe par la définition d'une fonctionnelle appelée Lagrangien :

Definition 23.27. *(Lagrangien)*
Soit J une fonctionnelle d'un ensemble X dans $\mathbb{R}$, et K un ensemble défini par N_u contraintes d'inégalité et N_e contraintes d'égalité :

$$K = \{v \in X\,,\ \varphi_i(v) \leq 0\,,\ \psi_j(v) = 0 \quad \forall i\,,\ j\,,\ 1 \leq i \leq N_u\,,\ 1 \leq j \leq N_e\}$$

Le Lagrangien associé au problème de minimisation de J sur K est défini par

$$(u, p, q) \in X \times \mathbb{R}_+^{N_u} \times \mathbb{R}^{N_e} \longmapsto L(u, p, q) = J(u) + \sum_{i=1}^{N_u} p_i \varphi_i(u) + \sum_{j=1}^{N_e} q_j \psi_j(u). \tag{23.5}$$

Conformément à la définition 23.25, on dira que $(u,p,q) \in X \times \mathbb{R}_+^{N_u} \times \mathbb{R}^{N_e}$ est point-selle du Lagrangien défini par (23.5) si

$$L(u,\tilde{p},\tilde{q}) \leq L(u,p,q) \leq L(\tilde{u},p,q) \quad \forall \tilde{p} \in \mathbb{R}_+^{N_u}\,,\ \tilde{q} \in \mathbb{R}^{N_e}\,,\ \tilde{u} \in X.$$

Proposition 23.28. *On considère une fonctionnelle d'un ensemble X dans $\mathbb{R}$, et l'on suppose que le Lagrangien associé au problème de minimisation de J sur*

$$K = \{v \in X\,,\ \varphi_i(v) \leq 0\,,\ \psi_j(v) = 0 \quad \forall i\,,\ j\,,\ 1 \leq i \leq N_u\,,\ 1 \leq j \leq N_e\}$$

admet un point-selle $(u,p,q) \in X \times \mathbb{R}_+^{N_u} \times \mathbb{R}^{N_e}$, c'est à dire que

$$J(u)+\sum_{i=1}^{N_u} \tilde{p}_i\varphi_i(u)+\sum_{j=1}^{N_e} \tilde{q}_j\psi_j(u) \leq J(u)+\sum_{i=1}^{N_u} p_i\varphi_i(u)+\sum_{j=1}^{N_e} q_j\psi_j(u) \leq J(\tilde{u})+\sum_{i=1}^{N_e} p_i\varphi_i(\tilde{u})+\sum_{j=1}^{N_e} q_j\psi_j(\tilde{u})$$

$$\forall \tilde{p} \in \mathbb{R}_+^{N_u}\,,\ \tilde{q} \in \mathbb{R}^{N_e}\,,\ \tilde{u} \in X.$$

Alors u minimise J sur K, et l'on a $p_i\varphi_i(u) = 0$ pour tout i.

Si X est un ouvert d'un espace de Hilbert, et que les fonctions J, φ_1, ..., ψ_{N_e} sont dérivables, alors on a de plus

$$\nabla J(u) + \sum_{i=1}^{N_u} p_i \nabla\varphi_i(u) + \sum_{j=1}^{N_e} q_j \nabla\psi_j(u) = 0.$$

Démonstration. D'après la première inégalité du point-selle, la quantité $\sum \tilde{q}_j\psi_j(u)$ est bornée sur $\mathbb{R}^{N_e}$, on a donc nécessairement $\psi_j(u) = 0$ pour tout j. De la même manière, la quantité $\sum \tilde{p}_i\varphi_i(u)$ est bornée sur $\mathbb{R}_+^{N_u}$, on a donc nécessairement $\varphi_i(u) \leq 0$ pour tout i. On montre ainsi $u \in K$. On a par ailleurs (en utilisant encore cette première inégalité avec $\tilde{p} = 0$ et $\tilde{q}_j = q_j$) $0 \leq \sum p_i\varphi_i(u)$. Comme il s'agit d'une somme de termes négatifs ou nuls, tous les termes sont nuls : $p_i\varphi_i(u) = 0$, et ainsi $p_i = 0$ dès que $\varphi_i(u) < 0$ (la contrainte n'est pas activée). On utilise maintenant la seconde inégalité :

$$J(u) = J(u) + \sum_{i=1}^{N_u} p_i\varphi_i(u) + \sum_{j=1}^{N_e} q_j\psi_j(u) \leq J(\tilde{u}) + \sum_{i=1}^{N_u} p_i\varphi_i(\tilde{u}) + \sum_{j=1}^{N_e} q_j\psi_j(\tilde{u})$$

qui est en particulier inférieur à $J(\tilde{u})$ pour tout $\tilde{u} \in K$.

Si X est un ouvert d'un espace de Hilbert et si les fonctions impliquées dans le problème (fonctionnelle à minimiser et fonctions définissant les contraintes) sont régulières, alors la fonctionnelle

$$v \longmapsto \nabla J(v) + \sum_{i=1}^{N_u} p_i \nabla\varphi_i(v) + \sum_{j=1}^{N_e} q_j \nabla\psi_j(v)$$

est régulière, et le fait que u la minimise implique que son gradient soit nul en u (proposition 23.3), ce qui conclut la démonstration.

□

Théorème 23.29. *(Kuhn et Tucker)*
On considère un ouvert convexe U de $\mathbb{R}^d$, J convexe différentiable sur U, et l'ensemble admissible

$$K = \{v\,,\ \varphi_i(u) \leq 0\,,\ 1 \leq i \leq N\}\,.$$

On suppose les φ_i différentiables et convexes sur U.

On suppose de plus qu'il existe $(u,p) \in (U \cap K) \times \mathbb{R}_+^N$ tel que

$$u \in U \cap K\,,\ p \cdot \varphi(u) = 0\,,\ \nabla J(u) + \sum_{i=1}^{N} p_i \nabla \varphi_i(u) = 0. \tag{23.6}$$

Le couple (u,p) est alors point-selle du Lagrangien

$$L(v,q) = J(v) + q \cdot \varphi(v)$$

sur $U \times \mathbb{R}_+^N$ et u minimise ainsi J sur $U \cap K$.

Démonstration. De la dernière condition de (23.6) on déduit que u minimise la fonctionnelle (convexe)

$$v \longmapsto J(v) + p \cdot \varphi(v),$$

sur le convexe U (voir proposition 23.4). Comme cette fonctionnelle est convexe, on en déduit la seconde inégalité du point-selle. On a par ailleurs, comme les $\varphi_i(u)$ sont négatifs,

$$J(u) + q \cdot \varphi(u) \leq J(u)$$

pour tout $q \in \mathbb{R}_+^N$. Mais on a aussi $J(u) = J(u) + p \cdot \varphi(u)$ par hypothèse (deuxième de (23.6)), d'où la première inégalité du point-selle. □

Corollaire 23.30. *Le théorème précédent s'applique au cas de contraintes d'égalité affines.*

Démonstration. Il suffit d'écrire chaque contrainte d'égalité comme deux contraintes d'inégalité. □

23.5 Compléments

Proposition 23.31. *On considère une fonctionnelle d'un ensemble X dans $\mathbb{R}$, et l'on suppose que le Lagrangien associé au problème de minimisation de J sur*

$$K = \{v \in V\,,\ \varphi_i(v) \leq \alpha_i\,,\ 1 \leq i \leq n\}\,,$$

admet un point-selle pour tout $\alpha = (\alpha_i)_{1 \leq i \leq n}$ dans un voisinage de 0, i.e.

$$J(u^\alpha) + \sum \tilde{p}_i^\alpha\,(\varphi_i(u^\alpha) - \alpha_i) \leq J(u^\alpha) + \sum p_i^\alpha\,(\varphi_i(u^\alpha) - \alpha_i) \leq J(\tilde{u}^\alpha) + \sum p_i^\alpha\,(\varphi_i(\tilde{u}^\alpha) - \alpha_i) \quad \forall \tilde{p} \geq 0\,,\ \tilde{u} \in X.$$

On note $m(\alpha)$ la valeur du minimum correspondant aux contraintes α. On a

$$m(\alpha) \geq m(0) - p^0 \cdot \alpha.$$

Si la fonction $\alpha \longmapsto m(\alpha)$ est dérivable, alors

$$p_i = -\frac{\partial m}{\partial \alpha_i}.$$

Démonstration. On a (d'après la seconde inégalité qui caractérise (u^0, p^0) comme point-selle)

$$m(0) = J(u^0) = J(u^0) + \sum_{i=1}^{n} p_i^0 \varphi_i(u^0) \leq J(u^\alpha) + \sum_{i=1}^{n} p_i^0 \varphi_i(u^\alpha) = J(u^\alpha) + \sum_{i=1}^{n} p_i^0 \left(\varphi_i(u^\alpha) - \alpha_i\right) + \sum_{i=1}^{n} p_i^0 \alpha_i$$

qui est (d'après la première inégalité qui caractérise (u^α, p^α) comme point-selle) plus petit que

$$J(u^\alpha) + \sum p_i^\alpha \left(\varphi_i(u^\alpha) - \alpha_i\right) + \sum_{i=1}^{n} p_i^0 \alpha_i = J(u^\alpha) + \sum_{i=1}^{n} p_i^0 \alpha_i$$

On obtient donc bien $m(\alpha) = J(u^\alpha) \geq m(0) - p^0 \cdot \alpha$.

Pour α fixé, ε petit, on a, si l'on admet la dérivabilité de m par rapport à α,

$$m(\varepsilon\alpha) = m(0) + \varepsilon \nabla m(0) \cdot \alpha + o(\varepsilon)$$

d'où

$$\nabla m(0) \cdot \alpha + o(1) \geq -p^0 \cdot \alpha,$$

pour tout α décrivant un voisinage symétrique de 0. On a donc bien $\nabla m = -p^0$.

□

23.6 Illustrations

Système masses - ressorts. Considérons une chaîne horizontale de $n+1$ masses 0, 1, 2, ..., n, reliées entre elles (0 reliée à 1, 1 à 2, etc...) par des ressorts de longueur au repos nulle et de raideur k. Les positions de ces masses sont représentées par le vecteur position $(x_0, x_1, \ldots, x_n) \in \mathbb{R}^{n+1}$. L'énergie potentielle du système s'écrit

$$J(x) = \frac{1}{2} k \sum_{i=1}^{n} |x_i - x_{i-1}|^2 = \frac{1}{2} k (Ax, x),$$

où A est (à une constante multiplicative près) la matrice du Laplacien discret avec conditions de Neuman. Tout point diagonal $(x, x, \ldots, x)$ de $\mathbb{R}^{n+1}$ minimise cette énergie. On s'intéresse maintenant à la situation où la masse 0 est fixée au point $x_0 = 0$, et la masse n au point $x_n = L > 0$. Il s'agit donc maintenant de minimiser J sur l'espace affine

$$E = \{x\,,\ x_0 = 0\,,\ x_n = L\} = X + \ker B\,, \quad \text{avec } B\ :\ x \in \mathbb{R}^{n+1} \longmapsto (x_0, x_n) \in \mathbb{R}^2.$$

La matrice B s'écrit

$$B = \begin{pmatrix} 1 & 0 & \ldots & 0 & 0 \\ 0 & 0 & \ldots & 0 & 1 \end{pmatrix}.$$

D'après ce qui précède, il existe donc $\lambda = (\lambda_0, \lambda_1) \in \mathbb{R}^2$ tel que

$$\nabla J(x) + B^\star \lambda = 0.$$

Écrivons les première et dernière lignes de ce système :

$$\begin{aligned} k(x_0 - x_1) + \lambda_0 &= 0 \\ k(-x_{n-1} + x_n) + \lambda_1 &= 0. \end{aligned}$$

Ces relations expriment l'équilibre des masses extrêmales, et permettent d'interpréter $-\lambda_0$ (resp. $-\lambda_1$) comme la force exercée par le support en 0 sur la masse 0 (resp. par le support en 1 sur la masse n). On peut préciser la configuration minimisante en notant que, pour $i = 1, \dots, n-1$, on a

$$x_{i+1} - x_i = x_i - x_{i-1},$$

de telle sorte que les longueurs des ressorts sont toutes identiques, égales L/n, et ainsi

$$\lambda_0 = -\lambda_1 = kL/n.$$

Cet exemple permet aussi d'illustrer et d'interpréter mécaniquement une méthode très utilisée en pratique, la méthode de pénalisation. Elle consiste à relaxer la contrainte, et à ajouter à la fonctionnelle à minimiser un terme supplémentaire qui pénalise la non vérification des contraintes. Dans l'exemple considéré, elle consiste à considérer la fonctionnelle

$$J_\varepsilon(x) = \frac{1}{2} k \sum_{i=1}^{n} |x_i - x_{i-1}|^2 + \frac{1}{2\varepsilon} \left(|x_0|^2 + |x_n - L|^2 \right).$$

Noter que cela revient à supposer les masses 0 et n attachées à des supports respectivement en 0 et L par des ressorts dont la raideur $1/\varepsilon$ tend vers l'infini.

Remarque 23.32. *Noter que la manière d'écrire les contraintes n'est pas unique. On peut rajouter par exemple $x_n - x_0 = L$. On aura alors un troisième multiplicateur de Lagrange, qui correspondrait à la tension (positive ou négative) au sein d'une barre rigide qui relierait les points extrêmaux. La non unicité met en évidence le fait concret qu'il est a priori impossible de prévoir la tension effective au sein de ce raidisseur, ainsi que l'effort au niveau des supports. Dans la réalité, il peut se produire par exemple que seuls les supports fixes soient actifs, jusqu'à ce que l'un d'entre eux se détériore et finisse par lâcher, pour être relayé par le raidisseur, sans que rien ne transparaisse au niveau de ce que nous appelerons par la suite les variables primales (i.e. les positions des ressorts). On parlera dans un contexte mécanique de situation* hyperstatique *(il y a trop de contrainte), par opposition aux situations* isostatiques *(jeu minimal de contraintes assurant l'unicité des multiplicateurs de Lagrange). On notera qu'il y a un lien fort entre l'expression mathématique d'un ensemble de contraintes et les moyens que l'on pourrait se donner pour les réaliser en pratique.*

L'exemple du pont rigide entre les points extrémaux évoqué plus haut est un peu caricatural car la troisième contrainte est manifestement redondante. Dans des situations plus compliquées pourtant, il peut ne pas être aisé de supprimer des contraintes pour parvenir à un jeu minimal équivalent qui assurera l'unicité des multiplicateurs de Lagrange (comme dans le modèle de prise en compte de la congestion pour les foules, présenté dans la section 3.2, page 39, en lien avec la figure 3.4). D'autre part certains systèmes réels très courants conduisent à une non unicité. Ainsi, pour la chaise à 4 pieds posés sur un sol horizontal, on aura un multiplicateur de Lagrange associé à chacun des 4 contacts avec le sol. Or 3 contacts suffisent pour que la chaise ne rentre pas dans le sol (nous ne considérons pas ici les questions de stabilité). Il est ainsi impossible de prévoir, même si l'on dispose de toutes les informations, quel est l'effort au niveau de chacun des pieds d'une chaise parfaitement équilibrée. Dans la pratique, ces efforts sont susceptibles de changer au cours du temps de façon très irrégulière.

Remarque 23.33. *Cet exemple permet d'illustrer et d'interpréter mécaniquement une méthode très utilisée en pratique, la méthode de pénalisation. Elle consiste à relaxer la contrainte, et à ajouter à la fonctionnelle à minimiser un terme supplémentaire qui pénalise la non vérification des contraintes. Dans l'exemple considéré, elle consiste à considérer la fonctionnelle*

$$J_\varepsilon(x) = \frac{1}{2}k\sum_{i=1}^{n}|x_i - x_{i-1}|^2 + \frac{1}{2\varepsilon}\left(|x_0|^2 + |x_n - L|^2\right).$$

Noter que cela revient à supposer les masses 0 *et* n *attachées à des supports respectivement en* 0 *et* L *par des ressorts dont la raideur* $1/\varepsilon$ *tend vers l'infini.*

Exercice 23.2. On considère un "agent" à qui est offerte la possibilité d'acquérir des biens 1,...,n. Les biens sont caractérisés par des fonctions d'*utilité* $p \mapsto u_j(p)$ qui quantifient la satisfaction qu'il retire en consacrant la part p de son capital à l'achat de biens de type j. On considère qu'il dispose d'un capital P, et qu'il cherche à maximiser sa satisfaction maximale

$$\max \sum u_j(p_j)\,,\ \sum_{j=1}^{n} p_j \leq P.$$

(On pourra intégrer la possibilité de conserver une partie de son capital en définissant un bien "vide" qui correspond à l'absence d'achat, ou tout du moins à la préservation d'une partie du capital.) Faire l'étude de ce problème d'optimisation.

On pourra notamment étudier le cas où les fonctions d'utilité sont concaves régulières croissantes sur $[0, +\infty[$, nulles en 0, par exemple $u_j(p) = \alpha_j \log(1 + p)$, et étudier comment la stratégie optimale varie en fonction de P.

A Compléments théoriques

A.1 Calcul différentiel, formules d'intégration par parties

On rappelle ici quelques formules d'intégration par partie. On supposera tous les champs réguliers . L'extension de ces formules à des champs scalaires ou vectoriel moins réguliers doit faire l'objet d'une vérification qui n'est pas traitée ici.

Soit $u = (u_1, u_2)^T$ un champ de vecteur. Sa divergence est

$$\nabla \cdot u = \begin{pmatrix} \dfrac{\partial}{\partial x_1} \\ \dfrac{\partial}{\partial x_2} \end{pmatrix} \cdot \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \frac{\partial u_1}{\partial x_1} + \frac{\partial u_2}{\partial x_2}.$$

Soit $u = (u_1, u_2)^T$ un champ de vecteur, son gradient est la matrice

$$\nabla u = \begin{pmatrix} \dfrac{\partial u_1}{\partial x_1} & \dfrac{\partial u_1}{\partial x_2} \\ \dfrac{\partial u_2}{\partial x_1} & \dfrac{\partial u_2}{\partial x_2} \end{pmatrix}$$

Pour tout vecteur n, on a

$$\nabla u \cdot n = \begin{pmatrix} \dfrac{\partial u_1}{\partial x_1} & \dfrac{\partial u_1}{\partial x_2} \\ \dfrac{\partial u_2}{\partial x_1} & \dfrac{\partial u_2}{\partial x_2} \end{pmatrix} \begin{pmatrix} n_1 \\ n_2 \end{pmatrix} = \begin{pmatrix} \dfrac{\partial u_1}{\partial x_1} n_1 + \dfrac{\partial u_1}{\partial x_2} n_2 \\ \dfrac{\partial u_2}{\partial x_1} n_1 + \dfrac{\partial u_2}{\partial x_2} n_2 \end{pmatrix},$$

qui est la dérivée de u dans la direction n, de telle sorte que

$$u(x + \varepsilon n) = u(x) + \varepsilon \nabla u \cdot n + o(\varepsilon).$$

Si n est un vecteur unitaire [115], on écrit $\nabla u \cdot n = \partial u / \partial n$.

Soit u un champ de vecteur. Son Laplacien Δu est le vecteur

$$\Delta u = \begin{pmatrix} \Delta u_1 \\ \Delta u_2 \end{pmatrix}.$$

Pour $A = (a_{ij})$ et $B = (b_{ij})$ des matrices, $A : B$ représente le scalaire

$$A : B = \sum_{i,j} a_{ij} b_{ij}.$$

115. Cette hypothèse reflète le caractère assez peu naturel de cette notation. C'est un peu comme si, pour une fonction $x \longmapsto f(x)$, avec $x = (x_1, x_2) = x_1 e_1 + x_2 e_2 \in \mathbb{R}^2$, on écrivait $\partial f / \partial e_1$ la dérivée de f par rapport à x_1. Pour pousser plus loin cette remarque, précisons qu'il existe une situation dans laquelle cette notation serait justifiée, mais pour désigner quelque chose de très différent à l'usage. On considère une partie de $\mathbb{R}^d$, strictement convexe au sens où tout point de la frontière est extrémal, et une fonction définie sur cette frontière §que l'on suppose régulière, même si cela n'est pas vraiment nécessaire). Du fait de la stricte convexité, si l'on se donne un vecteur unitaire , il existe un unique point de la frontière tel que la normale en se point corresponde à ce vecteur, on peut donc écrire la fonction comme une fonction de n, et considérer la différentielle de f par rapport à n.

Noter que $|A| = (A : B)^{1/2}$ est une norme euclidienne sur l'espace des matrices (appelée norme de *Frobenius*). Pour u et v deux champ de vecteurs

$$\nabla u : \nabla v = \begin{pmatrix} \dfrac{\partial u_1}{\partial x_1} & \dfrac{\partial u_1}{\partial x_2} \\ \dfrac{\partial u_2}{\partial x_1} & \dfrac{\partial u_2}{\partial x_2} \end{pmatrix} : \begin{pmatrix} \dfrac{\partial v_1}{\partial x_1} & \dfrac{\partial v_1}{\partial x_2} \\ \dfrac{\partial v_2}{\partial x_1} & \dfrac{\partial v_2}{\partial x_2} \end{pmatrix} = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{\partial u_i}{\partial x_j} \frac{\partial v_i}{\partial x_j}.$$

La notation $|\nabla u|^2$ est utilisée pour désigner $\nabla u : \nabla u$.

Soit σ un champ de matrices (ou de tenseurs). Sa divergence est un vecteur, dont chaque composante est la ligne de la matrice correspondante

$$\nabla \cdot \sigma = \nabla \cdot \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} = \begin{pmatrix} \dfrac{\partial \sigma_{11}}{\partial x_1} + \dfrac{\partial \sigma_{12}}{\partial x_2} \\ \dfrac{\partial \sigma_{21}}{\partial x_1} + \dfrac{\partial \sigma_{22}}{\partial x_2} \end{pmatrix}$$

Soit $u = (u_1, u_2)^T$ un champ de vecteur, on note $u \otimes u$ la matrice $(u_i\, u_j)_{i,j}$.

Si $\nabla \cdot u = 0$, on a

$$\nabla \cdot (u \otimes u) = (u \cdot \nabla)\, u = \begin{pmatrix} u_1 \dfrac{\partial u_1}{\partial x_1} + u_2 \dfrac{\partial u_1}{\partial x_2} \\ u_1 \dfrac{\partial u_2}{\partial x_1} + u_2 \dfrac{\partial u_2}{\partial x_2} \end{pmatrix}$$

Toujours sous la condition $\nabla \cdot u = 0$,

$$(\nabla \cdot (u \otimes u)) \cdot u = ((u \cdot \nabla)\, u) \cdot u = \nabla \cdot \left(\frac{|u|^2}{2} u \right).$$

Si $\nabla \cdot u = 0$, alors

$$\nabla \cdot {}^t\nabla u = 0.$$

En conséquence, si $\nabla \cdot u = 0$, alors

$$\nabla \cdot \left(\nabla u + {}^t\nabla u \right) = \nabla \cdot \nabla u = \Delta u.$$

Intégration par parties

Soit v un champ de vecteurs. on a

$$\int_\Omega \nabla \cdot v = \int_\Gamma v \cdot n \tag{A.1}$$

Soit σ un champ de matrices. on a

$$\int_\Omega \nabla \cdot \sigma = \int_\Gamma \sigma \cdot n \tag{A.2}$$

Soit q un champ scalaire. On a

$$\int_\Omega \nabla \mathrm{q} = \int_\Gamma \mathrm{q} n. \tag{A.3}$$

Soit v un champ de vecteurs, et q un champ scalaire. On a

$$\int_\Omega \mathrm{q}\, \nabla \cdot u + \int_\Omega u \cdot \nabla \mathrm{q} = \int_\Gamma \mathrm{q} u \cdot n. \tag{A.4}$$

Soient u et v des champs scalaires. On a

$$\int_\Omega v \Delta u = \int_\Omega \nabla u \cdot \nabla v + \int_\Gamma v \frac{\partial u}{\partial n}, \tag{A.5}$$

où n est la normale sortante au domaine.

Soit u un champ de vecteurs, et q un champ scalaire.

$$\int_\Omega \mathrm{q}\, \nabla \cdot u + \int_\Omega u \cdot \nabla \mathrm{q} = \int_\Gamma \mathrm{q} u \cdot n. \tag{A.6}$$

Soient u et v des champs de vecteurs. On a

$$\int_\Omega \Delta u \cdot v + \int_\Omega \nabla u : \nabla v = \int_\Gamma v \cdot \frac{\partial u}{\partial n}. \tag{A.7}$$

Si en outre $\nabla \cdot u = 0$, on a

$$0 + \int_\Omega {}^t\nabla u : \nabla v = \int_\Gamma v \cdot \left({}^t\nabla u \cdot n \right) \tag{A.8}$$

En conséquence, si $\nabla \cdot u = 0$, alors

$$\int_\Omega \Delta u \cdot v + \int_\Omega \nabla u : (\nabla v + {}^t\nabla v) = \int_\Gamma v \cdot \left(\nabla u + {}^t\nabla u \right) \cdot n. \tag{A.9}$$

Pour tous champs vectoriels u et v, on a

$$\int_\Omega \nabla u : {}^t\nabla v = \int_\Omega {}^t\nabla u : \nabla v, \tag{A.10}$$

de telle sorte que

$$\int_\Omega \nabla u : (\nabla v + {}^t\nabla v) = \frac{1}{2} \int_\Omega (\nabla u + {}^t\nabla u) : (\nabla v + {}^t\nabla v) \tag{A.11}$$

Dérivation d'une intégrale sur un domaine en mouvement

Soit ω un système matériel advecté par le champ de vitesse $u(x,t)$, et $F(x,t)$ une fonction scalaire. On a

$$\frac{d}{dt} \int_{\omega(t)} F(x,t) = \int_{\omega(t)} \frac{\partial F}{\partial t} - \int_{\partial\omega(t)} F(x,t) u \cdot n. \tag{A.12}$$

Proposition A.1. *Soient u et v deux champs de vecteurs réguliers définis sur Ω. On suppose que u est à divergence nulle. On a alors*

$$0 = -\int_\omega {}^t\nabla u : \nabla v + \int_{\partial\omega} v \cdot \left({}^t\nabla u \cdot n\right)$$

Démonstration. On écrit

$$\begin{aligned}\int_{\partial\omega} v \cdot \left({}^t\nabla u \cdot n\right) &= \int_{\partial\omega} n \cdot (\nabla u \cdot v) = \int_\omega \nabla \cdot (\nabla u \cdot v) \\ &= \sum_i \partial_i \sum_j v_j \partial_j u_i = \sum_i \sum_j \partial_i v_j \partial_j u_i + \sum_j v_j \partial_j \sum_i \partial_i u_i.\end{aligned}$$

Le second terme ci-dessus est nul car u est à divergence nulle, d'où l'on déduit l'identité annoncée. □

Proposition A.2. *Soient u et : v deux champs réguliers sur Ω. On a*

$$\int_\Omega \nabla u : {}^t\nabla v = \int_\Omega (\nabla \cdot u)(\nabla \cdot v) + \int_\Gamma (\nabla \cdot u) v \cdot n - \int_\Gamma (\nabla u \cdot v) \cdot n$$

Démonstration: On a

$$\begin{aligned}\int_\Gamma (\nabla u \cdot v) \cdot n &= \int_\Omega \nabla \cdot (\nabla u \cdot v) \\ &= \int_\Omega \sum_i \partial_i \sum_j v_j \partial_j u_i \\ &= \int_\Omega \sum_i \sum_j ((\partial_i \partial j u_i) v_j + \partial_j u_i \partial_i v_j) \\ &= \int_\Omega v (\nabla \nabla \cdot u) + \int_\Omega \nabla u : {}^t\nabla v \\ &= \int_\Omega (\nabla \cdot u) v \cdot n - \int_\Omega (\nabla \cdot u)(\nabla \cdot v) + \int_\Omega \nabla u : {}^t\nabla v\end{aligned}$$

A.2 Cercles de Gerchgorin

Definition A.3. *Une matrice $A = (a_{ij}) \in \mathcal{M}_n(\mathbb{C})$ est dite à diagonale strictement dominante si*

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}| \quad \forall i = 1, \ldots, n.$$

Proposition A.4. *(Gerschgorin)*
Soit $A = (a_{ij}) \in \mathcal{M}_n(\mathbb{C})$. Soit Sp(A) *l'ensemble des valeurs propres de A. On a*

$$\mathrm{Sp(A)} \subset \bigcup_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{D(a_{ii}, r_i)}\,,\ \mathrm{r_i} = \sum_{\mathrm{j \neq i}} \mathrm{|a_{ij}|}\ ,$$

où $D(a, r) \subset \mathbb{C}^2$ désigne le disque fermé de centre a et de rayon r.

A.3 Chaines de Markov

Soit V un ensemble fini et $K(\,\cdot\,,\,\cdot\,) \in \mathbb{R}_+^{V\times V}$ tel que

$$\sum_{y\in V} K(x,y) = 1 \quad \forall x \in V.$$

En numérotant les points de V : 1, 2, ..., N, on peut voir K comme une matrice de $\mathcal{M}_N([0,1])$. La somme des éléments de chaque ligne vaut 1, une telle matrice est dite *stochastique*.

Definition A.5. *(Chaîne de Markov)*
On appelle chaîne de Markov associée à K une suite de variables aléatoires X_0, X_1, ...$\in V$, avec

$$\begin{aligned}\mathbb{P}(X_{n+1} = y | X_n = x, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0) &= \mathbb{P}(X_{n+1} = y | X_n = x) \\ &= \mathbb{P}(X_1 = y | X_0 = x) = K(x,y).\end{aligned}$$

La matrice ${}^tK$ peut être interprétée de la façon suivante : si l'on considère une variable aléatoire X_0 suivant la loi $p = {}^t(p_1, p_2, \ldots, p_n)$, que l'on note X_1 la variable obtenue après un pas construit suivant les probabilités de transition définie ci-dessus, alors

$$\mathbb{P}(X_1 = i) = \sum_j K(j,i)\mathbb{P}(X_0 = j) = \sum_j K(j,i)p_j,$$

c'est à dire que X_1 suit la loi $q = {}^t(q_1, q_2, \ldots, q_n)$, avec

$$q = {}^tKp.$$

Noter que toute application de l'ensemble à N éléments vers lui même peut être représentée par une matrice qui ne contient que des 0 et des 1, avec exactement un "1" par colonne. La transposée d'une matrice stochastique est ainsi combinaison convexe de telles matrices : on peut voir toute chaine de Markov sur un espace à n états comme la généralisation d'une application qui a un point associe un point : chaque point (considéré comme étant de masse unitaire) peut être distribué sur plusieurs autres points, de façon à ce que la masse soit conservée. Cette matrice exprime ainsi un transport de mesure : si l'on se donne une mesure de probabilité sur l'ensemble à n points, $p = {}^t(p_1, p_2, \ldots, p_n)$, la mesure ${}^tKp$ est la mesure image (ou *push-forward*) de p par le transport.

Definition A.6. *(Irréductibilité)*
On dit que la chaîne de Markov est irréductible si, pour tous x, y, il existe n et m tels que

$$\mathbb{P}(X_n = y | X_0 = x) > 0 \text{ et } \mathbb{P}(X_m = x | X_0 = y) > 0.$$

Definition A.7. *(Mesure stationnaire)*
Une mesure π sur V est dite stationnaire pour la chaine K si

$${}^tK\pi = \pi.$$

Si l'on se place dans le cas où le point initial X_0 suit la loi associée à π, alors X_1 suite la même loi, ainsi que tous les X_n (sans bien sûr que que les X_n soient indépendants).

Théorème A.8. *(Perron-Frobenius)*
Soit K la matrice de transition d'une chaîne de Markov irréductible. Alors toutes les valeurs propres de K sont de module inférieur ou égal à 1, 1 est valeur propre de ${}^tK$, et c'est une valeur propre simple. Elle admet pour vecteur propre une mesure π sur V, avec $p(x) > 0$ pour tout $x \in V$, qui se trouve de fait être l'unique mesure stationnaire.

Noter que, dans le théorème précédent, il peut exister d'autres valeurs propres de module égal à 1. L'unicité de 1 comme v.p. de plus grand module est en revanche assurée si l'on suppose la matrice *primitive*, i.e. qu'il existe k t.q. A^k a tous ses coefficients strictement positifs.

Definition A.9. *(Réversibilité)*
Une chaîne de Markov K irréductible est dite réversible si sa mesure stationnaire vérifie

$$K(x,y)\pi(x) = K(y,x)\pi(y).$$

A.4 Spectre du Laplacien discret

La matrice

$$A = \begin{pmatrix} 2 & -1 & 0 & \cdot & \cdot & 0 \\ -1 & 2 & -1 & 0 & \cdot & \cdot \\ 0 & -1 & \cdot & \cdot & & \cdot \\ \cdot & & \cdot & \cdot & \cdot & \cdot \\ \cdot & & & \cdot & 2 & -1 \\ 0 & \cdot & \cdot & 0 & -1 & 2 \end{pmatrix} \in \mathcal{M}_{N-1}(\mathbb{R}) \tag{A.13}$$

possède $N-1$ valeurs propres distinctes

$$\lambda_k = 4\sin^2\left(\frac{k\pi}{2N}\right),\ k = 1,\ \ldots,\ N-1.$$

Le vecteur propre associé à la valeur propre λ_k s'écrit

$$u_k = {}^t\left(\sin\left(\frac{k\pi}{N}\right),\ \sin\left(\frac{2k\pi}{N}\right),\ \ldots,\ \sin\left(\frac{(N-1)k\pi}{N}\right)\right).$$

Références

[1] G. Allaire, Analyse numérique et optimisation, Publications Ecole Polytechnique, No 15, Ellipses Paris, 2005.

[2] H. Brezis, *Analyse Fonctionnelle, Théorie et Applications*, Masson 1983.

[3] H. Brezis, *Opérateurs maximaux monotones et semi–groupes de contraction dans les espaces de Hilbert*, North Holland publishing company 1973.

[4] V. Girault, P.A. Raviart, *Finite Element Methods for Navier-Stokes Equations- Theory and Algorithms* Springer Verlag, Berlin, 1986.

[5] B. Maury, *Analyse Fonctionnelle, exercices et problèmes corrigés*, Ellipses, Paris, 2004.

[6] P.-A. Raviart, J.M. Thomas, Introduction à l'Analyse Numérique des Équations aux Dérivées Partielles, Masson, Paris, 1983.

[7] F. Santambrogio, Optimal Transport for Applied Mathematicians, Progress in Nonlinear Differential Equations and Their Applications, Vol. 87, Birkhäuser Basel, 2015.

[8] C. Villani, Topics in optimal transportation, American Mathematical Soc, Vol. 58, 2003.