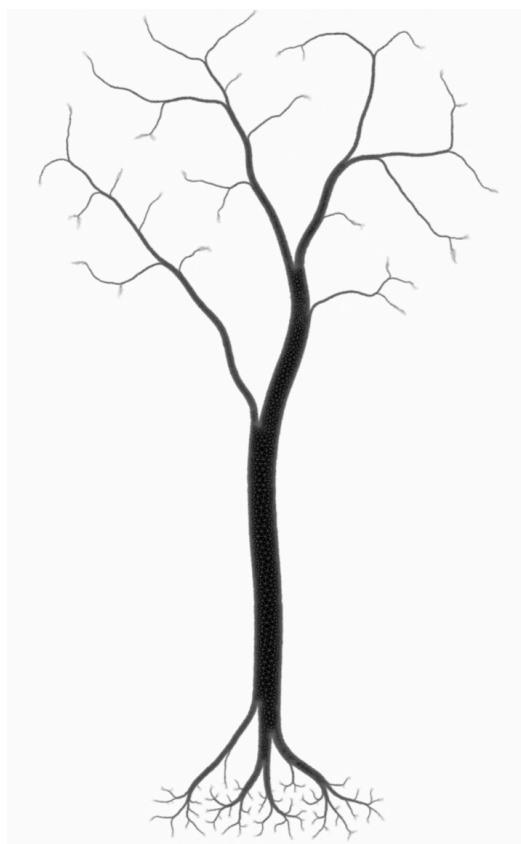


MODÉLISATION MATHÉMATIQUE



DMA, ENS PSL

B. MAURY

Table des matières

I Modélisation	7
1 Fondamentaux & notions	8
1.1 Lagrangien <i>vs.</i> eulérien	8
1.2 Grandeurs intensives et extensives	17
1.3 La notion de mesure, point de vue de la modélisation	20
1.4 Entropie d'une variable aléatoire discrète	25
1.5 Éléments de mécanique du point matériel et des systèmes	27
1.6 Éléments de thermodynamique, notion d'énergie	31
1.7 Exercices	34
2 Réseaux résistifs	37
2.1 Modèles, motivations	37
2.2 Cadre formel, problème de Laplace discret	39
2.3 Opérateurs discrets : gradient, divergence, et laplacien	45
2.4 Résistance équivalente d'un réseau	48
2.5 Cadre stochastique	51
2.6 Cadre abstrait et modélisation	55
2.7 Squelette métrique associé à un réseau résistif	57
2.8 Plongement dans l'espace euclidien	58
2.9 Premier pas vers le transport branché	59
2.10 Réseaux infinis	60
2.11 Remarques diverses	61
2.12 Exercices	62
3 Le poumon humain vu comme arbre résistif	65
3.1 Vue d'ensemble de l'appareil respiratoire humain	65
3.2 Modèle tuyau-ballon	66
3.3 Le poumon comme arbre résistif	69
3.4 Numérotation dyadique	70
3.5 Arbre résistif non symétrique	72
3.6 Arbre optimal ?	76
3.7 Vers un poumon infini	77
3.8 Particules et dépôt	78
4 Lois de conservation, transport et diffusion	83
4.1 Vecteur flux, équation de conservation	84
4.2 Transport	85
4.3 Diffusion	89
4.4 Transport - diffusion	95
4.5 Exercices	97
5 Modèles en vrac	99
5.1 Bilan radiatif	99
5.2 Montée / descente d'une rame de RER un jour de grève	101
5.3 Aérosols & concentration de CO ₂	102
5.4 Qualité de l'air dans un bâtiment	103

5.5	Inertie thermique & capacité thermique apparente	108
5.6	Modélisation de la survie d'un dialecte au sein d'une population	111
5.7	Positionnement de postes de secours	111
5.8	Affluence au supermarché	113
5.9	Densités de population et contacts induits	114
5.10	Modèles de mobilité sur réseaux	115
5.11	Cadre général, problématiques	116
5.12	Exercices	119
5.13	Mobilité individuelle	121
6	Propagation d'opinion sur réseaux sociaux	122
6.1	Modèle d'évolution	122
6.2	Cadre stochastique	128
6.3	Liens avec les schémas de discréttisation des EDP	131
6.4	Monitoring of a network though influencers / influence coefficients	132
6.5	Modèle continu et structure de flot de gradient	135
6.6	Réseaux charismatiques	140
6.7	Niveau de certitude	143
6.8	Clivage des opinions au sein d'une population	145
6.9	Propagation sur des réseaux du type "application"	146
6.10	Exercices	148
7	Modèles de propagation d'épidémies	149
7.1	Modèle SIR	149
7.2	Modèle stochastique orienté agent	154
7.3	Modèle déterministe portant sur des probabilités d'infection	155
7.4	Développements, extensions	162
7.5	Prise en compte de la perte d'immunité	163
7.6	Exercices	169
8	Modèles de trafic routier ou piéton	170
8.1	Le modèle FTL	170
8.2	Modèles d'ordre 2	184
8.3	Modèles granulaires de foules	191
8.4	Modèles macroscopiques de trafic routier	197
8.5	Modèles granulaires de foules	200
9	Éléments de mécanique des fluides	206
9.1	Tenseur des contraintes, équation générale du mouvement	206
9.2	Fluides parfaits	208
9.3	Fluides newtoniens	211
9.4	Bilan d'énergie pour les équations de Navier-Stokes	213
9.5	Écoulements en milieu poreux	215
9.6	Cadre mathématique pour le problème de Darcy	216
9.7	Cadre mathématique pour les équations de Stokes	217
9.8	Ecoulement de Poiseuille, notion de résistance	219
9.9	Ecoulement autour d'une sphère	221
II	Fondamentaux	222
10	Graphes	223
10.1	Définitions	223
10.2	Laplacien(s)	229
10.3	Exercices	231
11	Équations différentielles ordinaires	235
11.1	Théorème de Cauchy–Lipschitz	235

11.2 Points d'équilibre, stabilité	239
11.3 Compléments	241
11.4 Exercices	244
12 Espaces de Sobolev	247
12.1 Définitions, propriétés générales	247
12.2 Traces	249
12.3 Injections	253
12.4 Inégalités de Poincaré	253
12.5 Problèmes aux limites elliptiques	255
12.6 Espaces de Sobolev et transformation de Fourier	258
13 Éléments d'optimisation	260
13.1 Éléments d'analyse convexe	260
13.2 Existence d'un minimiseur, conditions d'optimalité	261
13.3 Contraintes unilatérales	264
13.4 Point-selle, théorème de Kuhn et Tucker	269
13.5 Formalisme de l'analyse non lisse, sous-différentiels	272
13.6 Dérivation du minimum par rapport aux contraintes	277
13.7 Cas de la dimension infinie	278
13.8 Contraintes non linéaires d'égalité	279
13.9 Illustrations	281
13.10 Exercices	283
14 Transport optimal discret	286
14.1 Problème d'affectation et problème de Monge Kantorovich discret	286
14.2 Formulation duale du problème de MK discret	289
14.3 Exemples d'applications	291
14.4 Métriques induites sur l'ensemble des mesures de probabilités sur un espace métrique fini	292
14.5 Métriques induites sur l'ensemble des mesures atomiques	293
14.6 Complétion de l'espace de Wasserstein discret	295
14.7 Régularisation entropique	296
14.8 Calcul effectif par Régularisation entropique	298
14.9 Calcul effectif par l'algorithme des enchères	300
14.10 Compléments, extensions	303
14.11 Compléments sur la régularisation entropique	305
14.12 Exercices	309
15 Analyse de sensibilité	311
15.1 Introduction	311
15.2 Sensibilité pour les problèmes d'optimisation	312
15.3 Méthode de l'état adjoint	315
15.4 Exercices	321
16 Méthode des différences finies	324
16.1 La méthode	324
16.2 Consistance, stabilité, convergence	326
16.3 Analyse des principaux schémas numériques	330
16.4 Symboles discret et continu des opérateurs différentiels	332
16.5 Interprétation probabiliste de schémas explicites	336
16.6 Extensions, développements	339
16.7 Implémentation effective	339
16.8 Exercices	342
17 Méthode des éléments finis	344
17.1 Formulation variationnelle du problème de Poisson	344
17.2 Méthode des éléments finis	349
17.3 Estimation d'erreur pour la méthode des Éléments Finis	351

17.4	Éléments finis et réseaux résistifs	356
18	Espaces de Hilbert	358
18.1	Définitions, principales propriétés	358
18.2	Convergence faible	364
18.3	Somme Hilbertienne, bases Hilbertiennes	366
18.4	Décomposition spectrale des opérateur auto-adjoints compacts	367
18.5	Problèmes d'évolution	370
18.6	Minimisation de fonctionnelles convexes	371
19	Compléments	374
19.1	Inégalités	374
19.2	Théorème d'Arzela Ascoli	375
19.3	Théorèmes de point fixe	375
19.4	Théorème de Krein-Milman	375
19.5	Théorèmes des fonctions implicites et d'inversion locale	376
19.6	Convergence faible et compacité	379
19.7	Espaces de Sobolev et traces : le point de vue de la modélisation	382
19.8	Introduction aux flots de gradient dans l'espace de Wasserstein	388
19.9	Calcul différentiel, formules d'intégration par parties	391
19.10	Cercles de Gerchgorin	395
19.11	Spectre du Laplacien discret	395
19.12	Théorème spectral généralisé	395
19.13	Entiers p -adiques, espaces ultra-métriques	396
19.14	Distance de Gromov-Wasserstein	404
19.15	Dendrogrammes	405
20	Problèmes	408
20.1	Conditions aux limites de Robin sur graphe	408
20.2	Propagation de la chaleur sur un réseau	411
20.3	Résistance de l'hypercube	413
20.4	Clivage	414
20.5	Charisme	416
20.6	Stabilité du poumon humain	418
20.7	Poumon non linéaire	420
20.8	Phénomène de limitation du débit expiratoire	423
20.9	Mobilité et équilibres de Wardrop	424
20.10	Optimisation d'une préparation de concours	426
20.11	Modèles de foules de type Nash	428
20.12	Mouvement de véhicules autonomes	431
20.13	Transport partiel	432
20.14	Transport sous contraintes	434
20.15	Décomposition polaire discrète	435
20.16	Entropie relative	436
20.17	Décroissance de l'entropie pour les schémas de différences finies	439
20.18	Flots de gradients discrets dans l'espace de Wasserstein, équation de la chaleur comme flot de gradient de l'entropie	440

Première partie

Modélisation

Chapitre 1

Fondamentaux & notions

Sommaire

1.1	Lagrangien <i>vs.</i> eulérien	8
1.1.1	La vision lagrangienne, données et modèles	8
1.1.2	La vision eulérienne	10
1.1.3	Transport & conservation : le point de vue eulérien	12
1.1.4	Au delà de l'espace physique	14
1.1.5	Dérivées lagrangienne et eulérienne	17
1.2	Grandeur intensives et extensives	17
1.3	La notion de mesure, point de vue de la modélisation	20
1.4	Entropie d'une variable aléatoire discrète	25
1.5	Éléments de mécanique du point matériel et des systèmes	27
1.6	Éléments de thermodynamique, notion d'énergie	31
1.7	Exercices	34

1.1 Lagrangien *vs.* eulérien

Les qualificatifs lagrangien et eulérien font référence à deux manières d'appréhender des phénomènes de transport, ou de mouvement, au sens le plus général de ces termes¹.

Ils sont couramment utilisés en mécanique des fluides, mais peuvent s'appliquer dans tous les domaines qui impliquent des entités évoluant dans un espace (espace physique ou espace plus abstrait représentant des caractéristiques de ces entités autres que leur position). Ils permettent de qualifier à la fois la manière dont peuvent être recueillies des données de positions ou d'état des entités considérées, et l'approche adoptée pour écrire des modèles visant à décrire ces phénomènes de mouvement.

1.1.1 La vision lagrangienne, données et modèles

Dans le contexte de la récolte de données, en vue de récupérer des informations sur le mouvement d'entités, le caractère lagrangien signifie que l'on suit les entités dans leur mouvement. S'il s'agit par exemple de personnes qui se déplacent dans un territoire, les données GPS récupérées à partir d'un téléphone portable sont lagrangiennes : même si le droit ne permet pas de les stocker accompagnées

1. Nous verrons qu'il est possible de considérer des populations structurées par autre chose que la position dans l'espace physique, et que l'on peut étendre la notion de transport à une évolution de "proche en proche", chez les entités considérées, pour la variable structurante. À titre d'exemple, comme il est détaillé plus loin, le vieillissement que subit tout mortel peut se représenter comme un transport le long de la flèche du temps, à la vitesse de 1 seconde par seconde.

de l'identité de la personne suivie, un identifiant anonymisé permet d'associer les positions successives à un même individu.

Selon ce même principe, maintenant instancié dans le domaine de la *modélisation*, l'approche lagrangienne consiste à suivre des entités dans leur mouvement. Si l'on considère par exemple un ensemble de N particules évoluant dans l'espace physique \mathbb{R}^d , la donnée de leurs positions au cours du temps $t \mapsto x(t) = (x_i(t))$, avec $x_i(t) \in \mathbb{R}^d$, est lagrangienne.

Pour i fixé, l'ensemble des positions $x_i(t)$ est la *trajectoire* de la particule i . La dérivée en temps $\dot{x}_i(t)$ est la vitesse de la particule i au temps t . La forme lagrangienne de l'équation décrivant le transport de particules se réduit à la définition même de la dérivée, on écrit simplement que la dérivée de la position x_i est égale à la vitesse u_i :

$$\frac{dx_i}{dt} = u_i \quad i = 1, \dots, N. \quad (1.1)$$

La collection des vitesses (u_i) est elle-même de nature lagrangienne. Il peut sembler artificiel de considérer cette collection d'identités comme un système d'équations, du fait que, si l'on se donne les vitesses au cours du temps, on obtient les positions par simple intégration :

$$x_i(t) = x_i(0) = \int_0^t u_i(s) ds,$$

mais nous verrons que la description *eulérienne* de ce même phénomène de transport conduit à une équation délicate à formaliser.

Système de particules en interaction

Un modèle lagrangien archétypal est le système d'équation exprimant le principe fondamental de la dynamique (second principe de Newton, d'essence lagrangienne, puisqu'il s'applique à un système matériel) en mécanique classique, pour des particules en interaction. Ce système s'écrit, si l'on note m_i la masse de la particule i , et f_{ji} la force exercée par j sur i ,

$$m_i \frac{d^2 u_i}{dt^2} = \sum_{j \neq i} f_{ji} \quad \forall i = 1, \dots, N.$$

On se reportera à la section 1.5 pour une présentation plus complète de ce type de modèles.

Modèles de trafic

Il est également courant de modéliser de façon lagrangienne des entités vivantes comme des personnes. Le modèle microscopique le plus simple de trafic routier est ainsi basé sur le suivi des positions de N véhicules sur l'axe des réels, $x = (x_1, \dots, x_N) \in \mathbb{R}^N$. Considérer que la vitesse instantanée d'un conducteur ne dépend que de la distance au véhicule précédent conduit au modèle

$$\dot{x}_j = \varphi(x_{j+1} - x_j) \quad j = 1, \dots, N.$$

La fonction φ , définie de \mathbb{R}_+ dans \mathbb{R}_+ , encode le comportement des conducteurs. Il est naturel de considérer que φ s'annule en-dessous d'une valeur-seuil w_{min} correspondant à la taille des véhicules, qu'elle est croissante, et majorée par une vitesse limite U , par exemple

$$\varphi(w) = U \left(1 - \exp \left(\frac{(w - w_{min})_+}{w_s} \right) \right).$$

On notera que ce modèle ne suit pas le principe d'action-réaction, au sens où les interactions entre les conducteurs sont asymétriques : i est influencé par $i + 1$, qui n'est pas influencé par i . Une étude détaillée de modèles de ce type fait l'objet du chapitre 8.

Remarque 1.1. La description lagrangienne des entités permet d'adapter le modèle de comportement aux individus. On peut ainsi considérer que la vitesse limite dépend de i (U remplacé par U_i ci-dessus), ainsi que la fonction de comportement $\varphi = \varphi_i$.

Exercice 1.1. Préciser les limites du modèle ci-dessus, et proposer des extensions plus réalistes. On cherchera en particulier à élaborer un modèle d'ordre 2 en temps.

Le caractère lagrangien du modèle permet une prise en compte fine des *personalités* des acteurs impliqués. On peut par exemple considérer que chaque conducteur i est caractérisé par une fonction φ_i qui lui est propre, ce qui permet de modéliser l'effet de la coexistence de conducteurs plus ou moins prudents.

1.1.2 La vision eulérienne

L'approche eulérienne est basée sur l'observation d'une ou plusieurs zones de l'espace dans lequel le mouvement s'effectue, et à suivre "ce qui se passe" dans cette zone, sans prise en compte de l'identité des entités observées. En termes de récolte de données, on pourra penser à la mesure au cours du temps du taux d'occupation d'un espace déterminé (une salle, un bâtiment, une gare ...). Ou, pour parler de données plus localisées, l'information récoltée par un capteur de flux installé au niveau d'un accès, qui va compter les passages dans un sens et dans l'autre. Considérons par exemple une ligne de train dont toutes les gares sont équipées de tels compteurs. La connaissance de l'ensembles des nombres de voyageurs qui sont montés et descendus à chaque arrêt permet d'estimer l'affluence dans les trains à chaque instant, mais cette information n'est pas lagrangienne, elle ne permet pas, sauf dans des cas très dégénérés, d'inférer les trajets effectués par les voyageurs (on pourra se reporter à l'exercice 5.1, page 119 pour une analyse plus poussée de l'impossibilité de retrouver les parcours effectifs des personnes). À titre d'illustration, la figure 1.1 propose une représentation de données eulériennes sur Paris (en 1889), flux selon les différents moyens de transport (train, tramway et bateau sur la Seine), ainsi que des données d'affluence pour les différentes gares.

Dans le contexte de la modélisation, l'approche eulérienne consiste à introduire des quantités variables en temps afférente à une position (ou une zone) fixe de l'espace. Il est frappant de constater que la situation la plus simple envisagée dans le paragraphe précédent, à savoir le suivi du mouvement de N particules dans l'espace physique, est assez délicat à écrire de façon eulérienne. Cela est néanmoins possible si l'on utilise la notion de masse ponctuelle : à une particule située en un point x de l'espace, on associe la *mesure* δ_x , qui est une application de l'ensemble des parties de l'espace dans \mathbb{R}_+ (ou simplement \mathbb{N} si l'on se limite à compter des entités). Pour $A \subset \mathbb{R}^d$, on dit que $\delta_x(A)$ est égal à 1 si x est dans A , à 0 sinon. On peut sommer ces objets pour obtenir une *mesure* qui encode l'ensemble des positions des particules au temps t

$$\mu_t = \sum_{i=1}^N \delta_{x_i(t)}.$$

Cette identité exprime² que, pour tout ensemble A , $\mu_t(A)$ est le nombre de personnes qui se trouvent dans A à l'instant t .

Le fait que 2 particules puissent occuper la même position ne pose pas de problème, on a une règle simple de sommation : $\delta_x + \delta_x = 2\delta_x$, qui est une masse ponctuelle affectée d'un poids 2.

On prendra en revanche garde au fait que, si μ_t est univoquement définie à partir des $x_i(t)$, l'information contenue dans μ est dégradée par rapport à la collection des x_i , puisque les labels ont été *perdus* dans la sommation : on ne sait pas qui est où. Plus précisément, si les positions sont distinctes deux à deux, une même mesure μ_t correspond à un grand nombre de distributions possibles des entités. De façon plus formelle, pour toute permutation $\varphi \in S_N$ (groupe des bijections sur l'ensemble à N

2. Cette identité exprime un passage du lagrangien vers l'eulérien, sa simplicité est liée au fait que cette opération est toujours simple, puisqu'elle consiste à *dégrader* l'information que l'on a sur la population, comme précisé plus loin.

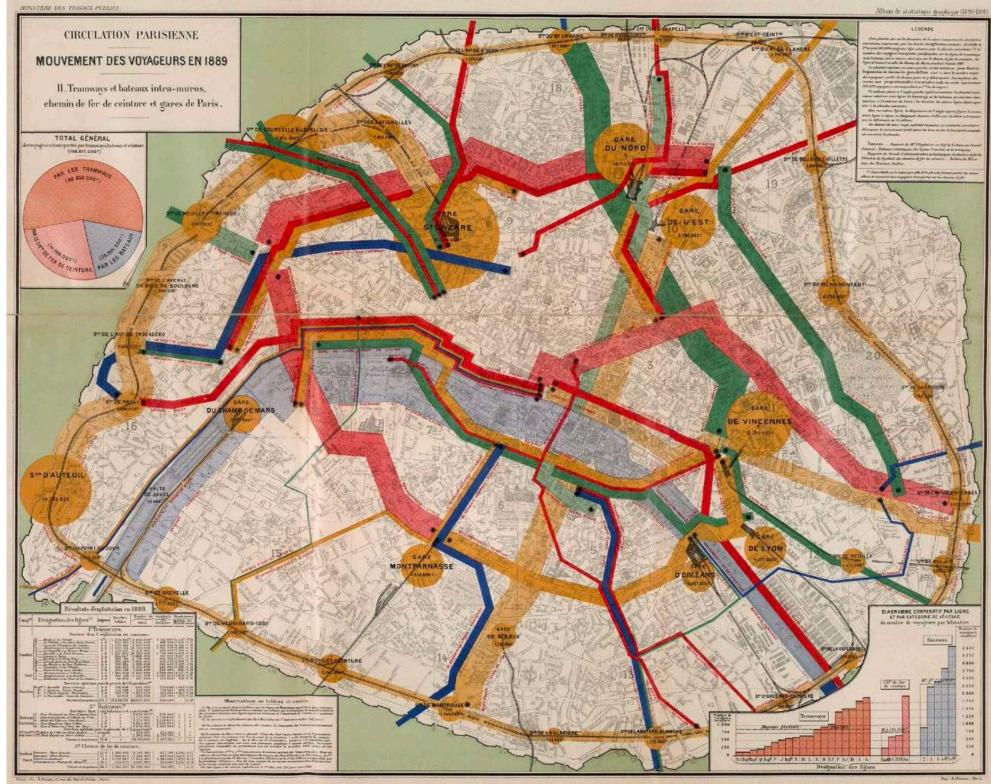


FIGURE 1.1 – Mobilité parisienne d'un point de vue eulérien (1889)

éléments), on a

$$\sum_{i=1}^N \delta_{x_i(t)} = \sum_{i=1}^N \delta_{x_{\varphi(i)}(t)}.$$

Si les points sont distincts deux à deux, la quantité d'information marginale entre les descriptions lagrangienne et eulérienne correspond à la connaissance d'une permutation particulière dans S_N , dont le cardinal est $N!$. On peut évaluer cette quantité d'information (voir section 1.4 pour plus de détails) en termes de bits d'information, i.e. la taille des mots de 0 et de 1 qu'il faudrait utiliser pour désigner³ une permutation particulière parmi les $N!$ possibles, qui vaut $\log_2(N!)$. La formule de Stirling permet d'obtenir un équivalent de ce nombre :

$$N! \sim (N/e)^N \sqrt{2\pi N} \implies \log_2(N!) \sim N \log_2 N.$$

Remarque 1.2. On se trouve souvent dans une situation intermédiaire entre les deux visions, en termes d'information. Par exemple, si l'on considère la densité de population (résidence principale) dans la région parisienne d'un côté, et de l'autre la liste des personnes y habitant, la connaissance de “qui habite où” repose sur une quantité d'information de l'ordre de $N \log_2 N$. Si l'on dispose pour chaque habitant de son niveau de revenu, mettons pour simplifier riche ou non-riche, et que l'on sait que seuls les riches peuvent habiter intra-muros, cela permet de limiter le champ des possibles, et de lever une part de l'indétermination eulérienne, donc de se rapprocher du lagrangien. L'exercice 1.2 ci-dessous précise dans une situation simplifiée le gain apporté par une information partielle.

Exercice 1.2. On considère N points dans un espace, avec la dégénérescence suivante : on a p_1 points confondus en y_1 , p_2 points confondus en y_2 , ..., p_k points confondus en y_k , avec les y_k distincts deux

3. On peut aussi concevoir cette information suivante : si je suis en face d'une personne qui détient l'information (i.e. qui connaît l'indexation des points), combien de questions binaires (réponse oui ou non) dois-je lui poser pour récupérer cette information ?

à deux, et $p_1 + \dots + p_k = N$. Calculer la quantité d'information séparant le lagrangien de l'eulérian. Estimer cette quantité pour une distribution uniforme des regroupements : $p_i = p$ constant, $kp = N$, en supposant $1 \ll p \ll N$.

Si l'on considère maintenant la mesure μ_t associé à une même population à plusieurs instants successifs, on notera qu'il est possible à partir de cette information d'inférer le mouvement lagrangien des entités, sous certaines hypothèses. Si l'on suppose par exemple que les entités évoluent à une vitesse dont le module est d'ordre u , que la distance typique entre entités est d'ordre L , et que la résolution temporelle Δt des différentes "images" est significativement inférieure à L/u , on pourra aisément reconstruire chacune des trajectoires (L) à partir de la collection des photographies aux instants successifs.

De façon général, passer d'un description eulérienne à une description lagrangienne est un problème mal posé, qui a une importance crucial dans certains domaines, comme celui de la mobilité urbaine. On pourra se reporter à l'exercice 5.1 pour une démarche de reconstruction (non unique) de données lagrangiennes (plans origine-destination) à partir d'observations eulériennes (mesure des entrées-sorties dans un bus).

Si l'on se donne les vitesses des particules u_i , exprimer le phénomène de transport (qui s'écrivait simplement (1.1) dans le cadre lagrangien) dans un cadre eulérien, nécessite des outils assez sophistiqués. Nous présentons ci-dessous comment on peut écrire une équation de transport, ou plus généralement une équation de conservation, de façon eulérienne, et dans un cadre monodimensionnel.

1.1.3 Transport & conservation : le point de vue eulérien

On considère des entités (particules, personnes, véhicules, ...) en mouvement, en nombre suffisant pour qu'il soit pertinent d'introduire une fonction qui mesure la densité locale de ces entités. Dans le développement informel qui suit, nous supposons que cette densité est une fonction régulière. On se place dans un cadre monodimensionnel, et l'on note $\rho(x, t)$ la densité au voisinage du point x , au temps t : la quantité

$$\int_x^{x+\delta x} \rho(x, t) dx$$

est la quantité (exprimée en masse, nombre d'entités, en mol si ce nombre est très grand comme pour des molécules) dans l'intervalle⁴ $]x, x + \delta x[$ à l'instant t . On suppose que le champ de vitesse est donné en tout point et en tout temps $u(x, t)$, de telle sorte que le flux d'entité qui traverse x par unité de temps, au temps t , est $\rho(x, t)u(x, t)$. On écrit alors le bilan de masse instantané sur l'intervalle $]x, x + \delta x[$, c'est-à-dire que la dérivée en temps de la quantité sur cet intervalle est égale au flux rentrant moins le flux sortant :

$$\frac{d}{dt} \int_x^{x+\delta x} \rho(x, t) dx = -(\rho(x + \delta x, t)u(x + \delta x, t) - \rho(x, t)u(x, t)).$$

Le premier terme peut s'écrire (quand δx tend vers 0)

$$\int_x^{x+\delta x} \partial_t \rho(x, t) dx \sim \partial_t \rho(x, t) \delta x$$

et on a par ailleurs

$$(\rho(x + \delta x, t)u(x + \delta x, t) - \rho(x, t)u(x, t)) \sim \partial_x (\rho(x, t)u(x, t)) \delta x.$$

Il vient (on omet la dépendance en (x, t) pour alléger l'écriture)

$$\partial_t \rho + \partial_x (\rho u) = 0, \quad (1.2)$$

qui est l'équation de transport conservatif, parfois appelée *équation de continuité*.

4. Nous utilisons conformément à l'usage l'intervalle ouvert, mais cela ne changerait rien si l'on prenait l'intervalle fermé : la densité étant supposée régulière, une quantité non nulle de matière ne peut se concentrer sur un singleton.

Remarque 1.3. On peut appliquer ce formalisme à des densités non régulières, voire à des mesures qui n'admettent pas de densité par rapport à la mesure de Lebesgue, à l'aide de la notion de solution faible (voir définition 4.9, page 88). Cette extension du sens donné à l'équation permet notamment une description eulérienne (a priori peu adaptée) du transport d'une masse ponctuelle dans l'espace.

La même démarche peut être étendue à un cadre plus général. Si l'on considère connu en chaque point le flux $f(x, t)$ au travers de x (compté positivement vers les x croissants), on obtient de la même manière l'équation (on omet les dépendances explicites en (x, t) pour alléger l'écriture)

$$\partial_t \rho + \partial_x f(x) = 0. \quad (1.3)$$

L'équation ci dessus n'est pas un modèle à proprement parler tant que l'on ne s'est pas donné une expression du flux f . Le transport est un cas particulier de cette équation : on écrit comme indiqué précédemment $f = \rho u$, où u est donné, pour obtenir l'équation (1.2).

On se reportera au chapitre 4 pour la construction et l'étude de telles équations en dimension d'espace plus grande que 1.

Modèles eulériens de trafic routier ou piéton

De façon plus générale, on peut supposer que f est une fonction de ρ . Dans le cas du transport routier ou piéton par exemple, il est raisonnable de considérer que ce flux s'écrit ρu comme pour le transport, avec une vitesse u qui dépend de ρ : $u = u(\rho)$. Pour le trafic routier, on considérera par exemple que la vitesse des véhicules à basse densité (les véhicules sont éloignés les uns des autres) prend une certaine valeur $U > 0$ (vitesse maximale autorisée), et décroît avec la densité jusqu'à s'annuler lorsque la densité maximale ρ_{max} est atteinte (pare-choc contre pare-choc). Un choix possible est

$$u(\rho) = U \left(1 - \frac{\rho}{\rho_{max}} \right), \quad \partial_t \rho + \partial_x (\rho u(\rho)) = 0. \quad (1.4)$$

La représentation graphique de la correspondance $\rho \mapsto f = \rho u$ est appelée *diagramme fondamental*. On notera que, dans cette approche eulérienne, la prise en compte d'individus différenciés est délicate, puisqu'elle nécessiterait de reconstruire les trajectoires individuelles à partir du champ de vitesse $u(\rho)$ inconnu a priori. On se reportera à la section 8.4, page 197, pour une étude plus détaillée de ce type de modèle.

Dans le cadre lagrangien, il était immédiat de prendre en compte des différences de comportement entre individus (voir remarque 1.1, page 10). Si l'on souhaite différencier les comportements individuels, il est nécessaire ici d'introduire un champ de "labels" qui permet de suivre les individus dans leurs mouvements. Ce champ est le pendant continu de la suite d'index $1, \dots, N$. Notons $\alpha = \alpha(x, t)$ ce champ de labels, que nous considérons pour fixer les idées à valeurs dans $[0, 1]$. Pour tout (x, t) , $\alpha(x, t)$ est le label de la personne (parmi l'infinité non dénombrable de personnes) située en x au temps t , et si $\alpha(x', t') = \alpha(x, t)$, cela signifie que cette personne se trouve en x' au temps t' . Pour obtenir une équation (eulérienne) sur α , on considère les caractéristiques associées au champ de vitesse $u(x, t)$:

$$\frac{\partial}{\partial t} X_s(x, t) = u(X_s(x, t), t), \quad X_s(x, s) = x.$$

La fonction $t \mapsto X_s(x, t)$, que nous noterons pour simplifier $t \mapsto X(t)$, est la trajectoire d'un individu. On écrit alors simplement que α est constant le long de cette trajectoire :

$$0 = \frac{d}{dt} \alpha(X_s(x, t), t) = \frac{\partial \alpha}{\partial t} \alpha(X_s(x, t), t) + \frac{\partial X_s}{\partial t}(x, t) \frac{\partial \alpha}{\partial x}(X_s(x, t), t) = \left(\frac{\partial \alpha}{\partial t} + u \frac{\partial \alpha}{\partial x} \right) (X_s(x, t), t),$$

d'où l'équation sur α

$$\partial_t \alpha + u \partial_x \alpha = 0. \quad (1.5)$$

Cette équation est aussi appelée équation de transport, ou équation de transport non conservative, pour souligner le fait qu'elle n'exprime pas la conservation d'une quantité. Coupler cette nouvelle

équation (1.5) à l'équation du trafic routier (1.4) permet de différentier le comportement des conducteurs, en considérant par exemple que le modèle de comportement qui exprime la vitesse en fonction de la densité dépend aussi de l'individu

$$u(\rho, \alpha) = U(\alpha)\Psi(\rho, \alpha).$$

Remarque 1.4. La présence du terme $\partial_x \alpha$ dans l'équation (1.5) ci-dessus est assez troublante : α représente une quantité qui n'a aucune raison d'être une fonction lisse de la variable d'espace. Pire encore, même si nous avons supposé qu'elle prenait ses valeurs dans $[0, 1]$, c'est très artificiel puisqu'elle prend ses valeurs a priori dans un ensemble d'"étiquettes", qui n'a aucune raison d'être muni de la moindre structure⁵, en particulier métrique ni même topologique. On peut rester d'ailleurs sur l'intervalle $[0, 1]$, mais en considérant ses points comme des noms écrits comme suites infinies de lettres (les chiffres 0, 1, ..., 9), mais sans que la notion d'addition par exemple n'ait plus de sens que d'ajouter Jean-Pierre à Marguerite. Il s'agit d'une variable *essentiellement intensive*, i.e. une variable intensive qui n'est pas construite comme densité d'une mesure par rapport à une autre (voir section 1.2 ci-après). La nature intrinsèquement lagrangienne de cette quantité rend périlleuse la modélisation de son évolution sous un angle eulérien. De fait, l'équation de transport sera en général construite par une méthode dite *des caractéristiques*, qui revient au caractère lagrangien du phénomène. Dans cette optique, on aura tendance à écrire l'équation de transport (1.4) sous la forme

$$\frac{D\alpha}{Dt} = 0,$$

qui fait intervenir la dérivée *particulaire* (appelée aussi dérivée *totale*, ou *lagrangienne*), et qui permet de s'affranchir sur terme $\partial_x \alpha$, purement eulérien, qui est d'une certaine manière dénué de sens dans ce contexte.

Remarque 1.5. (Lien micro-macro)

Les ingrédients principaux des modèles eulérien (macro) et lagrangien (micro) sont les mêmes, il s'agit de la donnée d'une vitesse fonction de la densité locale. Le choix $u = U(1 - \rho/\rho_{max})$ en eulérien correspond par exemple La densité étant l'inverse de la distance entre véhicules, on a

$$\varphi(w) = u(1/w) = U(1 - w_{min}/w).$$

Ce cadre conduit également à l'*équation de la chaleur*. Si l'on suppose que le flux est donné par la loi de Fick $f = -D\partial_x \rho$, où D est appelé *coefficient de diffusion*, on obtient

$$\partial_t \rho - D\partial_{xx} f = 0.$$

Cette équation décrit notamment l'évolution de la densité de particules indiscernables qui suivent des mouvements browniens indépendants. Même dans le cadre mono-dimensionnel que nous avons choisi, elle représente une réalité bi- ou tri-dimensionnelle (avec invariance dans une ou deux directions), de telle sorte que l'*ordre* de positionnement des particules sur l'axe des réel, contrairement au cas du trafic routier, n'est absolument pas préservé. On se reportera au chapitre 4 pour une présentation détaillée de cette équation, ainsi qu'à la section 8.4, page 197, pour une introduction succincte à l'étude de ce type d'équations aux dérivées partielles.

1.1.4 Au delà de l'espace physique

Bien qu'il ne s'agisse pas d'une pratique courante, cette distinction lagrangien-eulérien nous semble pouvoir s'étendre à d'autres situations que celles considérées précédemment, où l'on s'intéressait au suivi de positions (L) ou d'occupation de l'espace (E) au cours du temps. Considérons par exemple la collection des N assujettis sociaux français. La donnée des âges afférents aux différents individus (qui est d'ailleurs contenue dans la liste des labels si on identifie chaque personne par son numéro de sécurité sociale) est de nature lagrangienne, puisque la variable âge est donnée pour chaque individu.

5. A part éventuellement une *mesure* définie sur un tribu de l'ensemble des valeurs, qui permettrait de "compter" les gens appartenant à un certain ensemble.

Comme pour les particules transportées du paragraphe précédent, chaque individu est soumis à une équation de vieillissement, que l'on peut voir comme une équation de transport dans l'espace des temps, à la vitesse de 1 seconde par seconde : l'âge $a_i(t)$ de la personne i vérifie l'équation $da_i/dt = 1$, qui s'intègre simplement entre t_i (date de naissance de i) et t , pour donner

$$a_i(t) = a_i(t_i) + \int_{t_i}^t 1 ds = t - t_i.$$

On peut à partir de cette donnée à un instant donné établir un histogramme, que l'on peut encoder par un tableau de nombres n_0, n_1, \dots, n_{122} , où n_k est le nombre de personnes dont l'âge est compris entre k et $k + 1$. Cet histogramme s'appelle une *pyramide des âges*. On peut considérer cette pyramide des âges comme étant de nature eulérienne, du fait que les identités des personnes ont disparu. La notion de zone géographique en laquelle on observe une population d'individus indifférenciés est remplacée ici par un intervalle de temps, entre deux âges successifs.

Cette vision eulérienne conduit à écrire une équation en temps continu sur la densité $\rho(a, t)$ du nombre de personnes d'âge a au temps t . Plus précisément, $\int_a^{a+\delta a} \rho(a', t) da'$ est le nombre de personnes, au temps t , dont l'âge est compris entre a et $a + \delta a$. On note $\mu = \mu(a)$ le taux de mortalité, tel que $\mu(a)\rho(a, t)\delta t$ est le nombre de gens d'âge a au temps t qui décèdent pendant $[t, t + \delta t]$. On note $\beta = \beta(a)$ le taux de fécondité, tel que $\beta(a)\rho(a, t)\delta a$ est le nombre d'enfants "produits" par les personnes d'âge entre a et $a + \delta a$. On obtient une équation de transport non homogène (pris en compte de la mortalité), où la variable d'âge a joue le rôle de variable d'espace,

$$\begin{cases} \partial_t \rho + \partial_a \rho = -\mu \rho, \\ \rho(0, t) = \int_0^{+\infty} \beta(a) \rho(a, t). \end{cases} \quad (1.6)$$

Noter que la condition en $a = 0$ (deuxième ligne du système), bien qu'elle semble consister en la prescription de la valeur de ρ en 0, est en fait une condition de flux : $\rho(0, t)$ doit être lu $1 \times \rho(0, t)$, où 1 est la vitesse (de 1 année par année) à laquelle les nouveaux arrivants apparaissent dans ce modèle structuré en âge.

Exercice 1.3. Écrire un modèle d'évolution de population structuré en âge à deux populations (hommes et femmes).

Exercice 1.4. On considère le système (1.6), et l'on suppose que $\rho(\cdot, 0)$ est supportée par un intervalle $[0, A]$. On se donne un âge initial $a_0 \in [0, A[$, et l'on introduit la fonction $\varphi(t) = \rho(a + t, t)$.

a) Écrire l'équation différentielle vérifiée par φ , et donner une condition suffisante sur la mortalité μ pour que l'âge d'un individu ne puisse jamais dépasser la valeur A .

b) Donner l'expression de l'espérance de vie des personnes d'âge a_0 à l'instant initial.

Exercice 1.5. a) On considère le modèle (1.6) en supposant que μ et β ne dépendent pas de a . En supposant que ρ a toujours un support borné (inclus dans $[0, A[$), établir une équation différentielle vérifiée par la population totale $N(t)$ (intégrale de $\rho(a, t)$ sur $[0, A]$), et préciser le comportement en temps long de cette population en fonction des paramètres.

b) On suppose toujours le taux de mortalité μ constant, mais β est supposé dépendre de l'effectif global N , selon l'expression

$$\beta = \beta(N) = \beta_0 \left(1 - \frac{N}{N_{max}}\right).$$

écrire la nouvelle équation sur N , et décrire le comportement des solutions lorsque $\beta_0 > \mu$.

Exercice 1.6. (Modélisation, définition et estimation de l'espérance de vie)

On considère une population fermée (les changements d'effectifs ne sont dus qu'aux décès), dont la pyramide des âges est représentée par $\rho(a, t)$, supposé connue *dans le passé*, relativement à un instant τ donné.

a) Dans le cas d'une distribution stationnaire, proposer une expression pour ce que l'on appelle l'espérance de vie à l'âge A , notée $E(A)$. Calculer la dérivée de S par rapport à A , et proposer une interprétation graphique de cette quantité. Cette dérivée est elle toujours négative ?

b) Proposer une manière d'estimer cette espérance de vie à l'âge A au temps τ dans le cas général (distribution non stationnaire), définie selon le principe suivant : on estime en fonction des données connues le taux de mortalité pour tous les âges supérieurs à A , et l'on prédit le futur de la génération A en supposant que ces taux de mortalité sont figés en temps.

Discuter des limites de cette définition.

Exercice 1.7. (Modélisation, extensions du modèle d'évolution d'une pyramide des âges)

Compléter ou modifier le modèle (1.6) de façon à prendre en compte les phénomènes suivants :

- (i) Émigration ou immigration, guerre, épidémie, ...
- (ii) Dépendance des taux de mortalité et de fécondité vis à vis de la population globale (dans un contexte de ressources limitées par exemple).
- (iii) Prise en compte différenciée des populations de femmes et d'hommes.

Exercice 1.8. (Modélisation, taux de fécondité apparent)

Explorer comment un modèle de type (1.6) pourrait expliquer pourquoi le fait que les individus font des enfants de plus en plus tard peut donner l'impression que le taux de fécondité est en baisse.

Remarque 1.6. La connaissance de la pyramide des âges à des instants successifs, permet de remonter à des informations (partiellement) lagrangiennes. Du fait que, contrairement au cas de particules qui auraient un mouvement erratique dans l'espace, on connaît l'équation d'évolution sous-jacente avec une parfaite précision (équation de vieillissement, i.e. transport sur la ligne du temps à vitesse 1), si l'on considère la pyramide des âges associée à une même population (sans apport extérieur) à deux instants (années) successifs, T et $T + 1$, on sait que les personnes emplissant la case $[k, k + 1[$ au temps $T + 1$ étaient dans la case $[k - 1, k[$ précédemment. Si $n_k < n_{k-1}$, cela signifie qu'une partie des personnes considérées est sortie de la population (décès ou émigration). Il s'agit d'une information de nature lagrangienne, puisqu'on suit le "mouvement" (sur la ligne du temps) d'un ensemble de personnes, information partiellement dégradée puisqu'on ne sait pas quelles personnes sont sorties du circuit.

Remarque 1.7. Au delà de l'espace et du temps, cette dichotomie eulérien-lagrangien est pertinente dans tout contexte où la notion d'individu ou *entité* est définie. L'approche lagrangienne consiste à privilégier le point de vue des entités, l'approche eulérienne correspondant à la vision d'un observateur extérieur qui "fixe" une zone (de l'espace ou du temps) donnée. Par exemple dans le domaine des idées, la biographie d'un intellectuel (l'*entité* en question), qui retrace ce qu'il a pu écrire sur différentes thèmes ou événements, pourrait être qualifiée de lagrangienne. Un ouvrage de type *monographie*, portant sur un thème donné et confrontant les points de vue de différents auteurs, est par opposition eulérienne. En acceptant une définition plus abstraite de la notion d'*entité*, on peut également étendre ces termes à la description de la musique. Si l'on considère qu'une mélodie ou un thème, même au cœur d'une pièce instrumentale, correspond à ce qui pourrait être chanté par une personne, l'approche centrée sur l'étude des différentes voix (d'une fugue de Bach par exemple), appelée *contrepoint*, est lagrangienne. Si l'on se place à un instant donné, les notes forment un accord, l'objet de base de l'*harmonie*, qui est de nature eulérienne. Noter que, se basant sur la vision d'un morceau de musique comme un multigraphie dont l'abscisse est le temps, et l'ordonnée la hauteur des notes (conformément à l'écriture des *partitions* dans la culture occidentale), on parle d'approche horizontale pour la première, et verticale pour la seconde.

Remarque 1.8. Pour revenir à des notions plus proche des mathématiques, l'Analyse Fonctionnelle, basée sur la définition de distances (issues de normes) entre fonctions d'un même espace, qui permet de donner un cadre aux équations aux Dérivées Partielles, correspond à une vision essentiellement eulérienne. Ainsi les distances usuelles définies, par exemple, pour les fonctions à valeurs réelles sur un intervalle $I \subset \mathbb{R}$, du type

$$\sup_I |f(x) - g(x)|, \int |f(x) - g(x)|, \left(\int |f(x) - g(x)|^p \right)^{1/p}, \left(\int |f'(x) - g'(x)|^2 \right)^{1/2}, \dots$$

sont de nature eulériennes, puisqu'elles sont entièrement basées sur des différences de valeurs $f(x) - g(x)$ en un même point x de l'espace. La distance de Wasserstein (voir chapitre 14) définie pour des couples de mesures de probabilité (ou mesures de même masse), est en revanche d'inspiration lagrangienne, ou *horizontale*.

1.1.5 Dérivées lagrangienne et eulérienne

On s'intéresse ici au mouvement d'entités dans l'espace physique, en nombre suffisamment grand⁶ pour qu'il soit pertinent de définir une vitesse locale $u(x, t)$ en chaque point x de l'espace, à chaque instant t . On note φ une quantité afférente à ces entités transportées, par exemple leur température, la masse volumique locale, ou la vitesse elle-même. L'écriture $\varphi = \varphi(x, t)$ est eulérienne, du fait que x représente une position fixe, en laquelle on se place en quelque sorte pour observer l'évolution au cours du temps $t \mapsto \varphi$ de la quantité φ , qui est afférente à des entités qui changent au cours du temps. La dérivée partielle par rapport au temps

$$\frac{\partial \varphi}{\partial t}(x, t),$$

dérivée *verticale* dans l'espace-temps $\mathbb{R}^d \times \mathbb{R}$ est elle-même eulérienne.

Considérons maintenant les *trajectoires* des entités, qu'on appelle aussi le *flot* associé au champ $u(x, t)$. Nous allons conformément à l'usage identifier les particules par leur position dans l'espace physique à un instant donné. Dans cet esprit, on note $s \mapsto X_s(x, t)$ la trajectoire de l'entité située en x au temps t :

$$\begin{cases} \frac{\partial X_s}{\partial s}(x, t) &= u(X_s(x, t), s) \\ X_t(x, t) &= x. \end{cases} \quad (1.7)$$

On notera que le ∂t ci-dessus correspond maintenant à une dérivée lagrangienne, puisque l'on suit précisément la particule dans son mouvement. Ces trajectoires nous permettent de définir ce que l'on appelle la dérivée lagrangienne (on parle aussi parfois de dérivée *particulaire*, ou dérivée *totale*). Si l'on se place en un point x , au temps t , la dérivée particulière en ce point est la dérivée en temps de la quantité φ *telle que vécue* par l'entité qui est en x au temps t , définie plus précisément par

$$\frac{D\varphi}{Dt}(x, t) = \frac{\partial}{\partial s} (\varphi(X_s(x, t), s))_{s=t}.$$

Elle s'exprime

$$\frac{D\varphi}{Dt}(x, t) = \frac{\partial \varphi}{\partial t}(x, t) + \underbrace{(\partial_s X_s(x, t))_{s=t}}_{u(x, t)} \cdot \nabla \varphi(x, t) = \frac{\partial \varphi}{\partial t}(x, t) + u(x, t) \cdot \nabla \varphi(x, t). \quad (1.8)$$

Exercice 1.9. On considère une tronçon de route en aval d'un péage. On suppose que tous les véhicules accélèrent de la même manière à partir de la sortie du péage, en adoptant une vitesse qui dépend de la distance x à cette sortie, selon l'expression

$$u(x) = U \left(1 - e^{-x/\delta} \right).$$

Exprimez les dérivées en temps lagrangienne et eulérienne de la vitesse, en un point en aval du péage représenté par sa position $x > 0$.

1.2 Grandeur intensives et extensives

Definition 1.9. (Variables intensives et quantités extensives)

On appelle *extensive* (E) une grandeur afférente à un système matériel qui vérifie la propriété de

6. On définit implicitement champ de vitesse comme un continuum, en gardant à l'esprit qu'à une échelle qui serait inférieure à la distance typique entre deux des entités considérées, cette notion est une vue de l'esprit, détachée de la réalité.

sommation suivante : si l'on “double” le système, en considérant qu'on lui ajoute une copie de lui-même, la quantité est elle-même multipliée par deux. Une telle grandeur est associée à la notion de *quantité*, comme la masse ou le volume, ou l'énergie interne (voir section 1.6). Une grandeur est dite *intensive* (I) si elle n'est pas modifiée par l'expérience virtuelle de doublement décrite ci-dessus.

Une variable intensive peut être définie ponctuellement (on peut se ramener à des systèmes de plus en plus petits en divisant au lieu de doubler), ce qui conduit à la notion mathématique de *fonction*, ou *champ*, qui, à un point de l'espace physique associe un nombre réel ou un vecteur. Elle *qualifie* le voisinage du point considéré.

Une grandeur extensive est en revanche afférente à un objet ou à une zone déterminée de l'espace, elle en *quantifie*, au sens le plus large, la taille (il peut s'agir de volume, de masse, de nombre d'habitants sur une zone géographique, d'énergie interne, etc.). La notion mathématique sous-jacente est celle de *mesure* (voir section 1.3 ci-après), définie précisément comme une application qui à une partie associe un nombre positif ou nul, en vérifiant une règle de sommation qui en garantit la nature extensive. Cette règle de sommation stipule que la mesure d'une union de parties disjointes est la somme des mesures des parties⁷.

Le produit d'une grandeur intensive par une variable extensive donne une grandeur extensive. Ainsi, si l'on considère un objet homogène de densité ρ (variable I), de volume V (grandeur E), le produit des deux est la masse m , qui est de nature extensive. Si la variable intensive n'est pas uniforme sur l'objet, le produit doit être remplacé par la notion d'intégrale : on intègre la variable $\rho = \rho(x)$ contre la mesure volume, ce qui s'écrit

$$m(A) = \int_A \rho(x) dx,$$

l'intégrale pouvant être interprétée comme une somme infinie de contribution infinitésimales, dx désignant un petit volume au voisinage d'un point x , et $\rho(x)$ la densité en ce point. Comme détaillé plus loin, la masse peut elle-même être considérée comme une mesure, contre laquelle on peut intégrer une variable intensive. Par exemple si l'ensemble A est rempli d'un fluide en mouvement, décrit part un champ de vitesse $u = u(x)$ (variable I), la quantité de mouvement du fluide dans A est l'intégrale de u contre la mesure de masse, ce que l'on écrira

$$p(A) = \int_A u(x) dm.$$

Le produit de variables intensives est une nouvelle variable intensive.

Le produit de deux quantités extensives n'a pas de sens a priori, en tout cas il ne rentre pas dans la classification proposée.

Exercice 1.10. Interpréter en termes de quantités extensives et intensives cette maxime populaire :

Plus y'a de gruyère, plus y'a de trous, plus y'a de trous, moins y'a de gruyère.

Parmi les variables intensives les plus couramment utilisées, un grand nombre correspond à une *densité*, au sens le plus général du terme, qui exprime un lien entre deux quantités extensives. La démarche de construction de ces densités est la suivante : on considère une mesure de références, en général la mesure de volume λ (mesure de Lebesgue sur l'espace physique), qui à un ensemble A associe son volume $\lambda(A)$. On considère maintenant une autre mesure, par exemple la mesure de la masse $m(A)$ contenue à un instant donné dans une zone A de l'espace. Si A est rempli de façon homogène, on estime la densité par $m(A)/\lambda(A)$, en kg m^{-3} . De façon plus générale (lorsque la masse n'est pas répartie uniformément), un grand théorème de théorie de la mesure (dit de Radon–Nikodyn, qui dépasse le cadre de ce cours) assure que, si la mesure de masse est absolument continue par rapport à la mesure de volume⁸ alors $m(\cdot)$ admet une densité par rapport à λ , c'est-à-dire qu'il existe une

7. Comme indiqué plus loin, dans la définition mathématique, on demande que cette règle de sommation s'applique à des unions potentiellement infinies de partie, sous réserve que cet infini soit dénombrable.

8. C'est-à-dire que $\lambda(A) = 0$ implique $m(A) = 0$: ce qui n'occupe pas de place ne pèse rien, ce qui exclut les concentrations, en particulier l'existence de masse ponctuelles.

fonction ρ de l'espace courant dans \mathbb{R}_+ telle que $dm = \rho d\lambda$. Pour tout ensemble A mesurable, on aura alors

$$m(A) = \int_A \rho d\lambda.$$

Une densité de population est obtenue ainsi à partir de la mesure 'nombre de gens dans la zone' et de la mesure de superficie. On peut aussi interpréter la variable intensive par excellence, la température, comme une densité d'agitation des molécules par unité de volume. On peut pousser la démarche jusqu'à considérer un champ de vitesse associé à un continuum de matière en mouvement comme une densité de quantité de mouvement (variable extensive vectorielle, i.e. mesure vectorielle) relativement à la mesure de masse.

Au delà de ces variables extensives rendues intensives par cette approche de type Radon Nikodym (construction d'une densité relativement à une autre mesure), on peut concevoir des variables *essentiellement* intensives, qui seraient qualifiées de *qualitatives* dans le langage commun, dont l'intégration contre une mesure n'a aucun sens, au moins en tant que quantité, comme par exemple une couleur de cheveux, une propension chez un individu à voter pour tel ou tel candidat à une élection, ou un goût plus ou moins prononcé pour les plats épices. De telles variables peuvent être discrètes : comme le nom et le genre des personnes dans une population, une équipe de football supportée. Elles vivent dans des ensembles sans structure mathématique clairement identifiée, en particulier la notion d'interpolée entre deux valeurs n'a pas forcément de sens clair, ce qui rend difficile l'élaboration de modèles d'évolution pour ces variables.

Variables extensives / extensives et sommation

Dans l'esprit des considérations précédentes, on peut dire que les variables extensives sont les variables que l'on peut sommer lorsqu'elles sont afférentes à deux objets dont on considère la réunion, et les variables intensives celles que l'on ne peut pas sommer de cette manière. Précisons en premier lieu que cette sommation entre réels mérite d'être distinguée de la loi de composition interne qui fait de \mathbb{R} un groupe additif. Il s'agit là de sommer des quantités positives, éventuellement de les soustraire (plus précisément de soustraire une quantité à une quantité plus grande). D'un certain point de vue, l'espace \mathbb{R}_+ des quantités doit être vu dans ce cadre comme un *monoïde*⁹, comme l'espace des résultats d'une *mesure* (voir section suivante). Concernant la non-sommabilité des variables intensives, précisons que ces variables ne peuvent être en effet sommées comme des quantités absolues, mais qu'on peut être amenés à effectuer la somme de telles quantités dans certains contextes, par exemple :

- une personne marche à la vitesse v vers l'avant d'un train qui se déplace à la vitesse V . La vitesse du piéton un observateur extérieur au train est $v + V$, qui est bien la somme des deux vitesses, malgré le caractère intensif de la vitesse, mais il ne s'agit pas d'une somme conséquente à la réunion de deux systèmes, il s'agit d'une formule de cinématique détachée des notions de quantité.

- dans un contexte thermique, si l'on dit que la température du jour est 2 degrés au dessus des normales saisonnières, on considère que cette température est du type $T + 2^\circ \text{C}$, qui est bien a priori la somme de deux températures. mais ces températures ne jouent pas un rôle symétrique, la première est absolue, la seconde est une *variation de température*, qui pourrait d'ailleurs être inférieure au zéro absolu.

Dans les deux cas ci-dessus, les variables en question peuvent être vues comme résultant du quotient de deux variables extensives, elle en sont pas déconnectées de la notion de quantité : la vitesse est une quantité de mouvement par unité de masse, et la température d'un objet, à une constante multiplicative près (capacité thermique), peut être vue comme une quantité d'énergie par unité de volume). Pour les variables *essentiellement intensives* évoquée plus haut, la sommation n'a aucun sens. Remarquons d'ailleurs que l'on peut être amené à considérer que, par exemple, $1 + 3$ n'a aucun sens. Si 1 et 3 correspondent à des labels pour indiquer le type d'une entité (par exemple 1 = salarié, 2 = chômeur, 3 = retraité), la somme de ces deux types n'a absolument aucun sens.

9. Ensembe muni d'une loi de composition interne associative, et admettant un élément neutre.

1.3 La notion de mesure, point de vue de la modélisation

Nous revenons ici sur la notion de *mesure*, objet mathématique associé aux variables extensives telles que décrites dans la section précédente, du point de vue de la modélisation. Comme cadre conceptuel d'appréhension du réel, cette notion unique de mesure répond à deux enjeux, qu'il nous paraît important de distinguer malgré le fait qu'ils correspondent à la même notion mathématique.

En premier lieu, une mesure permet de structurer le fond d'un espace destiné à accueillir de la matière. Par espace nous entendons par exemple l'espace euclidien usuel, qui est en dimension 3 un modèle de l'espace physique dans lequel nous vivons, sur lequel il peut être pertinent de définir des *champs* (champ de densité, de concentration d'un polluant, de température, de densité de population, ...). Considérons par exemple un milieu occupant une certaine zone de l'espace euclidien, milieu dont on connaît la densité. Si l'on suppose la densité constante sur une zone A , la masse portée par A est le produit entre cette valeur de densité et le volume de la zone. Il est donc essentiel de savoir estimer le volume des zones susceptibles d'accueillir de la matière, pour pouvoir estimer la masse correspondante. Définir une mesure consiste précisément à concevoir une procédure pour associer à une zone son volume. Même s'il n'est pas dans les usages d'affecter une unité physique aux grandeurs mathématiques, on pourra concevoir cette mesure comme s'exprimant en unité de volume (ou d'aire s'il s'agit de l'espace bi-dimensionnel, ou de longueur s'il s'agit d'un espace à une dimension). Il s'agit d'une donnée *statique* associée à l'espace considéré. Dans le cas de l'espace euclidien, ce volume est canoniquement défini dans le cas de formes simples : longueur d'un intervalle, aire d'un rectangle, volume d'un parallélogramme. La notion d'intégration d'une fonction constante sur de tels ensembles est basée sur le simple produit de la valeur à intégrer par le volume. Si, suivant l'intuition associée à la notion de volume, on décrète que le volume de la réunion de deux zones disjointes est la somme des volumes des zones élémentaires, on peut estimer le volume de toutes les zones qui peuvent se construire comme réunion de ces formes simples. Définir le volume de n'importe quel ensemble est plus délicat et même, d'une certaine manière, impossible, comme nous le verrons. La construction de la mesure de Lebesgue permet de définir un tel volume pour une classe très générale de zones de l'espace euclidien, et permettra de construire un cadre définissant la notion d'intégrales pour des fonctions très générales.

Les mesures ont également vocation à représenter des quantités absolues de matière (fluides, matériau solide, cellules, individus, ...), distributions d'une certaine substance susceptible d'évoluer en temps, d'être transportée, supprimée, développée. L'objet mathématique associé est le même, mais la nature de la réalité qu'il a vocation à représenter est différente. Il sera ici naturel de penser la mesure associée comme exprimée en kg, en moles, qui mesurent des quantités de matières associées à des principes de conservation.

Nous proposons dans les paragraphes qui suivent quelques exemples de situations réelles qui illustrent les deux types de mesures évoqués ci-dessus et les liens qu'elles entretiennent : mesures de type *volume*, qui formalisent la capacité de parties de l'espace sous-jacent à accueillir de la matière, et mesures de type *masse*, qui représentent des quantités de choses réelles. Nous nous restreignons dans ces exemples à des ensembles finis, de telle sorte que les objets mathématiques sont très simples à définir.

Superficies, densités, et nombre d'habitants.

Considérons l'ensemble $X = \llbracket 1, N \rrbracket$ des grandes villes françaises, numérotées de 1 à N . On note $\mu_i > 0$ la superficie de la ville i . À la collection des μ_i est naturellement associée une application μ de l'ensemble $\mathcal{P}(X)$ des parties de X dans \mathbb{R}_+ ;

$$\mu : A \in \mathcal{P}(X) \longmapsto \mu(A) = \sum_{i \in A} \mu_i \in \mathbb{R}_+. \quad (1.9)$$

Cette application est additive, au sens où si A et B sont disjoints, $\mu(A \cup B) = \mu(A) + \mu(B)$. Cette application définit donc une variable *extensive*.

Il s'agit d'une mesure au sens volumique évoqué ci-dessus, qui structure l'ensemble des villes en termes de capacité d'accueil. Notons maintenant ρ_i la densité d'habitants dans la ville i . Il s'agit là

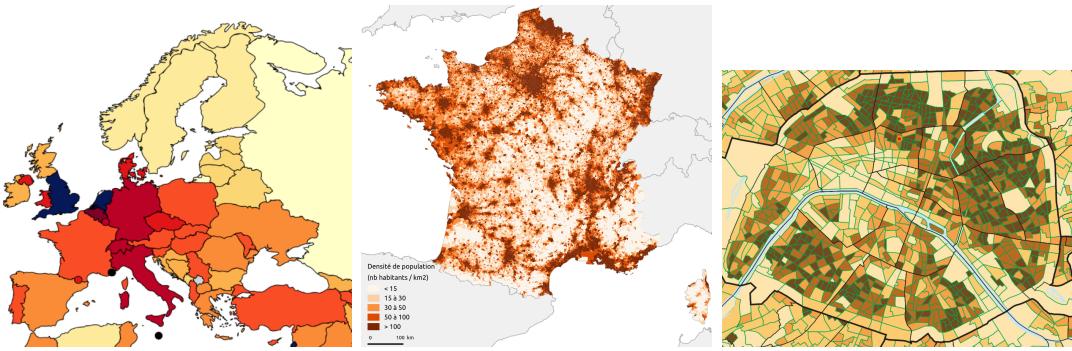


FIGURE 1.2 – Densités de population à différentes échelles

d'une variable *intensive*¹⁰. Le produit $m_i = \rho_i \mu_i$ est le nombre d'habitants dans cette ville i . On peut, comme précédemment pour les μ_i , associer à la collection des villes une application m de $\mathcal{P}(X)$ dans \mathbb{R} , additive par construction. Cette application est une nouvelle mesure sur X , de type “masse”. Le nombre total d'habitants dans le sous-ensemble $A \subset X$ de villes peut s'écrire comme un produit de dualité¹¹ noté $\langle \rho, \mu \rangle_A$ entre les collections de superficies et de densités

$$m_A = \langle \rho, \mu \rangle_A = \sum_{i \in A} \rho_i \mu_i.$$

Il s'agit là de la version discrète d'une intégrale, construite par mise en dualité d'une mesure volume (μ , version discrète de la mesure de Lebesgue construite plus loin) et d'une variable intensive (densité ρ , qui joue le rôle d'une fonction à intégrer sur un domaine). La mesure masse m est la variable sommable, produit de la variable extensive μ et la variable intensive ρ .

On peut aussi définir, dans le cas présent d'une collection finie de villes, des mesures qui correspondent à des probabilités. Prenons l'exemple d'un crime commis à Paris à l'heure H d'un jour J. Vingt-quatre heures après, l'assassin court toujours, et les enquêteurs cherchent à estimer dans quelle ville il pourrait être. L'état de leur opinion concernant la position du fugitif peut être encodé par une mesure $m = (m_i)$. Si l'on sait qu'il ne dispose pas de véhicule et que l'on considère que prendre le train était risqué pour lui, on considérera que la probabilité associée à Paris est de 0.75. Si l'on sait qu'il a des contacts à Lyon, on évaluera à 0.15 la probabilité qu'il y soit, le complément étant distribué sur le reste du pays en fonction des informations que l'on peut avoir. On a ici l'exemple typique d'une mesure (ici de probabilité, c'est à dire normalisée à 1) qui évolue au cours du temps, en fonction des informations reçues.

L'intérêt d'introduire la notion de *mesure* pour les exemples ci-dessus, alors que les objets manipulés se ramènent à des tableaux de nombres réels, n'est pas immédiat. Nous verrons qu'il est néanmoins fécond de considérer par exemple la collection $\mu = (\mu_i)$ des superficies comme une application qui, à un ensemble de villes $I \subset \llbracket 1, N \rrbracket$, associe la population totale des villes concernées, selon l'expression (1.9). Cette application attribue 0 à l'ensemble vide, et vérifie par construction la règle de sommation suivante : l'image de la réunion de deux ensembles disjoints est la somme des images (on dira que l'application est *additive*), ce qui peut s'écrire

$$A \cap B = \emptyset \implies \mu(A \cup B) = \mu(A) + \mu(B).$$

Nous définirons une mesure comme une application qui à une partie associe un réel positif, et qui vérifie des conditions du type de celles qui précédent.

10. La densité associée à la réunion de deux villes de même densité ρ est ρ , et pas 2ρ .

11. Un produit de dualité entre deux espaces vectoriels E et F est simplement une application bilinéaire de $E \times F$ dans \mathbb{R} . On dit que cette application met les espaces en dualité. Un espace euclidien est ainsi en dualité avec lui-même par le biais de son produit scalaire.

Densité de population

Les développements précédents sont basés sur une vision discrète, considérant les zones d'habitation comme des points. Nous décrivons ici la démarche de construction d'une densité "continue" (en fait, elle n'a pas de raison d'être continue au sens strict, on parlera plutôt de densité *diffuse* dans ce contexte). La construction repose sur deux mesures, une de type volume et l'autre de type masse (conformément à la distinction faite en début de section). La première est la mesure d'aire, variable extensive (E) donc, qui est la mesure de Lebesgue λ définie à la surface de la terre (que l'on considérera plane ici), "vue du dessus" au sens où l'élément de surface correspond à une sphère virtuelle qui approcherait au mieux la surface du globe¹². L'aire d'une zone est donc en fait un angle solide (relativement au centre de la terre) qui découpe cette zone, multiplié par le rayon moyen de la terre au carré. On a par ailleurs une mesure μ de type masse, qui est la mesure de population : si A est une zone géographique (assimilée à l'angle solide qui la découpe), $\mu(A)$ est le nombre de personnes qui résident¹³ dans cette zone, en supposant que chaque personne a un point d'ancrage bien défini. La densité de population (variable de type I) est alors construire comme la densité (au sens mathématiques) de la mesure μ relativement à la mesure λ . Bien entendu, cette construction ne peut se faire rigoureusement au sens mathématique, si l'on considère la mesure population à l'échelle la plus fine comme une mesure atomique, qui n'est pas absolument continue¹⁴ par rapport à la mesure de Lebesgue.

La mesure de population est au départ une mesure de comptage, que l'on peut assimiler à une somme de masses ponctuelles aux points de résidence (assimilable aux coordonnées GPS du point de résidence) ; cette mesure atomique n'est pas, au sens strict, absolument continue par rapport à la mesure de Lebesgue. On peut néanmoins définir cette densité à une certain échelle, en décomposant la zone géographie en cellules suffisamment grandes pour contenir un nombre significatif de personnes. On définit alors une densité constante sur chaque cellule C , égale à $\mu(C)/\lambda(C)$. La figure 1.3 présente le résultat d'une telle approche pour différentes granularités.

Il peut être pertinent d'affiner cette approche en évaluant la densité relativement à la surface habitable. On peut ainsi considérer la mesure ν qui à une zone A affecte la totalité de la surface de plancher qu'elle contient. Il s'agit d'une mesure "statique", ou au moins quasi-statique, qui évolue sur de grandes échelles de temps. La densité de cette mesure relativement à la densité de terrain telle que définie précédemment s'appelle *densité du bâti*, ce qui incite à nommer ν *mesure du bâti*. La densité de la mesure population relativement à la mesure du bâti est une densité de population en un sens différent de la précédente, plus représentative de l'*entassement* des personnes. Dans les zones où elle est non nulle, l'inverse de cette densité est le nombre de mètres carrés de plancher dont chaque personne dispose en moyenne¹⁵. La figure 1.3 représente cette densité à Paris en 1999. Il s'agit d'une quantité sans dimension, de type comptage, qui correspond pour un bâtiment au nombre d'étages. On prendra garde en revanche que, définie sur une zone comme un quartier ou un arrondissement, cette densité prend en compte les zones non habitées entre les bâtiments (voies, trottoirs, parcs, ...), qui affichent une surface habitable nulle, elle est donc inférieure en général au nombre moyen d'étages des bâtiments de la zone considérée.

Aérosols.

On considère maintenant une collection de N micro-gouttelettes sphériques flottant dans l'air. Si l'on note μ_i le volume de la gouttelette i , on peut définir une application de l'ensemble des parties de $X = \llbracket 1, N \rrbracket$ dans \mathbb{R}^+ associant à une sous-collection de gouttelettes son volume total. Si l'on note ρ

12. On pourrait, et cela est pertinent dans certains contextes, définir la mesure de surface au sol prenant en compte les aspérités du terrain, il s'agit alors de définir une mesure d'aire sur une variété, potentiellement non régulière, ce n'est pas ce qui est fait ici.

13. Nous admettrons que la notion de résident est bien définie, c'est à dire qu'il existe, pour la population considérée, une application de l'ensemble des personnes vers la zone géographique considérée. Même s'il peut être parfois délicat d'affecter un lieu unique à chaque personne (étudiants, militaires, retraités aisés alternant entre résidence principale et résidence secondaire, ...), il est important dans ce contexte de ne pas compter les gens deux fois, de façon à ce que, si l'on intègre la densité ainsi définie contre la mesure d'aire, on retrouve bien la population totale.

14. En effet, le point correspondant à la position d'une personne a une mesure de Lebesgue nulle.

15. On prendra néanmoins garde au fait que la densité du bâti ne distingue pas les surfaces effectivement habitées des surfaces inoccupées, il s'agit plus d'une potentialité de logement que d'une habitation effective. Si une grande partie de la surface de plancher est de fait inoccupée, la densité de population au mètre carré habitable peut être très en dessous de ce qu'elle est effectivement.

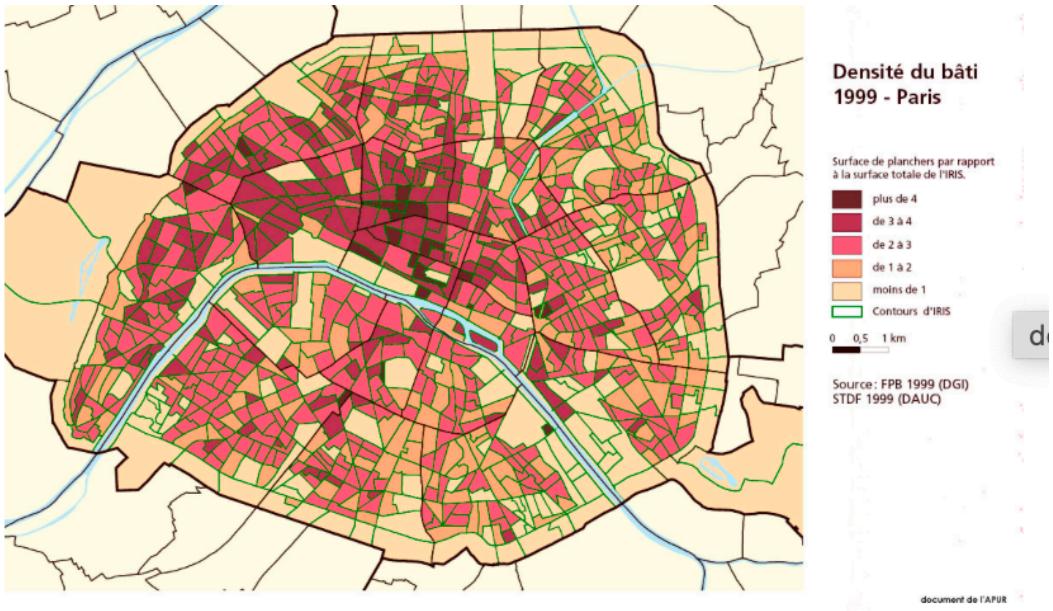


FIGURE 1.3 – Densité du bâti à Paris (1999, source APUR)

la densité du fluide considéré, on peut associer à la collection une nouvelle mesure, de type masse, simplement définie par ses valeurs en chaque entité, $m_i = \rho\mu_i$, la mesure associée, définie comme application de $\mathcal{P}(X)$ dans \mathbb{R}_+ , s'en déduisant simplement par additivité. On a ainsi construit une nouvelle mesure exprimant une variable extensive, construite comme produit d'une première mesure volume avec une variable intensive. On peut dans ce contexte continuer l'empilement des mesures en considérant que chaque particule est animée d'une vitesse u_i . Cette collection de vitesses peut être vue comme une fonction sur X . Cette variable vectorielle intensive peut être adossée avec la mesure m (extensive) pour former une nouvelle variable extensive (la quantité de mouvement), construite selon $p_i = m_i u_i$. Il s'agit de la version discrète de ce que l'on appellera une *mesure vectorielle*. C'est une variable extensive (la quantité de mouvement d'un système est la somme des quantités de mouvement de ses constituants). Dans ce contexte, on dira que la vitesse est mesurable m -presque partout. Ici, l'ensemble étant fini, cela signifie simplement que cela n'a pas de sens de définir la vitesse d'un objet qui n'a pas de masse, puisque cette vitesse sans masse ne pourrait intervenir daucune manière dans un modèle mécanique cohérent.

On remarquera que la variable intensive vitesse peut être intégrée selon cette nouvelle mesure vectorielle, pour former une quantité scalaire qui représente l'énergie cinétique

$$E_A = \langle u, p \rangle_A = \frac{1}{2} \sum_{i \in A} m_i u_i^2.$$

Vers la mesure de Lebesgue

Les cadres présentés ci-dessus peuvent être étendu assez naturellement à des ensembles dénombrables, on remplace alors les sommes finies par des sommes infinies, des séries, de nombres positifs, en acceptant éventuellement que la série puisse prendre la valeur $+\infty$. On remarquera néanmoins que, s'il est possible d'affecter une masse à chaque point d'une collection dénombrable de façon à ce que la masse totale soit *finie*, la distribution est forcément inégalitaire, ou identiquement nulle. En effet, si chaque point de notre ensemble dénombrable a une masse m , on a l'alternative suivante : si $m > 0$ la masse totale est infinie, et si $m = 0$ la masse totale est nulle. Une version temporelle de cet énoncé, qui évoque le paradoxe d'Achille et de la tortue, pourrait être : disposant d'un temps fini, on peut faire une infinité de choses qui chacune prend un certain temps, mais c'est impossible en attribuant un temps

identique à chacune des tâches. On retrouvera cet argument très simple au cœur de la construction d'un des ensembles pathologiques évoqués ci-après.

Les véritables difficultés commencent lorsque l'on s'intéresse à des ensembles qui ont ce que l'on appelle la *pouissance du continu*, comme la droite réelle, ou l'espace physique \mathbb{R}^3 . Considérons pour fixer les idées le cas de l'intervalle réel $X =]0, 1[$. On cherche à définir sur cet ensemble une notion de volume (il s'agit plutôt en l'occurrence d'une notion de longueur, que nous verrons ici comme un volume monodimensionnel). Plus précisément, on cherche à construire une *mesure*, c'est-à-dire une application μ qui à une partie A de $]0, 1[$ associe un nombre réel positif ou nul, et qui généralise à des ensembles quelconques la notion de longueur. On souhaite donc en particulier que $\mu(]a, b[) = b - a$. Le caractère extensif de la notion de longueur impose une propriété d'additivité. On demande donc que la mesure d'une union d'ensembles disjoints soit égale à la somme des mesures des ensembles. Comme nous le verrons plus loin, il est nécessaire pour aboutir à une notion "utilisable" que cette propriété s'étende à des collections dénombrables de parties, on parlera de σ -additivité. L'intervalle fermé $[a, b]$ étant l'intersection des intervalles $]a - 1/n, b + 1/n[$, sa longueur est la même que celle de l'intervalle ouvert. On en déduit que la mesure des singletons (comme les extrémités de l'intervalle) est nulle. On peut étendre immédiatement cette mesure à des réunions dénombrables d'intervalles, mais on se heurte ensuite à un mur. Pour des raisons assez profondes qui tiennent à la nature même de la droite réelle, et malgré l'apparente simplicité du problème, il est *impossible* de définir une telle application, qui affecterait aux intervalles leurs longueurs, qui serait σ -additive (manière distinguée de dire que cela correspond à une variable extensive), qui affecterait à une partie quelconque de l'intervalle¹⁶ $]0, 1[$ ce qu'il conviendrait alors d'appeler sa longueur.

On peut contourner le problème par le haut en suivant un principe inhérent à la notion intuitive de volume : si un ensemble est inclus dans un autre, ce dernier a un plus gros volume. Si l'on se donne $A \subset]0, 1[$, on peut considérer l'ensemble des collections dénombrables d'intervalles (on s'affranchit du caractère disjoint des collections) qui recouvrent A . Si l'on était capable de définir une mesure pour A , cette mesure serait inférieure ou égale à la mesure de toute collection qui recouvre A , qui est elle-même inférieure à la somme des longueurs des intervalles. Il est ainsi naturel de considérer la quantité $\mu^*(A)$ définie comme l'infimum de la somme des longueurs des intervalles, infimum sur l'ensemble des collections qui recouvrent A . On appellera cette quantité la *mesure extérieure de Lebesgue* de A . Cette démarche conduit néanmoins à un problème : il apparaît qu'il existe des parties de X qui vérifient des propriétés que nous qualifierons de *bizarres*. Il existe en effet des ensembles B , dont le complémentaire dans X est noté B^c , qui conduisent à une violation de la propriété d'additivité que l'on souhaite voir vérifier par la mesure. Plus précisément, il existe certaines parties B telles que, pour certaines parties A , l'identité

$$\mu^*(A) = \mu^*(A \cap B) + \mu^*(A \cap B^c)$$

n'est pas vérifiée. Plus précisément $\mu^*(A)$ est strictement inférieur à la somme des mesures des parties disjointes $A \cap B$ et $A \cap B^c$ qui le constituent. Le mathématicien se retrouve dans la position d'un arpenteur étudiant une région A , composée exclusivement de 2 propriétés A_1 et A_2 sans recouvrement, imbriquées l'une dans l'autre de façon extrêmement complexe, et telle que l'aire estimée de A selon la méthode évoquée ci-dessus est *strictement inférieure* à la somme des aires de A_1 et A_2 .

Il n'existe pas de manière complètement satisfaisante de régler ce nouveau problème. La démarche conduisant à des "monstres", on choisit simplement de les exclure de l'approche, et de se concentrer sur les parties B pour lesquelles l'identité ci-dessus est vérifiée pour toute partie A (parties appelées *mesurables*, et dont la collection s'appelle une *tribu* comme on le verra) pour définir une mesure. Cette mesure, qui est la restriction de la mesure extérieure ci-dessus à la collection \mathcal{A} des ensembles mesurables, vérifie alors de bonnes propriétés, au prix de l'*exclusion de certains ensembles pathologiques*, qu'il est d'ailleurs impossible de décrire explicitement¹⁷. Une fois cette construction réalisée, la définition de la notion d'intégrale s'ensuit naturellement. L'intégrale d'une fonction constante égale à ρ (que l'on peut voir ici comme une densité) sur une partie A est simplement le produit $\rho \times \mu(A)$, qui est alors la masse de la matière contenue dans A . On peut étendre facilement cette définition aux

16. On peut aussi formuler ce problème dans le plan \mathbb{R}^2 en considérant à la place des intervalles des rectangles, dont on sait calculer l'aire, ou dans l'espace physique \mathbb{R}^3 en considérant des pavés (i.e. parallélépipèdes), dont on sait calculer le volume.

17. La construction de ces contre-exemples nécessite l'*axiome du choix*, ce qui confère un caractère très abstrait à ces contre-exemples.

fonctions qui prennent un nombre fini de valeurs (fonctions dites *simples*, ou *étagées* dans le cadre de la théorie de la mesure) sur des parties mesurables, en sommant simplement les différentes contributions, comme pour calculer la masse d'un objet composite à partir des densités de ses constituants, et des volumes des différentes zones qu'ils occupent. On peut alors étendre cette notion d'intégrale à une classe très générale de fonctions (nous ne considérons pour l'instant que des fonctions positives), appelées *mesurables*, en définissant l'intégrale comme le supremum des intégrales des fonctions étagées qui sont partout inférieures ou égales à la fonction considérée.

1.4 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

$$p = (p_1, p_2, \dots, p_N), \quad p_i \geq 0, \quad \sum p_i = 1.$$

Definition 1.10. On définit¹⁸ 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.

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 probabilité non uniforme, cette interprétation en termes de *bits* d'information est plus délicate. Considérons l'exemple de la distribution

$$p = \left(\frac{1}{2}, \frac{1}{2(N-1)}, \dots, \frac{1}{2(N-1)}\right).$$

La variable a une chance sur deux de se trouver en première position, avec probabilité 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

18. 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.

à chaque fois la nécessité de la localiser en posant le minimum de questions binaires. La première question sera : est-elle en 1 ? cette question aura une réponse positive en moyenne 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 que nous l'avons définie.

Entropie relative

Si l'on se donne maintenant une loi de référence $\pi = (\pi_1, \dots, \pi_N) \in]0, +\infty[^N$ sur l'ensemble à N éléments, on définit comme suit l'entropie relative de la loi p relativement à π , appelée aussi divergence de Kullback-Leibler de p relativement à π .

Definition 1.11. (Entropie relative)

Soit φ une fonction strictement convexe de \mathbb{R}_+ dans \mathbb{R} . On définit l'entropie de p relativement à π (associée à la fonction φ)

$$S_\pi(p) = KL(p|\pi) = \sum \varphi\left(\frac{p_i}{\pi_i}\right) \pi_i$$

Dans le cas $u \mapsto \varphi(u) = u \log u$, on a

$$S_\pi(p) = KL(p|\pi) = \sum p_i \log\left(\frac{p_i}{\pi_i}\right).$$

Proposition 1.12. Pour toute loi p on a $S_\pi(p) \geq S_\pi(\pi)$, avec inégalité stricte pour $p \neq \pi$.

Démonstration. Par convexité de φ on a

$$\sum \varphi\left(\frac{p_i}{\pi_i}\right) \pi_i \geq \varphi\left(\sum \frac{p_i}{\pi_i} \pi_i\right) = \varphi\left(\sum p_i\right) = \varphi(1) = S_\pi(\pi).$$

L'inégalité est stricte dès que tous les points ne sont pas confondus, i.e. dès que $p \neq \pi$. \square

Remarque 1.13. L'entropie d'une loi définie en début de section comme la somme des $p_i \log p_i$ est l'entropie relative vis-à-vis de la mesure uniforme, à constante additive près, en effet, pour $\pi_i \equiv 1/N$, on a

$$S_\pi(p) = \sum \frac{p_i}{\pi_i} \log\left(\frac{p_i}{\pi_i}\right) \pi_i = \sum p_i \log\left(\frac{p_i}{\pi_i}\right) = \sum p_i \log p_i - \sum p_i \log \pi_i = S(p) + \log N = S(p) - S(\pi).$$

Information mutuelle

On considère 2 variables aléatoires X et Y à valeurs dans deux ensembles finis, on note $\gamma = (\gamma_{ij})$ la loi jointe, μ et ν les marginales. On a

$$\mu_i = \sum_j \gamma_{ij} \text{ et } \nu_j = \sum_i \gamma_{ij}.$$

Definition 1.14. (information mutuelle)

On appelle *information mutuelle* entre les deux variables¹⁹ la quantité

$$\begin{aligned} M(\gamma) = S(\gamma) - S(\mu) - S(\nu) &= \sum_i \sum_j \gamma_{ij} \log(\gamma_{ij}) - \sum_i \mu_i \log(\mu_i) - \sum_j \nu_j \log(\nu_j) \\ &= \sum_i \sum_j \gamma_{ij} \log\left(\frac{\gamma_{ij}}{\mu_i \nu_j}\right) = KL(\gamma|\mu \otimes \nu). \end{aligned}$$

Proposition 1.15. L'information mutuelle associé à une loi jointe γ est positive ou nulle, nulle en cas d'indépendance des deux variables aléatoires (i.e. pour $\gamma = \mu \otimes \nu$). Dans le cas d'ensembles de même cardinal et de marginales uniformes, cette information atteint son maximum pour tous les couplages bijectifs.

Démonstration. On a $M(\gamma) = KL(\gamma|\mu \otimes \nu)$, elle est donc positive et minimale pour $\gamma = \mu \otimes \nu$ d'après la proposition 1.12. Si les ensembles ont même cardinaux N , et les mesures sont uniformes ($\mu_i = \nu_j = 1/N$), alors on peut associer un plan de couplage à toute bijection σ du groupe symétrique S_N :

$$\sigma \in S_N \longmapsto \gamma^\sigma, \quad \gamma_{ij}^\sigma = \frac{1}{N^2} \sigma_{j,\sigma(i)}.$$

Pour tout tel γ^σ , on a

$$M(\gamma^\sigma) = -\log N + 2 \log N = \log N.$$

D'après la proposition 14.5, page 288, ces plans particuliers sont exactement les points extrémaux de $\Pi_{\mu,\nu}$. D'après le théorème de Krein-Milman (théorème 19.10, page 376), tout plan s'écrit comme combinaison convexe de tels points extrémaux. Comme $\gamma \mapsto M(\gamma)$, à marginales fixées, est une fonctionnelle strictement convexe, l'information mutuelle de tout plan diffus (i.e. pas associé à une bijection) est strictement inférieure à la valeur maximal $\log N$. \square

1.5 Éléments de mécanique du point matériel et des systèmes

L'objet essentiel de cette section, la masse ponctuelle, est l'idéalisation d'une certaine quantité de matière pesante occupant un espace très petit, de telle sorte que l'espace qu'elle occupe est assimilé à un point. On fera référence à cet objet idéalisé sous les dénominations de masse ponctuelle, masse, particule, ou simplement point. Comme nous le verrons, ce cadre peut être utilisé pour modéliser de "gros" objets, pour autant que leur forme n'ait pas d'incidence sur les phénomènes auxquels on s'intéresse.

On gardera à l'esprit que cette notion de masse ponctuelle perd toute pertinence lorsque la taille des objets réels devient comparable avec d'autres distances qui interviennent dans le modèle, ce qui est inévitablement le cas par exemple lorsque la distance entre deux masses tend vers 0.

On considère une masse m dont la trajectoire s'écrit $t \mapsto x(t)$ dans un référentiel donné. On note $u = \dot{x}$ la vitesse. On appelle quantité de mouvement de la masse m au temps t , dans le référentiel considéré, le vecteur

$$p(t) = mu(t).$$

Principe 1.16. (Principe Fondamental de la Dynamique, ou Deuxième Principe de Newton)
Il existe un référentiel tel que, si $t \mapsto x(t)$ représente la trajectoire d'une masse au cours du temps dans ce référentiel, et F la force¹⁹ s'exerçant sur cette masse, la relation suivante soit vérifiée :

$$\frac{dp}{dt} = m\ddot{x} = F.$$

Un référentiel dans lequel le principe fondamental de la dynamique s'applique à toute masse ponctuelle est dit *galiléen*. Sauf avis contraire, tous les vecteurs variables au cours du temps intervenant dans la suite seront exprimés dans un référentiel galiléen. Ainsi $t \mapsto x(t)$ désignera la trajectoire dans un référentiel galiléen d'une masse m soumise à une force F .

19. En fait cette information mutuelle ne porte que sur les lois de ces variables, de fait nous l'écrirons comme simple fonction de γ .

20. L'auteur de ces lignes doit reconnaître qu'il lui serait difficile de répondre à la question : "C'est quoi une force ?". Se contenter de dire que c'est une source de quantité de mouvements appauvrit de fait l'énoncé du premier principe.

On notera que tout référentiel en translation à vitesse constante par rapport à un référentiel galiléen est lui-même galiléen, de telle sorte que la vitesse n'est définie que relativement au référentiel choisi²¹.

Definition 1.17. (Énergie cinétique)

On considère une masse m dont la trajectoire s'écrit $t \mapsto x(t)$. On note toujours $u = \dot{x}$ la vitesse. On appelle énergie cinétique la quantité définie par

$$E_c = \frac{1}{2}m|u|^2.$$

Definition 1.18. (Travail, puissance)

Soit $F(t)$ une force s'exerçant en un point en mouvement $x(t)$. La puissance instantanée est définie par

$$\mathcal{P} = F \cdot u.$$

Le travail de la force entre les instants t_1 et t_2 est défini par

$$W_{t_1}^{t_2} = \int_{t_1}^{t_2} \mathcal{P} dt = \int_{t_1}^{t_2} F(t) \cdot u(t) dt = \int_{t_1}^{t_2} P(t) dt.$$

Le travail s'exprime en Joules (J), et la puissance en Watt (W), avec $1 \text{ W} = 1 \text{ Js}^{-1}$.

Théorème 1.19. La variation d'énergie cinétique entre deux instants est égale au travail de la force entre ces deux instants.

Démonstration. On multiplie la relation $m\ddot{x} = F$ par la vitesse \dot{x} . Il vient

$$m\ddot{x} \cdot \dot{x} = \frac{1}{2}m \frac{d|u|^2}{dt} = F \cdot \dot{x} = P(t).$$

On a donc

$$E_c(t_2) - E_c(t_1) = \int_{t_1}^{t_2} \frac{dE_c}{dt} dt = \frac{1}{2}m \int_{t_1}^{t_2} \frac{d|u|^2}{dt} dt = \int_{t_1}^{t_2} P(t) dt = W_{t_1}^{t_2}. \quad \square$$

Exercice 1.11. (Coût énergétique par passager d'un vol)

On cherche à estimer le coût par passager d'un vol en avion. On considère un avion dont la masse à vide est M , transportant N voyageurs de masse m (bagages compris), à une hauteur en vol de croisière de h . L'avion a une *finesse* (rapport portance sur trainée) notée f , et il vole à une vitesse V .

Estimer le coût énergétique *marginal* par passager supplémentaire, ainsi que le coût moyen par passager (en prenant en compte le poids de l'avion qu'il a fallu faire voler pour transporter les passagers).

Valeurs numériques pour l'A380 (valeurs approximatives)

$$M = 400 \text{ t}, \quad N = 500, \quad m = 100 \text{ kg}, \quad h = 10 000 \text{ m}, \quad L = 6 000 \text{ km}, \quad V = 1 000 \text{ kmh}^{-1}, \quad f = P/T = 22.$$

Definition 1.20. (Force dérivant d'un potentiel)

Soit V une fonction de \mathbb{R}^d dans \mathbb{R} . On dit que la force F dérive du potentiel V si l'on a

$$F = -\nabla V = -\left(\frac{\partial V}{\partial x_i}\right)_{1 \leq i \leq d}.$$

Théorème 1.21. On considère une masse m soumise à l'action d'une force découlant d'un potentiel V . La quantité

$$E_{tot} = E_c + V,$$

appelée énergie totale, est conservée au cours du temps.

21. La précaution initiale stipulant que les vitesses sont petites devant la vitesse de la lumière n'a donc pas un sens très clair si l'on s'en tient au formalisme de la mécanique classique. On se reportera aux ouvrages d'introduction à la relativité restreinte pour comprendre le sens de cette hypothèse.

Démonstration. On écrit

$$\frac{d}{dt} E_{tot} = \frac{d}{dt} \left(\frac{1}{2} m |u|^2 + V(x) \right) = mu \cdot \dot{u} + \nabla V \cdot u = (m\dot{u} - F) \cdot u = 0 \quad \square$$

Modèle 1.22. (Particule dans un fluide au repos)

On considère une particule se déplaçant dans un fluide visqueux au repos. La force exercée par le fluide sur la particule peut se modéliser par

$$F_{\text{fluide} \rightarrow \text{particule}} = -\alpha u(t),$$

où α est une constante positive qui dépend du rayon a (supposée petit²²) de la particule, et de la viscosité μ du fluide, selon la loi dite de Fáxen (équation (9.25), page 221) : $\alpha = 6\pi\mu a$.

Modèle 1.23. (Particule dans un fluide en mouvement)

On considère une particule baignant dans un fluide dont la vitesse au point x et à l'instant t s'écrit $U(x, t)$. Dans cette situation l'action du fluide sur la particule peut se modéliser par une force

$$F_{\text{fluide} \rightarrow \text{particule}} = -\alpha (\dot{x} - U(x(t), t)),$$

qui tend à rapprocher la vitesse de la particule de la vitesse du fluide environnant au point correspondant (voir exercice 1.12).

Exercice 1.12. On considère une particule soumise à la seule action d'un fluide visqueux dont la vitesse est U :

$$m\ddot{x} = \alpha(U(x, t) - \dot{x}).$$

On suppose la particule située en x_0 à l'instant τ , et l'on note $U_0 = U(x_0, \tau)$. Montrer que la dérivée instantanée de $|u - U_0|$ au temps τ est négative.

La propriété établie dans l'exercice précédent ne garantit pas que la vitesse de la particule tende vers celle du fluide, car si au temps τ la vitesse se rapproche de la vitesse locale du fluide, cette dernière évolue ensuite le long de la trajectoire de la particule, et il est possible que cette vitesse change significativement sans laisser à la particule le temps de s'adapter à la vitesse locale. On peut identifier deux régimes extrêmes. Dans le premier, avec une particule lourde (et / ou un fluide de viscosité très petite), le fluide n'exerce aucune action significative sur le fluide. Dans le second régime, particule légère et / ou très petite, et / ou viscosité très forte, la relaxation de la vitesse de la particule vers celle du fluide sera quasi-instantanée, de telle sorte que la particule devient ce qu'on appelle un *traceur passif*. Plus formellement, on peut montrer que quand α tend vers $+\infty$, les trajectoires du système précédent tendent à se rapprocher des trajectoires de particules purement convectées par le fluide, et qui sont solutions de l'équation d'ordre 1

$$\dot{y} = U(y, t).$$

Cette propriété fait l'objet de l'exercice 1.17, page 34. Le positionnement entre ces deux régimes extrêmes est conditionné par le *nombre de Stokes* (voir section 5.3, page 102).

Principe 1.24. (Principe de l'action et de la réaction, ou Troisième Principe de Newton)

Si un corps 1 exerce une force f sur un corps 2, alors le corps 2 exerce sur 1 la force $-f$.

Proposition 1.25. On considère un système de particules ponctuelles de masse m_1, \dots, m_N . Chaque masse i est soumise à une force f_i , et les particules interagissent entre elles. On note f_{ij} la force exercée par i sur j . La dérivée en temps de la quantité de mouvement totale, qui est égale à la quantité de mouvement du centre de gravité du système auquel on attribue la masse totale m , est égale à la somme f des forces extérieures.

22. La particule est supposée suffisamment petite pour que l'écoulement du fluide au voisinage soit régi par les équations de Stokes, c'est à dire que le nombre de Reynolds (voir définition 9.12, page 212) basé sur la vitesse relative de la particule par rapport au fluide soit petit devant 1.

Démonstration. On a pour chaque particule

$$m_i \ddot{x}_i = f_i + \sum_{j \neq i} f_{ji},$$

d'où, en sommant toutes ces équations,

$$\sum_i m_i \ddot{x}_i = + \sum_i f_i + \sum_i \sum_{j \neq i} f_{ji} = f$$

car les forces internes s'annulent deux à deux d'après le principe d'action-réaction ($f_{ij} + f_{ji} = 0$).

Le barycentre du système est défini par $X = (\sum m_i x_i) / m$ de telle sorte que l'identité ci-dessus peut s'écrire $m \ddot{X} = f$, ce qui termine la preuve. \square

On se reportera au chapitre 9 pour une présentation des équations régissant le mouvement d'un continuum de matière de type fluide, équations qui résultent du principe fondamental de la dynamique appliquée à un système élémentaire de fluide que l'on peut voir comme un petit paquet d'une infinité de particules regroupées au voisinage d'un même point.

Phénomène de suramortissement

Pour le système masse-ressort présenté ci-dessus, on parlera de suramortissement lorsque le système n'est pas en mesure d'emmagasiner une quantité significative d'énergie cinétique. Dans cette situation, le travail des forces (force de rappel élastique, ou force extérieure exercée sur la masse) est instantanément dissipée sous forme de chaleur. On obtient un tel comportement en faisant par exemple tendre la masse m vers 0. L'équation de conservation de la quantité de mouvement (1.11) devient alors une équation d'ordre 1 en temps, que l'on peut écrire

$$-\mu \dot{x} - kx - mg = 0,$$

qui correspond à un bilan instantané des forces (le système n'est pas en mesure de stocker de la quantité de mouvement).

On pourra écrire les systèmes de ce type sous la forme suivante :

$$\ddot{x} + \frac{1}{\tau} \left(\dot{x} + \frac{1}{\eta} x \right) = f, \quad (1.10)$$

qui fait bien apparaître 2 temps caractéristiques. Pour un système masse-ressort classique, de paramètres m , μ , et k , on a $\tau = m/\mu$ et $\eta = \mu m/k$.

Exercice 1.13. Décrire le comportement des solutions de l'équation (1.10) homogène (avec $f = 0$) en fonction des temps caractéristiques τ et η .

On s'intéresse maintenant à un système masses ressort en considérant que la rigidité du système est encodée par une matrice symétrique définie positive A . Décrire le comportement du système en fonction du spectre de A .

De façon générale, on appellera système suramorti un système pour lequel les effets inertiels sont négligeables, de telle sorte que le bilan énergétique qui détaille le Premier Principe ne contient pas de terme d'énergie cinétique.

Exercice 1.14. (Modélisation : la locomotion)

La locomotion peu être définie comme un ensemble d'actions permettant de se déplacer dans un milieu donné. On s'attachera ici à ne considérer que des effets résultants d'efforts internes au système considéré, i.e. vérifiant la loi de l'action et de la réaction au sein du système²³. Proposer des systèmes (les plus simples possible) de type masses-ressort modélisant les types de locomotion suivants :

23. Selon ce critère strict, une voiture ne peut pas être considérée comme un moyen de locomotion *pour la personne qu'il déplace*, puisque le véhicule exerce une force extérieure sur la personne, mais le système conducteur-véhicule est lui un moyen de locomotion pour lui-même

1. Nage en milieu inertiel
2. Nage en milieu visqueux non inertiel
3. Reptation (en appui sur un support supposé fixe)
4. Propulsion dans le vide

1.6 Éléments de thermodynamique, notion d'énergie

Premier principe de la thermodynamique

Le premier principe de la thermodynamique énonce que l'on peut associer à tout système isolé une quantité appelée *énergie totale*, qui se conserve au cours du temps. Cette énergie totale est la somme de l'énergie cinétique du système et d'une quantité appelée *énergie interne*. Pour un système non isolé, la variation de cette énergie totale est égale à la somme du travail des forces extérieures et de la quantité de chaleur échangée avec le monde extérieur (comptée positivement quand la chaleur est *reçue* par le système).

Exemple 1.6.1. (Système masse-ressort.)

L'exemple simple suivant illustre la manière dont l'expression du premier principe dépend du système que l'on considère. On considère une masse ponctuelle m attachée à un ressort de longueur au repos nulle et de raideur k , dont l'autre extrémité est fixée en O . On suppose la masse assujettie à se déplacer verticalement, et l'on note $x \in \mathbb{R}$ sa position verticale. On suppose la masse plongée dans un fluide visqueux qui exerce une force de résistance au déplacement d'intensité proportionnelle à la vitesse, qui s'écrit $-\mu\dot{x}$. Le principe fondamental de la dynamique s'écrit

$$m\ddot{x} = -kx - mg - \mu\dot{x}. \quad (1.11)$$

Considérons dans le premier temps le système *constitué de la seule masse*, considérée comme non susceptible d'emmagasiner la chaleur. En multipliant l'équation de conservation de la quantité de mouvement par la vitesse, on obtient un bilan énergétique instantané (en W) qui peut s'écrire

$$\frac{d}{dt} \left(\frac{1}{2}\dot{x}^2 \right) = -kx\dot{x} - \mu\dot{x}^2 - mg\dot{x}$$

qui exprime que la dérivée en temps de l'énergie totale (ici constituée de la seule énergie cinétique) est égale au travail des forces extérieures (force du ressort, force de frottement fluide, et poids).

Si l'on considère maintenant le système constitué de la masse et du ressort, on écrira

$$\frac{d}{dt} \left(\underbrace{\frac{1}{2}m\dot{x}^2}_{E_c} + \underbrace{\frac{1}{2}kx^2}_{E_k} \right) = -\mu\dot{x}^2 - mg\dot{x}.$$

L'énergie totale inclut maintenant une composante interne, qui est l'énergie potentielle élastique du ressort, les forces extérieures étant réduite à la friction et à la force de gravité.

Si l'on considère maintenant le système masse – ressort – terre, il est pertinent d'interpréter le travail du poids $mg\dot{x}$, comme la dérivée en temps d'une énergie potentielle gravitationnelle $E_g = mgx$, qui vient compléter l'énergie interne de notre nouveau système. Le bilan énergétique, qui exprime le premier principe appliqué à ce nouveau système, s'écrit

$$\frac{d}{dt} \left(\underbrace{\frac{1}{2}m\dot{x}^2}_{E_c} + \underbrace{\frac{1}{2}kx^2}_{E_k} + \underbrace{mgx}_{E_g} \right) = -\mu\dot{x}^2.$$

Le membre de gauche est la dérivée en temps de l'énergie totale, somme de l'énergie cinétique, de l'énergie potentielle élastique et de l'énergie potentielle gravitationnelle. Le membre de droite $-\mu\dot{x}^2$ s'interprète comme la puissance de la force exercée par le fluide (qui est *extérieur* au système), sur la masse. On notera qu'il est impossible de l'écrire comme la dérivée d'une fonction qui ne dépendrait que de la position (i.e. de l'état) du système. En effet la dérivée de $g(x)$ est de la forme $g'(x)\dot{x}$, qui dépend linéairement (et pas quadratiquement) de \dot{x} . On ne peut donc pas écrire ce terme comme l'opposé d'une fonction de l'état qui pourrait être interprété comme une composante de l'énergie interne du système.

Si l'on considère pour finir le système complet masse – ressort – terre – fluide, on est amené à interpréter ce terme comme une puissance dissipée : les forces de friction vont induire une production de chaleur au sein du fluide, au voisinage de la masse, qui vont faire augmenter la température. Si l'on note c_f la capacité thermique du fluide au voisinage de la masse, et par T_f sa température moyenne (prise égale à 0 à l'état initial), l'énergie thermique emmagasinée s'écrit $c_f T_f$.

Le bilan énergétique pour ce système global, considéré comme isolé, exprime alors la conservation de l'énergie totale

$$\frac{1}{2}m\dot{x}^2 + \underbrace{\frac{1}{2}kx^2 + mgx}_{\text{én. potentielle interne}} + \underbrace{c_f T_f}_{\text{én. thermique interne}} = .$$

Comme on le verra plus loin, le second principe établit une hiérarchie entre ces différentes énergies, excluant par exemple le transfert spontané d'énergie thermique vers les formes potentielle ou cinétique.

L'exercice 1.16 propose une généralisation de ce qui précède à un terme de rappel non linéaire.

Exemple 1.6.2. (La terre et ce qui l'entoure)

Le Premier Principe s'applique à tous les systèmes, y compris les systèmes si complexes qu'il est impossible d'en comprendre des détails. Considérons ici le système terre-atmosphère. On peut considérer que système n'échange de l'énergie avec l'extérieur que selon deux modalités : rayonnement reçu en provenance du soleil (essentiellement dans le spectre visible), et rayonnement émis vers l'extérieur (spectre infrarouge principalement). Ce second rayonnement, compté négativement dans le bilan (il s'agit d'énergie perdue par le système), est la somme d'un rayonnement réfléchi instantanément par la terre (surfaces recouvertes de glace par exemple), et rayonnement émis (essentiellement dans le spectre infrarouge), qui dépend de la température de surface, et dont la partie qui part effectivement vers l'extérieur (celle que l'on doit prendre en compte dans le bilan), dépend. Le bilan de ces énergie reçue et évacuée peut s'écrire $Q_{in} - Q_{out}$, on parle de *forçage radiatif*. Si cette quantité est strictement positive, le premier principe implique une augmentation de l'énergie totale du système. Ce principe ne décrit pas en détail les différentes formes d'énergie qui constituent l'énergie interne, et ne dit rien sur les importances respectives de ces énergies. Il garantit en revanche que, si l'on estime les variations des différents types d'énergie interne, et que la variation induite d'énergie totale ne correspond pas au forçage radiatif, c'est que l'on s'est trompé, par exemple en ne prenant pas en considération l'une des formes d'énergie interne. Notons en particulier que, même si l'énergie thermique est la composante principale de l'énergie interne, une part (petite en l'occurrence) du rayonnement reçu peut être par exemple stocké au sein de matière organique créés par synthèse chlorophyllienne. Cette énergie pourra convertie en chaleur et en énergie mécanique ultérieurement par la combustion de cette matière, ou l'ingestion de végétaux alimentaire par un être vivant, au sein duquel cette énergie sera convertie en force motrice et en chaleur. Les mouvements des océans et de l'atmosphère participent à des échanges permanents entre des énergies thermique et mécanique. La *convection naturelle* en particulier correspond à un mouvement vers le haut d'air chaud, et convertit de l'énergie thermique en énergie cinétique. Inversement la dissipation visqueuse au sein des fluides transforme de l'énergie mécanique en énergie cinétique.

Nous n'avons pas parlé de l'énergie cinétique de la terre dans sa globalité. A priori l'énergie totale de notre système devrait être considérée comme étant la somme de cette énergie cinétique liée au mouvement d'ensemble de notre planète, et des énergies évoquées ci-dessus. On peut vérifier que les flux d'énergie annuel en jeu sont très inférieurs à cette énergie cinétique d'ensemble²⁴. On traite

séparément ces deux énergies, plus précisément on ne prend pas du tout en compte l'énergie cinétique de la terre dans le bilan, du fait qu'il n'existe aucun mécanisme de transfert entre l'une et l'autre de ces énergies, de telle sorte qu'il est pertinent, dans ce contexte, de considérer la terre comme immobile. On se reportera à la section 5.1, page 99, pour un développement plus détaillé de ce bilan d'énergie.

Second principe de la thermodynamique

Le deuxième principe de la thermodynamique énonce l'existence d'une quantité, appelée *entropie* et noté S , qui croît²⁵ au cours de toute transformation réelle d'un système isolé.

Pour un système à la température T qui échange la quantité de chaleur δQ avec le milieu extérieur, la variation d'entropie est $\delta Q/T$.

Remarque 1.26. On prendra garde au fait que, dans l'expression ci-dessus, on considère que la température est exprimée en Kelvin (K), de telle sorte qu'une température nulle correspond bien au *zéro absolu*, c'est à dire à l'absence complète d'agitation interne.

Transfert de chaleur entre deux corps de température différente Considérons 2 objets en contact, aux températures respectives T_1 et T_2 , avec $T_1 < T_2$, en supposant le système 1 + 2 fermé vis à vis du monde extérieur. Notons δQ la quantité de chaleur échangée entre ces deux corps pendant un court instant. On se place du point de vue de 1, en considérant que δQ est positif si de l'énergie est transférée de 2 vers 1 (négatif dans le cas contraire). D'après le premier principe, qui assure la conservation de l'énergie globale du système, le corps 2 reçoit une quantité de chaleur $-\delta Q$. Les variations d'entropies respectives, et la variation d'entropie globale, s'écrivent

$$dS_1 = \frac{\delta Q}{T_1}, \quad dS_2 = -\frac{\delta Q}{T_2}, \quad dS = dS_1 + dS_2 = \delta Q \left(\frac{1}{T_1} - \frac{1}{T_2} \right).$$

Le second principe énonce une augmentation de l'entropie, d'où, du fait que $T_2 > T_1$, $\delta Q \geq 0$. Le second principe impose donc que la chaleur se propage de 2 vers 1, c'est à dire du *chaud vers le froid*. On peut aussi formuler ce principe de transfert du chaud vers le froid de façon plus informelle, comme étant le seul compatible avec le second principe : l'entropie du corps froid augmente, l'entropie du corps chaud diminue, mais moins que celle du corps froid n'augmente, du fait que la quantité de chaleur est divisée par la température du corps considéré.

Noter qu'un transfert d'énergie du froid vers le chaud ne violerait pas le premier principe, qui ne porte que sur une conservation globale de l'énergie, et ne dit rien sur les sens de transfert des énergies au sein du système.

L'exercice 1.19, page 35, porte sur les variations d'entropie associées au transfert de chaleur entre deux objets.

Les principes énoncés ci-dessus ne suffisent pas en général à écrire un modèle d'évolution pour un système. En revanche, lorsque l'on écrit un modèle, sa recevabilité physique est conditionnée au fait que les deux principes soient respectés. On pourra se reporter à l'exercice 20.1, page 411, qui porte sur une classe de modèles pour un univers limité à des échanges de chaleur entre objets, pour une illustration de la manière dont ces principes permettent de préciser les conditions de recevabilité de modèles de transfert thermique.

24. Cette énergie peut-être estimée à

$$\frac{1}{2}mv^2 = \frac{1}{2}6 \times 10^{24} \times (3 \times 10^4)^2 = 2.7 \times 10^{33} \text{ J} = 7.5 \times 10^{20} \text{ GWh}.$$

A titre de comparaison, le forçage radiatif dû au CO₂, de l'ordre de $+2 \text{ Wm}^{-2}$, correspond à un surplus d'énergie en 1 an de

$$2 \times (4 \times \pi \times (6.4 \times 10^6)^2) \times 3600 \times 24 \times 365 = 3.2 \times 10^{22} \text{ J} = 9 \times 10^{15} \text{ kWh} = 9 \times 10^9 \text{ GWh}.$$

La quantité totale d'énergie solaire reçue par la terre, estimée à 341 Wm^{-2} , correspond à une énergie reçue par année 341/2 fois supérieure, soit $1.5 \times 10^{12} \text{ GWh}$, qui reste très inférieure à l'énergie cinétique de la terre.

25. On prendra garde au fait que, dans un contexte mathématiques, l'usage est de définir l'entropie comme l'opposé de l'entropie physique, de telle sorte que l'entropie dite mathématique décroît pour un système fermé.

1.7 Exercices

Exercice 1.15. (Passage Lagrange → Euler et Euler → Lagrange pour un réseau linéaire)

On s'intéresse ici à une ligne de bus, dont on note $\{0, \dots, N\}$ les arrêts (on se focalise sur les déplacements vers les indices croissants). On note μ_j le nombre de personnes qui montent dans le bus à l'arrêt j , ν_j le nombre de personnes qui en descendent, et ρ_j le nombre de voyageurs dans le bus entre l'arrêt j et l'arrêt $j+1$. On notera $\mu, \nu \in \mathbb{R}^{N+1}$ et $\rho \in \mathbb{R}^N$ les vecteurs associés (N.B. : on pourra raisonnablement considérer que $\nu_0 = \mu_N = 0$). On note $\gamma = (\gamma_{ij})$ le plan origine-destination : γ_{ij} est le nombre de personnes montées en i qui descendent en j . On définit le plan origine destination normalisé $\tilde{\gamma}$ par

$$\tilde{\gamma}_{ij} = \frac{\gamma_{ij}}{\mu_i} \quad i < N, \quad i < j \leq N.$$

Ainsi défini, $\gamma_i = (\gamma_{ij})_j$ est une loi de probabilité supportée par $\{i+1, \dots, N\}$, qui décrit la distribution des destinations des voyageurs montant en i . On note enfin $b_j = \mu_j - \nu_j$ le bilan montée-descente à l'arrêt j .

1) Écrire $\rho = (\rho_j)$ en fonction de $b = (b_j)$, ainsi que la réciproque. (On écrira les relations entre les ρ_j et les b_j , puis on proposera une formulation matricielle).

2)a) Écrire la relation entre γ , μ , et b sous la forme

$$G\mu = b,$$

où G est une matrice qui dépend du plan renormalisé $\tilde{\gamma}$, dont on précisera l'écriture.

b) Montrer que si l'on connaît $\rho = (\rho_j)$ (historique de l'occupation du bus sur les différents tronçons) et $\tilde{\gamma}$, on peut reconstruire μ .

c) La reconstruction précédente est-elle stable ?

3)a) On ne fait plus d'hypothèse sur $\tilde{\gamma}$. On suppose que l'on est capable de mesurer les entrées-sorties à chaque station (i.e. μ et ν sont supposés connus). Décrire le plus précisément possible l'ensemble des plans OD renormalisés $\tilde{\gamma}$ compatibles avec les observations. (N.B. : on pourra introduire les quantités ρ_j^i qui représentent le nombre de voyageurs montés en i présents dans le bus sur le tronçon $[j, j+1]$).

b) Décrire un algorithme de reconstruction explicite de $\tilde{\gamma}$ sous une hypothèse First In First Out (FIFO), c'est à dire que tout individu sort avant tous les voyageurs montés après lui.

c) Même question sous une hypothèse Last In First out, i.e. le personnes qui sortent en premier sont celles qui sont montées le plus récemment.

4) Proposer une formulation continue du cadre précédent, i.e. la situation d'un bus imaginaire dont l'ensemble des arrêts est un continuum.

Exercice 1.16. On considère un ressort attaché en O , caractérisé par la relation $f = -\varphi(x)$, où x est la position de son autre extrémité, et φ est une fonction continue strictement croissante, nulle en 0. On considère le système masse-ressort associé, avec frottement :

$$m\ddot{x} + \mu\dot{x} + \varphi(x) = 0.$$

a) Établir le bilan énergétique de ce système masse-ressort.

b) A quelle condition peut-on affirmer que la position x de la masse reste bornée au cours du temps ?

Exercice 1.17. (★★)

On considère l'écoulement stationnaire d'un fluide décrit par son champ de vitesse $x \in \mathbb{R}^d \mapsto U(x)$, et l'on s'intéresse au mouvement d'une particule de masse m entraînée par ce fluide selon le modèle²⁶

$$m\ddot{x} = -\mu(\dot{x} - U(x)).$$

26. On se reportera à la loi de Faxen (équation (9.25), page 221), ainsi qu'à la remarque 9.25, sur la pertinence du modèle

On suppose la particule située en x^0 à $t = 0$, et animée d'une vitesse initiale u^0 . Montrer que, quand la masse m tend vers 0 (ou de façon équivalente quand le coefficient μ tend vers $+\infty$), la trajectoire $t \mapsto x(t)$ converge uniformément sur tout intervalle du type $[0, T]$ vers la trajectoire $t \mapsto y(t)$ correspondant à un traceur passif :

$$\begin{cases} \dot{y}(t) &= U(y(t)), \\ y(0) &= x^0. \end{cases}$$

Exercice 1.18. (Potentiel d'interaction)

On considère la fonction $D(\cdot)$ qui à $q = (q_1, q_2) \in \mathbb{R}^3 \times \mathbb{R}^3$ (attention, q_1 et q_2 désignent ici des points de \mathbb{R}^3) associe la distance entre les points q_1 et q_2 de \mathbb{R}^3 :

$$D(q) = D(q_1, q_2) = |q_2 - q_1|.$$

a) Montrer que $D(\cdot)$ est différentiable sur l'ouvert

$$U = \{ q = (q_1, q_2) \in \mathbb{R}^6, q_1 \neq q_2 \},$$

et exprimer son gradient (on pourra exprimer les gradients partiels ∇_{q_1} et ∇_{q_2}) en fonction du vecteur unitaire $e_{12} = (q_2 - q_1) / |q_2 - q_1|$.

b) On introduit un potentiel d'interaction sur le système de deux particules localisées en q_1 et q_2 sous la forme $V = V(q) = V(q_1, q_2) = \varphi(D(q_1, q_2))$, où φ est une fonction continûment dérivable de \mathbb{R}_+ dans \mathbb{R} . Montrer que V est différentiable sur U , et écrire son gradient.

c) Préciser les gradients partiels $\nabla_{q_1} V$ et $\nabla_{q_2} V$ si l'on prend pour φ le potentiel d'interaction gravitationnelle défini par $\varphi(D) = -1/D$.

d) On se replace dans le cas général d'un potentiel φ quelconque, et l'on considère maintenant un système de N particules dans \mathbb{R}^3 . On définit un potentiel d'interaction global de la façon suivante

$$V(q) = V(q_1, q_2, \dots, q_N) = \sum_{1 \leq i < j \leq N} \varphi(D(q_i, q_j)).$$

On s'intéresse au système résultant du principe fondamental de la dynamique, sous l'hypothèse de forces dérivant d'un potentiel (on prend des masses unitaires), c'est-à-dire

$$\frac{d^2q}{dt^2} = -\nabla V(q). \quad (1.12)$$

Écrire l'équation qui résulte de ce principe pour chacune des particules, qui s'écrit de façon abstraite

$$\frac{d^2q_i}{dt^2} = -\nabla_{q_i} V(q).$$

e)(*) On se place dans le cadre des notations de la question précédente. On suppose que l'on connaît une solution $t \in [0, T[\mapsto q(t) \in U$ de l'équation d'évolution (1.12). Montrer que l'on a conservation de l'énergie totale, c'est à dire que la quantité

$$E(t) = \sum_{i=1}^N \frac{1}{2} \left| \frac{dq_i}{dt}(t) \right|^2 + V(q(t))$$

est constante sur $[0, T[$.

Dans le cas du potentiel gravitationnel $\varphi(D) = -1/D$, peut-on en déduire que les vitesses sont majorées sur $[0, T[$? Même question pour le cas du potentiel coulombien entre charges identiques $\varphi(D) = 1/D$.

Exercice 1.19. On se place dans le cadre du système à deux corps (isolé en tant que système) présenté ci-dessus, aux températures initiales $T_1 < T_2$. On s'intéresse à des transformations telles que la température reste uniforme dans chaque corps. On suppose que l'énergie thermique dans chaque corps est produite d'une capacité thermique C commune aux deux corps et de la température courante.

- a) Pour quelle raison peut-on affirmer que les températures des corps à un instant donné peuvent s'écrire $T_1 + T$ et $T_2 - T$, avec $T \in \mathbb{R}$?
- b) Écrire l'entropie du système en fonction de T , et préciser à quelle situation correspond la valeur de T qui maximise cette entropie. Quels sont les états possibles du système (caractérisés par la température T) selon le second principe de la thermodynamique ?
- c) Que peut-on dire si les capacités thermiques des 2 corps sont différentes ?

Chapitre 2

Réseaux résistifs

Sommaire

2.1	Modèles, motivations	37
2.2	Cadre formel, problème de Laplace discret	39
2.3	Opérateurs discrets : gradient, divergence, et laplacien	45
2.4	Résistance équivalente d'un réseau	48
2.5	Cadre stochastique	51
2.6	Cadre abstrait et modélisation	55
2.7	Squelette métrique associé à un réseau résistif	57
2.8	Plongement dans l'espace euclidien	58
2.9	Premier pas vers le transport branché	59
2.10	Réseaux infinis	60
2.11	Remarques diverses	61
2.11.1	Réseau résistifs comme espace métrique mesuré	61
2.12	Exercices	62

2.1 Modèles, motivations

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).

Ce cadre est adapté à la description d'un grand nombre de situations, que nous décrivons ici succinctement en utilisant les notations habituelles pour chaque contexte, avant de passer à des notations unifiées pour la construction du cadre général (section 2.2 ci-après). Le terme de réseau résistif est lié au contexte électrique : on considère des fils électriques caractérisés par une certaine résistance, l'intensité I traversant chaque fil étant lié à la tension U (différence de potentiel électrique entre les extrémités) par la *loi d'Ohm* :

$$U = RI$$

où U est en *volt* (V), I en *ampère* (A), et R est la résistance du conducteur, en *ohms* (Ω). La loi des nœuds, ou loi de Kirchhoff, exprime un bilan d'intensité au niveau de chaque point de bifurcation.

Dans le cas de l'*écoulement d'un fluide visqueux incompressible*, c'est la *pression* aux noeuds qui jouera le rôle du potentiel, dont la différence induit un flux selon la loi dite de Poiseuille, qui exprime la proportionnalité entre le différentiel de pression entre les extrémités du tuyau et le flux de fluide au travers du tuyau :

$$P_{in} - P_{out} = RQ,$$

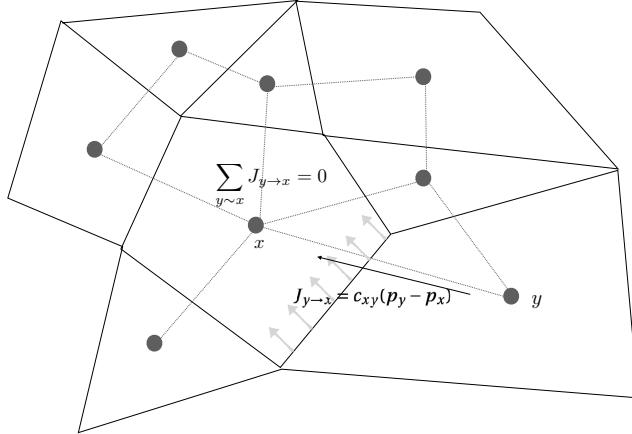


FIGURE 2.1 – Diffusion entre cellules.

où les pressions sont en *pascal* (Pa), qui sont des newton par mètre carré (Nm^{-2}), le débit en m^3s^{-1} , et la résistance en Pa m^{-3} s, où des unités équivalentes selon le contexte¹.

Dans un contexte de thermodynamique, on peut considérer des corps en contact. Chaque corps i est caractérisé par une température T_i (considérée comme étant uniforme sur l'objet), susceptible de varier au cours du temps. Si l'on considère que le flux de quantité de chaleur allant de j vers i obéit à une loi de Fourier discrète, c'est-à-dire qu'il est proportionnel à la différence des températures, on obtient une loi du type de celles qui précédent :

$$Q_{j \rightarrow i} = K_{ij}(T_j - T_i),$$

où K_{ij} est une conductance thermique, exprimée² en W K^{-1} . L'équilibre thermique correspondra à la situation où tous les flux afférents à un même corps s'équilibreront, on retrouve notre loi des noeuds

$$\sum_{j \sim i} Q_{j \rightarrow i} = 0.$$

Dans ce contexte, le problème hors équilibre est pertinent : si l'on considère que chaque corps a une capacité thermique C_i , l'équation d'évolution (en W) s'écrit

$$C_i \frac{dT_i}{dt} = - \sum_{j \sim i} K_{ij}(T_i - T_j).$$

Dans un contexte très voisin, on peut considérer un ensemble de cellules au sein desquelles une substance diffuse très rapidement, de telle sorte que la concentration dans chaque cellule puisse être considérée comme uniforme (voir figure 2.1). Ces cellules sont séparées par une membrane au travers de laquelle la diffusion de fait plus difficilement, selon une loi de type Fick : le flux d'une cellule à l'autre est proportionnel à la différences de concentrations³, avec une constante de proportionnalité qui dépend de l'air de la membrane de sa composition, et de son épaisseur. On définit le graphe non contient (x, y)

1. Ainsi, en pneumologie, du fait qu'historiquement les pressions étaient mesurées à l'aide d'un dispositif basé sur une colonne d'eau, les pressions sont encore couramment exprimées en centimètres d'eau (cmH_2O), un centimètre d'eau valant 100 Pa ($1 \text{ cmH}_2\text{O} = 1 \text{ hPa}$). Les volumes en jeu (pour la ventilation humaine) étant de l'ordre du litre, les débit sont exprimés en L s^{-1} , et subséquemment les résistances en $\text{cmH}_2\text{O L}^{-1} \text{ s}$.

2. On retrouve cette unité dans les spécifications des matériaux d'isolation pour la construction. Un revêtement permettant par exemple d'isoler thermiquement un mur d'habitation ou un toit sera caractérisé par une conductance par unité de surface, en $\text{W K}^{-1}\text{m}^{-2}$. On utilise en fait plus couramment son inverse, sous la forme d'une résistance (d'autant plus grande que le matériau isolant est de qualité), en KW^{-1}m^2 . La présence dans l'unité de m^2 est due au fait que la surface isolante peut être vue comme une infinité de résistances (sur des surfaces infinitésimales) en *parallèle*. L'avantage de cette quantité est qu'elle peut être sommée si l'on dispose de plusieurs couches de résistances différentes l'une sur l'autre (résistances en série).

si les cellules x et y partagent une interface. Si l'on note p_x la concentration (ou pression partielle) de la cellule x , et c_{xy} la perméabilité de la membrane (c_{xy} joue le rôle d'une *conductance* dans un réseau électrique ou fluide), le flux de y à x s'écrit

$$J_{y \rightarrow x} = c_{xy}(p_y - p_x).$$

Si l'on note C_x le volume de la cellule x , le produit $C_x p_x$ est la quantité présente dans x , et le bilan de matière en x s'écrit

$$V_x \frac{dp_x}{dt} = - \sum_{y \sim x} c_{xy}(p_x - p_y).$$

Cela conduit à l'équation globale (de type équation de la chaleur discrète)

$$C \frac{dp}{dt} + Lp = 0,$$

où, comme il sera détaillé plus loin, L est un laplacien discret, définit par

$$L : p \in \mathbb{R}^V \longmapsto Lp = \left(\sum_{y \sim x} c_{xy}(p_x - p_y) \right)_{x \in V}$$

Dans tous les cas, la démarche s'appuie sur une loi phénoménologique, qui exprime la proportionnalité en le flux d'une certaine quantité extensive et la différence entre les valeurs d'une variable intensive jouant le rôle de potentiel, et un principe de conservation de la quantité extensive. Comme nous le verrons, ces deux propriétés sont d'une certaine manière *adjointes* (ou *transposées*) l'une de l'autre, ce qui conférera au problème résultant une structure *variationnelle*, plus précisément on pourra identifier les champs de potentiels solutions aux minimiseurs d'une certaine fonctionnelle quadratique.

2.2 Cadre formel, problème de Laplace discret

Les modèles présentés ici reposent sur la notion de *graphe non orienté*. On se reportera au chapitre 10 pour un définition précise de cet objet.

Definition 2.1. (Réseau résistif)

Un réseau résistif fini est défini comme un triplet $\mathcal{N} = (V, E, r)$, où V est un ensemble fini de sommets (*Vertices*), $E \subset V \times V$ un ensemble d'arêtes (*Edges*) supposé symétrique⁴ :

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

et $r \in]0, +\infty[^E$ est le champ des résistances, défini sur E (avec $r_{xy} = r(y, x)$ pour tout $(x, y) \in E$). On notera $\mathcal{N} = (V, E, r, \Gamma)$ un réseau dans lequel on distingue comme frontière une partie non vide $\Gamma \subset V$, et $\mathcal{N} = (V, E, r, o, \Gamma)$ dans lequel on singularise et extrait un point particulier de Γ .

L'ensemble $V \setminus \Gamma$ ou $V \setminus (\{o\} \cup \Gamma)$ des sommets intérieurs est noté $\overset{\circ}{V}$, il correspond aux sommets (ou noeuds) 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 : $u_{xy} = -u_{yx}$.

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

$$p_x - p_y = r_{xy} u_{xy} = r_e u_e.$$

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

En tout point x intérieur, i.e. qui n'échange pas de matière avec l'extérieur, la loi de Kirchhoff (ou loi des noeuds), qui exprime la conservation de la matière (ou de l'intensité électrique) s'écrit

$$\sum_{y \sim x} u_{xy} = 0,$$

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

Proposition 2.2. On considère un champ de flux défini sur le réseau (V, E, Γ) (les résistances n'interviennent pas dans cette propriété), conservatif en tous les points intérieurs, i.e.

$$\sum_{y \sim x} u_{yx} = 0 \quad \forall x \in \mathring{V}.$$

Le bilan de flux au travers de Γ est alors nul.

Démonstration. On somme simplement l'identité de conservation ci-dessus sur tous les points intérieurs :

$$\sum_{x \in \mathring{V}} \sum_{y \sim x} u_{yx} = 0.$$

Dans la somme ci-dessus, chaque arête reliant deux points intérieurs apparaît 2 fois, dans des sens opposés, il ne reste donc que les termes sur les arêtes dont l'une des extrémités est dans Γ . On obtient donc

$$\sum_{y \in \Gamma} \sum_{x \sim y} u_{yx} = 0,$$

qui assure que le flux sortant au travers de Γ est nul. (N.B. : nous avons implicitement supposé ici que les points de Γ n'étaient pas reliés entre eux. Si certains le sont, alors l'expression ci-dessus doit être complétée par une somme de flux directs entre points de Γ , qui s'annulent deux à deux, ce qui ne change pas la propriété annoncée. \square

Remarque 2.3. Dans le cas d'un réseau enraciné, le flux rentrant au travers de Γ est égal au flux sortant au travers de la racine o .

On s'intéresse au problème consistant à calculer les pressions et les flux sur l'ensemble du réseau, quand les pressions sont prescrites sur Γ (il n'y a pas de lieu ici de distinguer une racine o). Les deux lois écrites ci-dessus se combinent pour former un problème de Darcy discret

$$\begin{cases} u_{xy} + c_{xy}(p_y - p_x) &= 0 \quad \forall (x, y) \in E, \\ \sum_{y \sim x} u_{yx} &= 0 \quad \forall x \in \mathring{V} \end{cases} \quad (2.1)$$

avec une pression imposée sur Γ .

Après élimination de la vitesse, en combinant les deux lois ci-dessus, on obtient un problème de Poisson discret pour la pression, avec conditions de Dirichlet :

$$\begin{cases} Lp &= 0 \quad \text{sur } \mathring{V}, \\ p &= P \quad \text{sur } \Gamma, \end{cases} \quad (2.2)$$

où P est une collection de pressions prescrites sur la frontière Γ , et L est l'opérateur de Laplacien discret associé au réseau, défini par

$$p \in \mathbb{R}^V \longmapsto Lp \in \mathbb{R}^V, \quad (Lp)_x = \sum_{y \sim x} c_{xy}(p_x - p_y). \quad (2.3)$$

Proposition 2.4. (Principe du maximum)

On se place sur un réseau (V, E, r, Γ) connexe. Soit $p \in \mathbb{R}^V$ un champ harmonique sur \mathring{V} , i.e. tel que $Lp = 0$ sur \mathring{V} . Le maximum de p est alors atteint sur la frontière Γ

Démonstration. C'est une conséquence directe de l'harmonicité, qui s'écrit, pour tout point $x \in \mathring{V}$,

$$\sum_{y \sim x} c_{xy} (p_x - p_y) = 0,$$

d'où

$$p_x = \frac{\sum c_{xy} p_y}{\sum c_{xy}} = \sum \theta_y p_y, \quad \theta_y \geq 0, \quad \sum \theta_y = 1, \quad (2.4)$$

ce qui exprime que p , en tout point de \mathring{V} , est combinaison convexe des valeurs aux points voisins. Si le maximum est atteint en un point intérieur, alors la valeur est la même en tous ses voisins. Si ces voisins sont tous intérieurs, on poursuit la démarche jusqu'à atteindre le bord (le nombre maximal d'étapes à effectuer est la longueur combinatoire du plus court chemin à la frontière). \square

Théorème 2.5. On suppose le réseau \mathcal{N} connexe. Le problème (2.2) est alors bien posé, et la correspondance $P \mapsto p$ est linéaire.

Démonstration. Nous indiquons ci-dessous deux approches permettant d'aborder ce problème, une première basée sur le principe du maximum, qui permet d'établir l'unicité de la solution, et donc aussi son existence du fait qu'il s'agit d'un problème linéaire avec autant d'équations que d'inconnue. Cette démonstration est fragile, car elle utilise le fait que l'on soit en dimension finie, ainsi que la linéarité du problème. La deuxième approche utilise des principes d'optimisation, en identifiant le problème aux conditions d'optimalité pour un problème de minimisation d'une fonctionnelle. Son principe peut s'étendre à la dimension finie, ainsi qu'à des problèmes non linéaires (voir exemple l'exercice 2.11, page 63).

1) La première approche consiste à utiliser le principe du maximum. Le système d'équations linéaires

$$\sum_{y \sim x} c_{xy} (p_x - p_y) = 0 \quad \forall x \in \mathring{V},$$

peut s'écrire sous forme matricielle

$$\mathring{L}\mathring{p} = b,$$

où \mathring{p} désigne ici le vecteur des pressions inconnues (pressions sur \mathring{V}), \mathring{L} est une matrice carrée (d'ordre égal au nombre de points intérieurs), et b est construit à partir des valeurs imposées sur Γ , on a plus précisément

$$(\mathring{L}\mathring{p})_x = \sum_{y \sim x, y \notin \Gamma} c_{xy} (p_x - p_y) + \sum_{y \sim x, y \in \Gamma} c_{xy} p_x, \quad b_x = \sum_{y \sim x, y \in \Gamma} c_{xy} P_y,$$

où P_y est la valeur de la pression imposée en $y \in \Gamma$ (il ne s'agit pas d'une inconnue). Montrons que \mathring{L} est injective. Si $\mathring{L}\mathring{p} = 0$, on prolonge \mathring{p} par 0 pour obtenir un champ p harmonique sur le réseau. Ce champ atteint donc son maximum sur le bord⁵, d'après la proposition 2.4, il est donc nul. Le raisonnement s'applique aussi au minimum, le champ est donc identiquement nul sur V . La matrice \mathring{L} est donc inversible, et le problème est bien posé.

2) Pour la seconde approche, nous introduisons la fonctionnelle d'énergie (qui correspond pour un réseau électrique ou hydraulique à la puissance instantanée dissipée)

$$\Phi(p) = \sum_e c_{xy} (p_x - p_y)^2 = \Phi(\mathring{p}, p_\Gamma).$$

Pour $p_\Gamma = P_\Gamma$ fixé, nous introduisons la fonctionnelle $J(\mathring{p}) = \Phi(\mathring{p}, P_\Gamma)$. Il s'agit d'une fonctionnelle quadratique. Montrons qu'elle est coercive, selon la définition 13.3. Du fait de son caractère quadratique, il suffit de vérifier que la matrice symétrique associée à sa partie quadratique est non dégénérée, ce qui peut se faire comme précédemment en montrant l'injectivité. Nous proposons une démonstration

5. C'est ici qu'intervient l'hypothèse de connexité. Si le réseau n'était pas connexe, on pourrait avoir des composantes connexes qui ne contiennent aucun point de Γ , donc aucun point du bord. On pourrait donc avoir une valeur constante arbitraire sur cette composante.

différente, qui présente l'avantage de s'étendre à des fonctionnelles plus générales. Elle se base sur une estimation des valeurs de p à partir des valeurs au bord, et de la valeur de la fonctionnelle. Il s'agit d'une version discrète de ce que l'on appelle *l'inégalité de Poincaré* dans les espaces de Sobolev⁶,

Soit $p = (\hat{p}, P_\Gamma) \in \mathbb{R}^V$, et x un sommet de \mathring{V} . Ce sommet x est relié à la frontière par un chemin

$$x_0 \in \Gamma \sim x_1 \sim \cdots \sim x_n = x.$$

On cherche maintenant à majorer $|p_x|$ par une quantité faisant apparaître la valeur de la fonctionnelle en p .

$$\begin{aligned} |p_x| &= \left| p_{x_0} + \sum_{j=1}^n (p_{x_j} - p_{x_{j-1}}) \right| \leq |p_{x_0}| + \sum_{j=1}^n |p_{x_j} - p_{x_{j-1}}| \\ &\leq |P_{x_0}| + \underbrace{\left(\sum_{j=1}^p c_{x_{j-1}x_j} |p_{x_{j-1}} - p_{x_j}|^2 \right)^{1/2}}_{\leq J(\hat{p})^{1/2}} \left(\sum_{j=1}^p \frac{1}{c_{x_{j-1}x_j}} \right)^{1/2}. \end{aligned}$$

Toute majoration de $J(\hat{p})$ induit donc une majoration des valeurs ponctuelles de \hat{p} , donc de n'importe quelle norme de p . La fonctionnelle J est donc coercive. Elle admet donc un minimiseur sur $\mathbb{R}^{\mathring{V}}$ (proposition 13.8, page 261). D'après la proposition 13.9, on a $\nabla J(\hat{p}) = 0$, soit

$$\sum_{y \sim x} c_{xy} (p_x - p_y) = 0 \quad \forall x \in \mathring{V},$$

qui correspond exactement au problème de Laplace discret. On a donc existence d'une solution. Pour l'unicité, en restant dans l'esprit d'une démonstration généralisable à d'autres situations, on montre la stricte convexité de la fonctionnelle. Pour p et p' deux champs différents (mais vérifiant tous deux les conditions aux limites), il existe au moins une arête sur laquelle $p_x - p_y \neq p'_x - p'_y$, le terme associé $c_{xy}(p_x - p_y)^2$ assure donc la stricte convexité de la fonctionnelle. On a donc unicité du minimiseur. Or toute solution du problème de Dirichlet est minimiseur (conditions suffisante d'optimalité pour une fonctionnelle convexe), on a donc bien unicité de la solution.

La linéarité est immédiate : pour P_0, P_1 dans \mathbb{R}^Γ , si l'on note p_0 et p_1 les solutions associées, alors pour tout $\lambda \in \mathbb{R}$, le champ $p_0 + \lambda p_1$ est solution du système avec la condition aux limites $P_0 + \lambda P_1$, donc la solution par unicité. \square

Remarque 2.6. (Formulation du problème de Dirichlet comme problème de minimisation sous contraintes) Le problème de Dirichlet a été interprété comme problème d'optimisation sur un espace affine (l'espace des champs qui vérifient la condition de Dirichlet). On dit que les conditions aux limites ont été prises en compte de façon *essentielle*, au sens où on les a intégrées à l'ensemble sur lequel on effectue la modélisation. Bien que l'approche présentée ci-dessous puisse ne pas être d'un intérêt qui saute aux yeux a priori, on peut formuler ce problème comme une minimisation sur l'espace \mathbb{R}^V entier, sous les contraintes de valeurs sur Γ imposées, et traiter ces contraintes par dualité. Plus précisément, on se place dans le cadre de la proposition 13.12, page 263. On définit

$$J : q \in \mathbb{R}^V \longmapsto J(q) = \frac{1}{2} \sum_e c_{xy} (q_x - q_y)^2$$

en imposant les valeurs au bord Γ de la façon suivante : on demande que q soit dans

$$K = P + \ker B,$$

6. Cette inégalité est une sorte de version en moyenne quadratique du théorème des accroissements finis : pour tout domaine Ω borné (au moins dans une direction), il existe une constante C telle que, pour toute fonction $u \in H_0^1(\Omega)$ (espace des fonctions dont la norme du gradient est dans L^2 , et nulles au bord de Ω)

$$\int_\Omega u^2 \leq C \int_\Omega |\nabla u|^2.$$

où B est l'application “trace”

$$B q \in \mathbb{R}^V \longmapsto q|_{\Gamma} \mathbb{R}^{\Gamma}.$$

La proposition 13.12 assure l'existence d'un vecteur de multiplicateurs de Lagrange $\lambda \in \mathbb{R}^{\Gamma}$ tel que, en le minimiseur p sur K ,

$$\nabla J(p) + B^* \lambda = 0.$$

Si l'on se place en un point x de Γ ,

$$(\nabla J(p))_x = \sum_{y \sim x} c_{xy} (p_x - p_y)$$

est le défaut de conservation en x , plus précisément le flux rentrant en ce point de la frontière. Par suite $(B^* \lambda)_x = \lambda_x$, le multiplicateur de Lagrange associé à la contrainte de pression imposée ponctuelle, est le flux *sortant* en x .

Au delà de l'interprétation des multiplicateurs de Lagrange dans ce cas particulier, cette formulation est parfois utilisée pour résoudre le problème de Dirichlet. L'un des avantages de l'approche est qu'elle permet de travailler avec la matrice du Laplacien sur l'ensemble du réseau, et permet de court-circuiter l'opération parfois délicate consistant à supprimer les lignes et les colonnes correspondant aux points de la frontière.

Remarque 2.7. Noter que, dans le problème d'optimisation intervenant dans la preuve précédente, on n'impose pas la loi des noeuds sur les flux associés à la pression p . La conservation au niveau des points intérieurs est *conséquence* du caractère minimisant de p .

Remarque 2.8. (Formulation variationnelle)

Pour faire un lien avec une méthode très féconde pour les EDP elliptiques⁷, on peut établir ce que l'on appelle une *formulation variationnelle* du problème, en considérant un champ $\in \mathbb{R}$ nul sur Γ , multipliant l'équation au point x par q_x , et en sommant sur les sommets de V . On obtient

$$\sum_x q_x \sum_{y \sim x} c_{xy} (p_x - p_y) = \sum_e c_e (p_y - p_x) (q_y - q_x) = 0,$$

qui est donc valable pour tout $q \in \mathbb{R}^V$ nul sur Γ . C'est simplement une autre manière d'écrire les conditions d'optimalité pour la fonctionnelle d'énergie, qui passe par la définition de la forme bilinéaire ci-dessus.

Conditions aux limites de Neuman / problème de Poisson

Imposer des conditions aux limites de Neuman consiste à imposer les flux (et non les valeurs de pression) au travers des points de la frontière. Comme l'équation sur les points intérieurs consiste aussi à imposer des flux, il n'y a pas lieu de distinguer les points intérieurs de ceux de la frontière⁸. On se donne simplement un champ de flux ponctuels (b_x) sur tout V (en gardant en tête que b_x doit être nul pour tout $x \in \mathring{V}$, si l'on veut garder l'idée de points intérieurs), et le problème s'écrit simplement

$$\sum_{y \sim x} c_{xy} (p_x - p_y) = b_x \quad \forall x \in V \tag{2.5}$$

On notera que, dans ce contexte discret, imposer des conditions aux limites de Neuman consiste à considérer un problème impliquant l'opérateur de laplacien discret avec un second membre non nul (on parle de problème non-homogène), ce qui confère au problème une structure de type problème de Poisson, version discrète du problème $-\Delta p = f$ dans un domaine euclidien.

7. Dans ce contexte, la formulation variationnelle du problème de Poisson $-\Delta p = f$ sur un domaine Ω avec conditions de Dirichlet homogènes s'écrit

$$\int_{\Omega} \nabla p \cdot \nabla q = \int_{\Omega} f q \quad \forall q \in H_0^1(\Omega),$$

où $H_0^1(\Omega)$ est l'espace de Sobolev des fonctions de carré intégrable, dont le gradient est lui aussi de carré intégrable, et de trace nulle.

8. Ce point distingue le problème discret du problème continu. Pour ce dernier, on considère que l'équation est valable en tout point intérieur au domaine (au sens topologique usuel), et que l'on a des conditions aux limites sur la frontière du domaine. Dans le cas discret, les notions d'intérieur et de frontière topologique perdent toute signification, il n'y a pas lieu de les distinguer.

Théorème 2.9. On suppose le réseau connexe, et les flux imposés globalement conservatifs, i.e.

$$\sum_x b_x = 0.$$

Le problème (2.5) admet une solution $p = (p_x)$ unique à une constante additive près.

Démonstration. Le problème s'écrit matriciellement $Lp = b$, où L est la matrice du laplacien sur le réseau total. Cette matrice est symétrique positive, non définie car tout vecteur constant annule L . Réciproquement, si $Lp = 0$, alors p est nécessairement constant. Le rang de L est donc $N - 1$, où $N = \sharp(V)$ est l'ordre du graphe, et le noyau est la droite des champs constants. D'après la condition sur les flux, b est dans l'orthogonal du noyau de L , il est donc dans l'image de L^T , donc dans l'image de L par symétrie. Le problème (2.5) admet donc une solution, qui est unique à un élément du noyau près, c'est-à-dire à une constante additive près. \square

Exercice 2.1. Écrire un énoncé précisant le caractère bien posé du problème de Neuman, dans le cas où le réseau n'est pas connexe.

Automorphismes

Nous précisons ici comment le fait qu'un graphe ait un groupe d'automorphismes (voir définition 10.11) non réduit à l'identité peut donner des informations sur la solution du problème de Dirichlet.

Lemme 2.10. Soit $\mathcal{N} = (V, E, r)$ un réseau résistif, et Φ un automorphisme⁹ de \mathcal{N} . Soit $p \in \mathbb{R}^V$, et L le laplacien associé au graphe (définition 2.3). On a

$$(Lp)_{\Phi(x)} = L(p \circ \Phi)_x \quad \forall x,$$

autrement dit $(Lp) \circ \Phi = L(p \circ \Phi)$.

Démonstration. Pour tout $x \in V$, on a

$$L(p \circ \Phi)_x = \sum_{y \sim x} c_{xy}(p_{\Phi(x)} - p_{\Phi(y)}) = \sum_{y \sim x} c_{\Phi(x)\Phi(y)}(p_{\Phi(x)} - p_{\Phi(y)}) = \sum_{y' \sim x'} c_{x'y'}(p_{x'} - p_{y'})$$

avec $x' = \Phi(x)$. \square

On en déduit la propriété suivante.

Proposition 2.11. Soit $\mathcal{N} = (V, E, r, \Gamma)$ un réseau connexe, avec frontière non vide, $P \in \mathbb{R}^\Gamma$ donné, et p la solution du problème de Dirichlet (2.2) associée à P . On note \mathcal{A} le groupe des automorphismes¹⁰ de \mathcal{N} , et l'on considère l'action de \mathcal{A} sur V . On suppose que $P \circ \Phi = P$, autrement dit que P est constant sur les orbites :

$$P_{\Phi(x)} = P_x \quad \forall x \in \Gamma, \forall \Phi \in \mathcal{A}.$$

Alors p est lui-même constant sur les orbites, i.e. $p \circ \Phi = p$ pour tout $\Phi \in \mathcal{A}$.

Démonstration. Pour tout $\Phi \in \mathcal{A}$, le champ $p \circ \Phi$ est lui-même solution du problème de Dirichlet (2.2), car il vérifie les conditions aux limites par hypothèse, et il est harmonique sur \mathring{V} d'après le lemme 2.10. Comme la solution au problème est unique, $p \circ \Phi$ s'identifie à p . \square

9. Conformément à la définition 10.11, on demande non seulement la correspondance des arêtes,

$$(\Phi(x), \Phi(y)) \in E' \iff (x, y) \in E,$$

mais aussi la correspondance des résistances :

$$r_{xy} = r_{\Phi(x), \Phi(y)} \quad \forall x, y \in V.$$

10. Conformément à la définition 10.11, on demande aux automorphismes de vérifier aussi $\Phi(\Gamma) = \Gamma$.

2.3 Opérateurs discrets : gradient, divergence, et laplacien

Nous introduisons dans cette section le pendant discret des opérateurs de divergence et de gradient, qui permettent de reformuler ce qui précéde, et d'établir un lien étroit la démarche suivie précédemment et les problèmes tels qu'ils se formulent dans un domaine euclidien.

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

$$\partial : u \in \mathbb{R}^E \mapsto \partial u \in \mathbb{R}^V \quad (2.6)$$

$$(\partial u)_x = - \sum_{y \sim x} u_{xy}. \quad (2.7)$$

Nous nous intéresserons dans la suite à des flux conservatifs, i.e. tels que $\partial u(x) = 0$ pour tout sommet x dans $\mathring{V} = V \setminus (\{o\} \cup \Gamma)$. On définit l'application ∂^* (équivalent discret de l'opérateur de gradient, dont le symbole est expliqué dans la proposition 2.12 ci-dessous) comme

$$\partial^* : p \in \mathbb{R}^V \mapsto \partial^* p \in \mathbb{R}^E \quad (2.8)$$

$$(\partial^* p)_{xy} = p_y - p_x. \quad (2.9)$$

Proposition 2.12. Les applications linéaires $\partial : \mathbb{R}^E \rightarrow \mathbb{R}^V$ et $\partial^* : \mathbb{R}^V \rightarrow \mathbb{R}^E$ sont mutuellement adjointes, au sens suivant :

$$\langle \partial u | p \rangle_V = \langle u | \partial^* p \rangle_E,$$

où $\langle \cdot | \cdot \rangle_V$ et $\langle \cdot | \cdot \rangle_E$ représentent les produits scalaires canoniques sur \mathbb{R}^V et \mathbb{R}^E , respectivement.

Démonstration. Pour tout $u \in \mathbb{R}^E$, $p \in \mathbb{R}^V$, on a

$$\langle \partial u | p \rangle_V = \sum_{x \in V} q_x \partial u(x) = \sum_{x \in V} q_x \sum_{y \sim x} u_{yx} = \sum_{(x,y) \in E} u_{xy} \underbrace{(p_y - p_x)}_{\partial^* q(x,y)} = \langle u | \partial^* p \rangle_E,$$

qui est la propriété annoncée. □

Remarque 2.13. (Retour sur la proposition 2.2)

On établit immédiatement un équivalent discret du théorème de la divergence

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

On considère exceptionnellement ici que V se partitionne en \mathring{V} et Γ , *sans que l'on ait nécessairement conservation sur les points de Γ* . On a, pour tout $e = (x, y) \in E$, $u_{xy} + u_{yx} = 0$, d'où, en sommant sur toutes les arêtes, et en écrivant la somme sur les sommets :

$$\sum_x (\partial u)_x = 0,$$

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

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

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 $u \cdot n$.

Remarque 2.14. On prendra garde au fait que, si l'on peut associer un champ de flux à un champ de pressions (en écrivant $u = -cd^*p$), la réciproque n'est pas vraie en général, et quand elle est vraie elle doit être justifiée. La possible non-existence d'un champ de pression associé à u vient des cycles. Considérer par exemple un réseau circulaire (discréétisation du cercle unité), avec un champ de flux constant qui fait tourner le fluide. Il est évident que ce flux ne peut résulter d'un champ de pression.

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

$$\begin{cases} u + c\partial^* p = 0 & \text{sur } E \\ \partial u = 0 & \text{sur } \mathring{V}. \end{cases} \quad (2.10)$$

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 sur Γ . On peut ré-écrire le problème de Dirichlet (2.2) avec ce nouveau formalisme :

$$\begin{cases} (\partial c\partial^* p)_x = 0 & \forall x \in \mathring{V}, \\ p_x = P_x & \forall x \in \Gamma, \end{cases} \quad (2.11)$$

ou, de façon plus concise encore,

$$\begin{cases} \partial c\partial^* p = 0 & \text{dans } \mathring{V}, \\ p = P & \text{sur } \Gamma, \end{cases} \quad (2.12)$$

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

Remarque 2.15. Le but de cette remarque est d'explorer de façon conjointe deux questions naturelles : (1) quels sont les ingrédients du modèle qui assurent que le principe du maximum (proposition 2.4) soit vérifié ?

(2) le fait que l'opérateur ∂^* (gradient discret, qui exprime la loi de Poiseuille) soit le transposé de l'opérateur ∂ (divergence discrète, qui exprime la loi des nœuds) exprime-t-il un principe universel ?

La deuxième question est en particulier justifiée par le fait que la formulation variationnelle utilisée dans la démonstration ci-dessous s'écrit

$$\sum_e c_e \partial^* p(e) \partial^* q(e) = 0$$

pour tout champ-test q nul au bord. Malgré la symétrie de cette formulation, les deux instances de l'opérateur ∂^* dans cette formulation variationnelle correspondent à des principes très différents. Le ∂^* de droite correspond à la loi de Poiseuille, le premier résulte de la sommation par parties de l'expression de la conservation locale :

$$\partial c\partial^* p = 0 \implies \sum_x \partial c\partial^* p q = 0 \implies \sum_e c\partial^* p \partial^* q = 0.$$

Comme dans le contexte de problème de Poisson sur un domaine de l'espace euclidien, l'opérateur ∂ de la loi de Krichhoff provient du fait que le bilan de matière s'exprime par une *sommation* des flux ou des quantités. Le ∂^* de la loi phénoménologique en revanche reflète une propriété particulière de linéarité du flux par rapport au saut de pression (ou de potentiel électrique), mais on peut vérifier (pour revenir à la première question) que d'autres lois pourraient être envisagées, sans perte du principe du maximum. L'essentiel est que le flux puisse s'écrire comme

$$u_{xy} = F(p_x - p_y),$$

où $F : \mathbb{R} \rightarrow \mathbb{R}$ est une fonction impaire croissante, strictement croissante dans un voisinage de 0. L'harmonicité d'un champ de pression prendrait alors la forme non linéaire suivante

$$\sum_{y \sim x} F(p_x - p_y) = 0.$$

Cette identité ne peut être réalisée que si $p_x - p_y$ est identiquement nul ou, si ça n'est pas le cas, si la somme contient des termes de signes différents. Dans tous les cas p_x est dans l'enveloppe convexe des p_y , ce qui exprime le principe du maximum.

Exercice 2.2. Ce petit exercice¹¹ est une préparation à la proposition 2.16 ci-après (il en est un cas particulier). On considère un réseau linéaire constitués de $N + 1$ sommets indexés par $0, 1, \dots, N$. On considère que chaque point j est connecté avec ses deux voisins $j - 1$ et $j + 1$, sauf les extrémités qui n'ont qu'un voisin. On cherche à minimiser (la moitié de) l'énergie dissipée

$$J(u) = \frac{1}{2} \sum_{j=1}^N r_{j-1,j} u_{j-1,j}^2$$

sous une contrainte de conservation aux point intérieurs, et de flux imposé aux extrémités :

$$u_{j-1,j} - u_{j,j+1} = 0 \quad \forall j = 1, \dots, N - 1, \quad -u_{0,1} = F_0, \quad u_{N-1,N} = F_N.$$

On note K l'ensemble des flux qui vérifient ces lois de conservation.

- a) Montrer que K est non vide si et seulement si $F_0 + F_N = 0$. On fait cette hypothèse, et l'on note $F = -F_0 = F_N$.
- b) En supposant la condition ci-dessus vérifiée, montrer que J admet un minimiseur unique sur K .
- c) Montrer que ce problème de minimisation rentre dans le cadre de la proposition 13.12, page 263. En déduire l'existence d'un champ de multiplicateurs de Lagrange $p = (p_0, \dots, p_N)$ qui vérifie une collection de relations avec les flux, et montrer que ce champ s'identifie à un champ de pression tel que défini précédemment.
- d) Le champ p est-il unique ?

La proposition suivante n'est pas nécessaire à la compréhension de la suite des développements, mais elle donne une interprétation des pressions comme un champ auxiliaire abstrait associé à un problème de minimisation de l'énergie dissipée au sein d'un réseau¹².

Proposition 2.16. On considère un réseau (V, E, r) , et l'on considère le problème de minimisation de (la moitié de) l'énergie dissipée exprimée à l'aide des flux $u \in \mathbb{R}^E$:

$$J : u \mapsto \frac{1}{2} \sum_e r_e u_e^2,$$

sous contraintes de bilan exprimées aux sommets :

$$\partial u = b,$$

où $b \in \mathbb{R}^V$ est globalement conservatif, i.e.

$$\sum_x b_x = 0$$

Ce problème admet une solution (unique si le réseau est connexe), et les conditions d'optimalité sous contraintes expriment la loi d'Ohm (ou loi de Poiseuille dans l'interprétation fluide du réseau).

Démonstration. En premier lieu remarquons que l'ensemble admissible est non vide : le théorème 2.9, page 44, applicable d'après la condition sur les flux, assure l'existence d'une solution au problème de Neuman, le champ de flux associé $u = -c\partial^*p$ est donc admissible. Le problème consiste en la

11. Il est en lui-même d'un intérêt abyssal, puisqu'il consiste à minimiser une fonctionnelle sur un singleton, mais si l'on passe outre ce fait, il permet de comprendre comment l'écriture de conditions nécessaires d'optimalité fait apparaître des pressions aux sommets comme multiplicateurs de Lagrange de la loi des noeuds.

12. On notera le caractère déroutant de cette proposition. Les lois d'Ohm et de Poiseuille sont des lois phénoménologiques faisant intervenir des variables scalaires (potentiel électrique ou pression) qui ont un sens physique clair, et qui peuvent faire l'objet de mesures expérimentales. L'approche présentée ici présente une vision différente, basée sur un principe de minimisation de l'énergie globale dissipée, sous contrainte de conservation, qui permet d'introduire ces potentiels/pressions comme des auxiliaires abstraits, multiplicateurs de Lagrange associés aux contraintes locales de conservation ou bilan au noeuds.

minimisation d'une forme quadratique définie positive, sous contraintes affines qui définissent un ensemble non vide. Il existe donc un unique minimiseur $u \in \mathbb{R}$. Ce problème rentre dans le cadre de la proposition 13.12, page 263, où $B \in \mathcal{L}(\mathbb{R}^E, \mathbb{R}^V)$ est l'expression matricielle de ∂ , c'est à dire que la ligne de B correspondant à la contrainte de conservation au sommet x exprime

$$\sum_{y \sim x} u_{yx} = b_x.$$

Chaque ligne de B contient donc des éléments qui valent 0 (si l'indice e de la colonne ne contient pas x), et ± 1 si e contient x , le signe dépendant de l'orientation (arbitraire) choisie pour désigner l'arête. Noter que chaque colonne $e = (x, y)$ de B contient exactement 2 éléments non nul : -1 pour x et $+1$ pour y . La proposition 13.12 assure donc l'existence d'un champ de multiplicateurs de Lagrange $p \in \mathbb{R}^V$ tel que

$$\nabla J + B^* p = 0 \iff ru + \partial^* p = 0$$

qui exprime exactement, pour toute arête (x, y) ,

$$r_{xy}u_{xy} - p_x + p_y = 0,$$

c'est à dire la loi d'Ohm / Poiseuille pour l'arête (x, y) . □

Remarque 2.17. Cette remarque est en quelque sorte duale de la proposition précédente, qui établissait que la loi d'Ohm/Poiseuille pouvait être interprétée comme une *conséquence* d'un principe de minimisation sur les flux. Si l'on considère maintenant le problème de minimisation de l'énergie dissipée exprimée en termes de pressions, à partir de la loi d'Ohm (on multiplie par $1/2$ par commodité) :

$$\mathcal{P} = \frac{1}{2} \sum_e c_{xy} (p_x - p_y)^2,$$

où l'on fixe des pressions sur Γ , les conditions d'optimalité s'écrivent

$$\sum_{y \sim x} c_{xy} (p_x - p_y) = 0$$

qui s'écrit $du(x) = 0$, pour tout $x \in \mathring{V}$. La conservation locale (loi des noeuds) est donc une *conséquence* de ce principe d'optimisation : il est optimal pour minimiser l'énergie dissipée d'assurer la conservation locale en tout point intérieur au réseau.

Remarque 2.18. Pour faire la synthèse entre l'approche initiale proposée, la proposition 2.16, et la remarque 2.17, on peut remarquer que la formalisation des phénomènes considérés ici repose sur trois piliers : (1) loi phénoménologique de type Ohm ou Poiseuille (2) Loi de conservation locale (loi de Kirchhoff) (3) principe de minimisation de l'énergie dissipée. Chacun de ces ingrédients peut en fait être déduit des deux autres. Nous avons privilégié l'approche (1)&(2) \Rightarrow (3), mais la proposition 2.16 assure (2)&(3) \Rightarrow (1), et la remarque 2.17 précise pourquoi (1)&(3) \Rightarrow (2).

2.4 Résistance équivalente d'un réseau

Nous définissons dans cette section la notion de résistance équivalente d'un réseau enraciné, où un sous-ensemble de sommets est défini comme la frontière Γ . Il est sous-entendu¹³ que cette résistance équivalente est définie entre o et Γ .

Definition 2.19. (Résistance équivalente d'un réseau)

Soit $\mathcal{N} = (V, E, r, o, \Gamma)$ un réseau connexe (selon la définition 2.1). On impose un champ de pression

13. On peut en tout généralité définir une résistance entre deux parties Γ_1 et Γ_2 disjointes de l'ensemble des sommets. Le principe consiste simplement à imposer une pression uniforme P_1 sur Γ_1 , une autre pression uniforme P_2 sur Γ_2 , à estimer le débit Q qui traverse le réseau (compté positivement de 2 vers 1, et à définir la résistance équivalente comme la constante R de proportionnalité entre Q et le saut de pression : $P_2 - P_1 = RQ$.

uniforme $P \equiv 1$ sur Γ . On note p la solution du problème de Dirichlet

$$\begin{cases} \partial c \partial^* p = 0 & \text{sur } \mathring{V}, \\ p_o = 0 \\ p_x = 1 & \forall x \in \Gamma, \end{cases} \quad (2.13)$$

et par $u = -c\partial^* p$ le flux associé. Le flux global Q est obtenu en sommant les flux rentrant au travers de Γ , ou de façon équivalente en considérant le flux qui sort par la racine o :

$$Q = \sum_{x \sim o} u_{xo} = (\partial u)_o. \quad (2.14)$$

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

Proposition 2.20. (Loi de Joule pour un réseau)

Soit $\mathcal{N} = (V, E, r, o, \Gamma)$ un réseau connexe, et p la solution du problème (2.13) associée à une pression uniforme P sur Γ . Le taux d'énergie dissipée dans le réseau, définie comme la somme des énergies $c_{xy}(p_x - p_y)^2$ dissipées dans les arêtes, vérifie la relation

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

où $Q = \partial u(o)$ est le flux sortant de Γ à travers o .

Démonstration. C'est une conséquence de la formule de Green discrète (sommation par parties). L'énergie dissipée s'écrit (on rappelle que dans la somme sur E on ne compte qu'une fois chaque arête)

$$\begin{aligned} \mathcal{P} &= \sum_E c_{xy}(p_x - p_y)^2 \\ &= \sum_{x \in \mathring{V}} p_x \underbrace{\sum_{y \sim x} c_{xy}(p_x - p_y)}_{=0} + \sum_{x \in \{o\} \cup \Gamma} p_x \sum_{y \sim x} \underbrace{c_{xy}(p_x - p_y)}_{=u_{xy}} \\ &= P \sum_{x \in \Gamma} \sum_{y \sim x} u_{xy} = PQ = RQ^2, \end{aligned} \quad (2.15)$$

ce qui termine la preuve □

La proposition suivante identifie la résistance équivalente d'un réseau à la puissance minimale dissipée pour les flux conservatifs sur \mathring{V} , qui assurent le passage d'un flux unitaire entre Γ et o .

Proposition 2.21. (Expression variationnelle de la résistance équivalente)

On considère un réseau enraciné (V, E, r, o, Γ) connexe. On note Λ_1 l'ensemble des champs de flux de \mathbb{R}^E conservatifs sur \mathring{V} , et tels que le flux entre Γ et o est unitaire, i.e.

$$(\partial u)_o = \sum_{x \sim o} u_{xo} = - \sum_{y \in \Gamma} (\partial u)_y = \sum_{y \in \Gamma} \sum_{x \sim y} u_{yx} = 1.$$

La résistance équivalente (entre o et Γ) s'exprime

$$R = \min_{u \in \Lambda_1} \sum_e r_e u_e^2. \quad (2.16)$$

Démonstration. En premier lieu, vérifions que Λ_1 est non vide. On considère pour cela la solution du problème de Dirichlet (2.2), pour la condition aux limites uniforme $P \equiv 1$ sur Γ . On note p la solution de ce problème, et l'on note $u = -c\partial^* p$ le champ de flux associé. Ce champ induit un certain flux $Q \neq 0$ de Γ vers o . Par linéarité du problème, le champ P/Q sur Γ induit un flux unitaire, et le champ

des flux, que l'on note \bar{u} , est conservatif sur les points intérieurs du fait de l'harmonicité du champ de pression sous-jacent sur \mathring{V} .

Le problème (2.24) consiste à minimiser une fonctionnelle J continue coercive (que l'on définit comme la moitié de la puissance dissipée) sur un ensemble fermé non vide, il admet donc un minimiseur, qui est unique par stricte convexité de la fonctionnelle et convexité de Λ_1 . L'ensemble admissible est un espace affine, ensemble des $u \in \mathbb{R}^E$ tels que

$$\sum_{y \in \Gamma} \sum_{x \sim y} u_{yx} = 1, \quad \partial u(x) = \sum_{y \sim x} u_{yx} = 0 \quad \forall x \in \mathring{V}.$$

Cet espace s'écrit $\bar{u} + \ker B$, où \bar{u} est le champ construit précédemment, et B est la matrice qui exprime les contraintes ci-dessus. Ce problème rentre dans le cadre de la proposition 13.12, page 263. On suit la démarche de la preuve de la proposition 2.16, page 47. la différence ici est que la contrainte de flux global fait intervenir l'ensemble des sommets de Γ . On aura donc un unique multiplicateur de Lagrange, c'est à dire une unique pression, pour l'ensemble des points de Γ . On a donc, comme pour la proposition 2.16,

$$\nabla J + B^* p = 0 \iff ru + \partial^* p = 0$$

c'est-à-dire, pour toute arête (x, y) ,

$$r_{xy}u_{xy} - p_x + p_y = 0,$$

avec maintenant p_x égal à une même valeur P pour tous les points de Γ . L'autre différence est qu'il n'y a pas de contrainte en la racine o , donc pas de multiplicateur de Lagrange associé à la racine, mais on peut néanmoins faire en sorte que l'identité ci-dessus soit bien vérifiée pour toutes les arêtes (y compris celles qui contiennent la racine), en posant simplement $p_0 = 0$. En éliminant maintenant les flux, on obtient que le champ p des multiplicateurs de Lagrange est harmonique sur \mathring{V} , prend la valeur 0 en o , la valeur P sur Γ , il s'identifie donc à la solution du problème de Dirichlet qui fonde la définition de la résistance équivalente (définition 2.19), multipliée par P . La loi de Joule sur le réseau (proposition 2.20) assure donc $\mathcal{P} = RQ^2 = R$, du fait que le débit est unitaire, où \mathcal{P} est la puissance dissipée associée à u . La résistance est égale à cette puissance dissipée minimale (comme u minimise $\mathcal{P}/2$, il minimise aussi \mathcal{P}). \square

Remarque 2.22. Précisons les similarités et différences entre ce cadre discret et le cadre continu. La formule de Green utilisée précédemment

$$\sum_E c_{xy}(p_x - p_y)(q_x - q(y)) = \sum_{x \in V} q_x \sum_{y \sim x} c_{xy}(p_x - p_y),$$

est analogue à la même formule dans un domaine continu sans bord (par exemple pour l'espace entier, ou un domaine périodique). De fait, la notion de frontière pour un réseau est arbitraire¹⁴, 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

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

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

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

où u_i est le flux rentrant en x_i . On a alors formellement

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

14. Contrairement au cas d'un domaine Ω (i.e. un ouvert non vide) de l'espace euclidien sur lequel on écrit un problème le Laplace pour représenter par exemple un phénomène de diffusion stationnaire sans terme source, considérant que le milieu n'échange de la matière avec le monde extérieur qu'au travers de sa frontière *topologique* $\overline{\Omega} \setminus \Omega$.

qui serait l'équivalent discret de (2.15). Le problème est que cette expression n'a pas de sens, car les points ont une capacité nulle en dimension $d \geq 2$.

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

$$\begin{aligned} \sum_E c_{xy}(p_x - p_y)(q_x - q_y) &= \sum_{x \in \vec{V}} q_x \underbrace{\sum_{y \sim x} c_{xy}(p_x - p_y)}_{=\partial c \partial^* p_x} \\ &\quad + \sum_{x \in \{o\} \cup \Gamma} q_x \sum_{y \sim x} c_{xy}(p_x - p_y) \\ &= \sum_{x \in \vec{V}} q_x \partial c \partial^* p_x - \sum_{e=(x,y) \in E^{\Gamma}} c_{xy} q_x \partial^* p(e), \end{aligned}$$

qui est maintenant l'équivalent discret de

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

2.5 Cadre stochastique

Soit un réseau $\mathcal{N} = (V, E, r)$ (voir définition 2.1), on considère la marche aléatoire sur V associée aux probabilités de transitions π_{xy} , définies par

$$\pi_{xy} = \frac{c_{xy}}{C_x}, \quad C_x = \sum_{y \sim x} c_{xy}, \quad (2.17)$$

où $c_{xy} = 1/r_{xy}$ 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 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 $\mathcal{N} = (V, E, r, \Gamma)$ connexe et la donnée d'un champ de pressions $(P_x)_{x \in \Gamma}$ sur la frontière. 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 :

$$X_0 = x, \quad X_1, \dots, \quad X_i \in \Gamma,$$

avec $X_j \notin \Gamma$ pour $0 \leq j < i$. 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.

Proposition 2.23. Le champ $p \in \mathbb{R}^V$ défini précédemment est la solution du problème (2.2).

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 $x \in \vec{V}$. On a

$$p_x = \sum_{y \sim x} \pi_{xy} p_y,$$

qui peut s'écrire (d'après (2.17))

$$C_x p_x - \sum_{y \sim x} c_{xy} p_y = 0,$$

de telle sorte que p est harmonique. Il s'agit donc nécessairement de l'unique solution du problème de Dirichlet (2.2). \square

On peut déduire de la proposition 2.23 ci-dessus une expression stochastique de la résistance entre o et Γ , pour un réseau enraciné. On considère le cas $P \equiv 1$. Le champ p solution du problème de Dirichlet (avec la valeur 0 en o) est tel que p_x est l'espérance de la valeur au bord au premier passage, qui s'écrit comme la probabilité d'atteindre Γ avant o (multipliée par la valeur 1). On parle de *probabilité d'échappement*.

Proposition 2.24. On considère une marche aléatoire sur $\mathcal{N} = (V, E, r, o, \Gamma)$ issue de o , avec des probabilités de transition données par (2.17). On a

$$\frac{1}{R} = C_o p_{esc}, \quad (2.18)$$

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. 2.19).

Démonstration. Soit p la solution du problème (2.2), 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 2.19, la résistance R est $1/\partial u(o)$. Par ailleurs on a

$$p_{esc} = \sum_{x \sim o} \pi_{ox} p_x = \frac{1}{C_o} \sum_{x \sim o} c(o, x) (p_x - p_o) = \frac{1}{C_o} \partial u(o) = \frac{1}{C_o} \frac{1}{R},$$

qui donne le résultat. \square

Remarque 2.25. La matrice de transition K associée à la marche aléatoire définie précédemment est reliée au Laplacien discret de la façon suivante :

$$K = (\pi_{xy})_{x, y \in V}, \quad \pi_{xy} = \frac{c_{xy}}{C_x} \text{ pour } (x, y) \in E,$$

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 coefficients diagonaux sont les C_x , on a la relation

$$L = \partial c \partial^* = C (\text{Id} - K). \quad (2.19)$$

On s'intéresse maintenant à l'évolution de la loi de probabilité de présence de la "particule" au fil des itérations successives.

Proposition 2.26. On considère la marche aléatoire sur un réseau connexe $\mathcal{N} = (V, E, r)$, dont les probabilités de transition sont définies par (2.17). Partant d'une loi de probabilité ρ^0 sur la position initiale, on note ρ^n la loi que suit la position de la particule à l'étape n , définie par

$$\rho_x^{n+1} = \sum_{y \sim x} \pi_{yx} \rho_y^n.$$

La relation de récurrence ci-dessus admet un unique point fixe (à une constante multiplicative près), qui correspond à la loi de probabilité C_x (après renormalisation appropriée).

Démonstration. On vérifie immédiatement que

$$\sum_{y \sim x} \pi_{yx} C_y = \sum_{y \sim x} \frac{c_{yx}}{C_y} C_y = \sum_{y \sim x} c_{yx} = C_x.$$

On note K la matrice des (π_{xy}) , de telle sorte que la relation s'écrit $\rho^{n+1} = K^T \rho^n$. Pour l'unicité du vecteur propre de K^T associé à la valeur propre 1, on montre que la dimension du sous-espace propre de K associé à la valeur propre 1 est 1. En effet, pour tout champ p tel que $p = Kp$, p_x s'écrit pour tout x comme combinaison barycentrique des valeurs aux points voisins. Si l'on considère le point x qui réalise le maximum des valeurs, alors ce maximum est également réalisé sur tous les voisins connectés, et donc par extension sur toute la composante connexe. Le réseau étant connexe, le champ p est nécessairement constant, le sous-espace propre est donc bien une droite.

N.B. : on peut aussi utiliser l'expression (2.19) reliant K à la matrice du laplacien, qui établit que tout point fixe de K est harmonique, donc constant du fait de la connexité du réseau. \square

La proposition suivante établit le caractère décroissant d'un large classe de fonctionnelles, appelées *entropie*. On se reportera à la section 1.4, page 25 pour une présentation de cette notion associée à une loi de probabilité discrète.

Proposition 2.27. Pour toute fonction φ de \mathbb{R}^+ dans \mathbb{R} convexe, la fonctionnelle d'entropie associée

$$S : \rho \mapsto \sum_{x \in V} \varphi \left(\frac{\rho_x}{C_x} \right) C_x$$

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

Démonstration. On a

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

Chaque terme de la somme s'écrit

$$\varphi \left(\frac{\rho_x^{n+1}}{C_x} \right) C_x = \varphi \left(\sum_{y \sim x} \frac{c_{xy}}{C_x} \frac{\rho_y^n}{C_y} \right) C_x \leq \sum_{y \sim x} \frac{c_{xy}}{C_x} \varphi \left(\frac{\rho_y^n}{C_y} \right) C_x$$

car φ est convexe.

On a donc finalement

$$S(\rho^{n+1}) \leq \sum_{x \in V} \sum_{y \sim x} c_{xy} \varphi \left(\frac{\rho_y^n}{C_y} \right) = \sum_y \left(\frac{\rho_y^n}{C_y} \right) \sum_{x \sim y} c_{xy} = \sum_y \left(\frac{\rho_y^n}{C_y} \right) C_y,$$

ce qui termine la preuve. \square

Corollaire 2.28. 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 :

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

Remarque 2.29. Si la fonction φ de la proposition 2.30 est strictement convexe, on peut s'attendre à une décroissance stricte de l'entropie tant que la mesure stationnaire n'est pas atteinte, de nature à assurer la convergence de la mesure vers la mesure stationnaire. On prendra garde au fait qu'il est pour cela nécessaire qu'au moins l'une des inégalités de convexité utilisée dans la démonstration soit stricte, ce qui nécessite que la combinaison convexe soit non triviale. On aura par exemple décroissance stricte dès qu'il existe un sommet qui communique avec tous les autres, i.e. $c_{xy} \neq 0$ pour tous les $y \neq x$.

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 (2.17). On note toujours ρ^n la loi d'un point évoluant suivant ces principes (en omettant la dépendance explicite en τ), on a

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

d'où

$$\frac{\rho_x^{n+1} - \rho_x^n}{\tau} = -\rho_x^n + \sum_{y \sim x} \pi_{yx} \rho_y^n,$$

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

$$\frac{d\rho_x}{dt} = -\rho_x + \sum_{y \sim x} \pi_{yx} \rho_y, \quad \text{i.e.} \quad \frac{d\rho}{dt} + (\text{Id} - {}^t K) \rho = 0.$$

On obtient une structure plus familière en considérant la variable μ exprimant la densité de ρ relativement à la mesure stationnaire C . i.e. $\mu_x = \rho_x/C_x$. En divisant l'équation précédente par C_x on obtient

$$\frac{d\mu_x}{dt} = -\mu_x + \frac{1}{C_x} \sum_{y \sim x} \frac{c_{xy}}{C_y} \rho_y = -\mu_x + \frac{1}{C_x} \sum_{y \sim x} \frac{c_{xy}}{C_y} \rho_y = -\mu_x + \sum_{y \sim x} \frac{c_{xy}}{C_x} \frac{\rho_y}{C_y} = -\mu_x + \sum_{y \sim x} \pi_{xy} \mu_y.$$

qui peut s'écrire matriciellement

$$\frac{d\mu}{dt} + (\text{Id} - K) \mu = 0. \quad (2.20)$$

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

On remarquera que K intervient pour l'équation de la chaleur associée à la variable intensive μ , et K^T pour l'équation associé à la variable extensive ρ . On se reportera à la remarque 6.15, page 129 pour des commentaires plus approfondis sur ces questions.

On peut montrer la décroissance de l'entropie relative à la mesure stationnaire pour cette équation d'évolution

Proposition 2.30. Pour toute fonction φ de \mathbb{R}^+ dans \mathbb{R} convexe C^1 , la fonctionnelle d'entropie associée

$$S : \rho \mapsto \sum_{x \in V} \varphi \left(\frac{\rho_x}{C_x} \right) C_x$$

est décroissante le long de la trajectoire discrète, i.e.

$$\frac{d}{dt} S(\rho(t)) \leq 0.$$

Démonstration. On a

$$\begin{aligned} \frac{d}{dt} S(\rho(t)) &= \sum_x \varphi' \left(\frac{\rho_x}{C_x} \right) \frac{1}{C_x} \frac{d\rho_x}{dt} C_x \\ &= \sum_x \varphi' \left(\frac{\rho_x}{C_x} \right) \left(-\rho_x + \sum_{y \sim x} \pi_{yx} \rho_y \right) \\ &= \sum_x \varphi' \left(\frac{\rho_x}{C_x} \right) \sum_{y \sim x} \left(-\frac{c_{xy}}{C_x} \rho_x + \frac{c_{yx}}{C_y} \rho_y \right) \\ &= - \sum_e c_{xy} \underbrace{\left(\varphi' \left(\frac{\rho_x}{C_x} \right) - \varphi' \left(\frac{\rho_y}{C_y} \right) \right)}_{\geq 0} \left(\frac{\rho_x}{C_x} - \frac{\rho_y}{C_y} \right) \\ &\leq 0. \end{aligned}$$

□

Remarque 2.31. Dans le cas $\varphi(\rho) = \rho \log \rho$, on retrouve l'entropie relative usuelle, qui est aussi la divergence de Kullback-Leibler entre la mesure courante et la mesure stationnaire

$$S(\rho) = \sum_{x \in V} \varphi \left(\frac{\rho_x}{C_x} \right) C_x = \sum_{x \in V} \frac{\rho_x}{C_x} \log \left(\frac{\rho_x}{C_x} \right) C_x = \sum_{x \in V} \rho_x \log \left(\frac{\rho_x}{C_x} \right) = KL(\rho_x | C_x).$$

La quantité

$$\begin{aligned} I_C(\rho) &= \sum_e c_{xy} \left(\varphi' \left(\frac{\rho_x}{C_x} \right) - \varphi' \left(\frac{\rho_y}{C_y} \right) \right) \left(\frac{\rho_x}{C_x} - \frac{\rho_y}{C_y} \right) \\ &= \sum_e c_{xy} \left(\log \left(\frac{\rho_x}{C_x} \right) - \log \left(\frac{\rho_y}{C_y} \right) \right) \left(\frac{\rho_x}{C_x} - \frac{\rho_y}{C_y} \right) \end{aligned}$$

est appelée *information de Fisher*.

2.6 Cadre abstrait et modélisation

Nous établissons dans cette section des liens entre ce qui précède et des notions élémentaires de topologie algébrique, en particuliers les notions de 0– et 1–chaînes, 0– et 1–cochaînes. Les opérateurs ∂ et ∂^* peuvent être représentés sur le diagramme suivant :

$$\begin{array}{ccccc} (\text{Extensive}) & \mathbb{R}^E & (1\text{--chaînes}) & \xrightarrow{\partial} & \mathbb{R}^V & (0\text{--chaînes}) & (\text{Extensive}) \\ (\text{Intensive}) & \mathbb{R}^E & (1\text{--cochaînes}) & \xleftarrow{\partial^*} & \mathbb{R}^V & (0\text{--cochaînes}) & (\text{Intensive}) \end{array} \quad (2.21)$$

Dans les différents contextes de modélisation évoqués au début de ce chapitre, il est pertinent de considérer les dualités verticales comme *non hilbertiennes* car, même si la dualité entre \mathbb{R}^E (1–chaînes) et \mathbb{R}^E (1 co–chaînes) est calquée sur le produit scalaire euclidien standard dans chacun de ces espaces, les objets de part et d'autre sont de natures différentes, ils n'ont en particulier pas la même unité. En premier lieu, dans un contexte de réseaux hydraulique (il en sera quasiment de même pour les réseaux hydrauliques), la dualité naturelle est à image dans les réels en unité de puissance, les Watts (W). L'espace \mathbb{R}^V en haut à droite du diagramme encode des collections de défauts de conservation ponctuelles, que l'on peut voir comme des flux ponctuels rentrant ou sortant du réseau au travers des points de bifurcation. On notera $g \in \mathbb{R}^V$ une telle collection de flux ponctuels. Une forme linéaire sur \mathbb{R}^V associe linéairement à un flux en (m^3s^{-1}) un nombre réel en watts (W). Le diagramme ci-dessus exprime que l'on a choisi d'identifier $(\mathbb{R}^V)'$ à (\mathbb{R}^V) (en Pa), selon le produit de dualité usuel

$$(p, g) \mapsto \sum_x p_x g_x.$$

On prendra garde au fait que, ici encore, cette dualité n'est pas hilbertienne à strictement parler, puisque les deux objets mis en dualité n'ont pas la même unité. De la même manière, on identifie $(\mathbb{R}^E)'$, l'espace des 1–cochaînes, dual de \mathbb{R}^E (en m^3s^{-1}), à l'espace \mathbb{R}^E (en Pa), où la valeur sur chaque arête est pensée comme une différence de pression entre les extrémités. Comme précédemment, cette dualité apparemment hilbertienne (elle le serait si l'on ne considérait pas les unités des variables impliquées) ne l'est pas. De fait, on a un produit scalaire hilbertien (ou euclidien, du fait que les réseaux sont finis) “naturel” sur \mathbb{R} , qui est donné par la puissance dissipée (loi de Joule) :

$$\langle u | u \rangle_{\mathbb{R}^E} = \sum_{e \in E} r_e u_e^2$$

qui correspond à la norme euclidienne canonique *pondérée* par les résistances.

Les dualités verticales à gauche et à droite du diagramme (2.21) mettent en correspondance des variables extensives et intensives¹⁵. À droite, la variable extensive est la 0–chaîne, mesure de bilan de flux, qui correspond à la variable intensive pression (0–cochaine). À gauche, la variable extensive est

15. Selon cette terminologie courante en physique, une variable extensive vérifie la propriété de doublement suivante : si l'on considère un système et la variable associée, la variable afférente au système composé du système de départ et d'une copie disjointe de ce système est deux fois plus grande. On parle de variable intensive si la valeur est inchangée par doublement du système. Les variables intensives correspondent à la notion mathématique de *mesure*, les variables intensives à la notion de fonction, typiquement la densité d'une mesure par rapport à une autre mesure donnée par le théorème de Radon-Nikodym.

la 1-chaîne, qui mesure les flux dans les arêtes, elle est en dualité avec la variable intensive différentiel de pression (1-cochaîne).

Remarque 2.32. Dans le cas où l'on considère un champ de pression solution du problème de Dirichlet (2.2) avec donnée au bord $P \in \mathbb{R}^\Gamma$, avec $u = -c\partial^* p$ le champ de flux associé, la relation qui définit ∂^* comme adjoint de ∂ prend un sens particulier. En effet

$$\langle u, \partial^* p \rangle = -\langle c\partial^* p, \partial^* p \rangle = -\sum_{e \in E} c_{xy} |p_y - p_x|^2$$

qui est l'opposé de la puissance dissipée au sein du réseau, alors que

$$\langle p, \partial u \rangle = \sum_{x \in V} p_x \sum_{y \sim x} u_{yx} = -\sum_{x \in \gamma} P_x \sum_{y \sim x} u_{xy}$$

est l'opposé de la somme sur la frontière du produit (pression \times flux rentrant), i.e. la puissance des forces extérieures. Cette identité exprime ainsi dans ce cadre particulier un bilan d'énergie : l'énergie injectée dans le système est instantanément dissipée sous forme de chaleur au sein du réseau.

On peut faire le lien entre ces deux dualités en considérant par exemple une forme linéaire φ sur \mathbb{R}^E définie comme un élément de l'image de ∂^* :

$$\langle \varphi, v \rangle = \sum_e v_e \partial^* p = \sum_{e=(x,y)} v_{xy} (p_y - p_x).$$

Le lien avec la dualité euclidienne est donné par le théorème de représentation de Riesz-Fréchet, qui assure l'existence d'un champ (1-chaine) $-u$ (nous l'écrivons $-u$ pour une raison qui s'éclaircira ci-dessous) tel que

$$\sum_{e=(x,y)} v_{xy} (p_y - p_x) = \sum_{e=(x,y)} r_{xy} v_{xy} (-u_{xy}),$$

de telle sorte que

$$u_{xy} = r_{xy} (p_x - p_y).$$

La loi d'Ohm apparaît ainsi comme le fruit d'une identification d'un élément de $(\mathbb{R}^E)'$ (1-cochaîne) avec un champ de flux (1-chaine) au travers du théorème d'identification de Riesz-Fréchet.

Dans le contexte des réseaux électriques, l'interprétation du diagramme ci-dessus est exactement la même, en remplaçant les flux par des intensités, en Ampères (A), les pressions par des potentiels électriques en Volts (V), l'unité de \mathbb{R} comme image des formes linéaires étant comme précédemment le W.

Dans le contexte de corps échangeant de la chaleur, l'interprétation est un peu différente, du fait que les 1-chaînes sont d'emblée des puissances (flux de chaleur traversant une arête). Les éléments du \mathbb{R}^V en haut à droite sont aussi des puissances en W, correspondant à des bilans de puissance ponctuels. Les 0-cochaînes correspondent à la variable intensive naturelle dans ce contexte : la température. La dualité entre 0-chaînes et 0-cochaînes, qui confère à ce modèle la même structure mathématique que les précédents, n'a pas un sens physique clair (l'unité de \mathbb{R} comme image des formes linéaires est non plus le W mais le WK). On a en revanche une relation naturelle entre les températures et les bilans de puissance, en introduisant explicitement la variable de temps, et une quantité statique extensive qui est la capacité thermique des sommets, que l'on notera C_x , en JK^{-1} . Le produit $C_x T_x$ correspond à l'énergie emmagasinée au point x , dont les variations correspondent à l'intégrale du bilan au même point.

Exercice 2.3. Dans le contexte des réseaux hydrauliques, proposer un dispositif qui permette de faire un lien direct entre les 0-cochaînes (pressions) et les 0-cochaînes (défaut de conservation), au travers d'une sorte de capacité volumique qui permet de transformer l'intégrale d'un bilan de flux en une quantité homogène à un volume, et proportionnelle à la pression. Proposer un tel dispositif dans le cas d'un réseau électrique.

Vocabulaire de l'homologie

On peut formaliser l'ensemble des considérations de ce chapitre en considérant des combinaisons linéaires formelles d'arêtes et de sommets. Plus précisément, on fait un choix d'orientation sur les arêtes, et l'on définit l'espace C_1 des 1-chaînes comme l'ensemble des combinaisons linéaires formelles finies à coefficients réels d'arêtes de \mathcal{N} . De la même manière l'espace C_0 des 0-chaînes est l'ensemble des combinaisons linéaires formelles finies à coefficients réels de sommets de \mathcal{N} . Toute arête (resp. tout sommet) peut être ainsi vu · e comme un élément de C_1 (resp. C_0), ce qui permet de définir l'opérateur de bord comme l'opérateur linéaire sur C_1 qui vérifie

$$e = (x, y) \in E \subset C_1 \longmapsto \partial e = y - x \in C_0.$$

Un champ de flux peut ainsi s'écrire

$$u = \sum_e u_e e,$$

et ∂u s'écrit

$$\partial u = \sum_{e=(x,y) \in E} u_{xy} \partial e = \sum_{e=(x,y) \in E} u_{xy} (y - x) = \sum_{x \in X} x \sum_{y \sim x} (-u_{xy}) = \sum_{x \in X} \left(\sum_{y \sim x} u_{yx} \right) x,$$

où le coefficient de x dans la décomposition ci-dessus correspond à $(\partial u)_x$ tel que défini par (2.6), page 45, ce qui explique qu'il soit raisonnable de garder la même notation, même si les deux notions de ∂ agissent sur des objets de natures différentes.

2.7 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 abstraite¹⁶ comme

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

On dira que la distance de $(1 - \theta)x + \theta y$ à x (resp. y) est θr (resp. $(1 - \theta)r$). Ce choix définit de façon immédiate une métrique sur la réunion des segments. On notera $\bar{\mathcal{N}}$ 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 \bar{p} continu sur $\bar{\mathcal{N}}$ affine par morceaux (sur chaque arête), et un champ de flux \bar{u} constant par morceaux. Si $u = -c\partial^* p$ (sur \mathcal{N}), on a immédiatement, sur chaque arête

$$\bar{u}(s) \equiv u_e = -\frac{1}{r_e} (p_y - p_x) = -\partial_s \bar{p}.$$

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

$$\sum_e r_e u_e^2 = \sum_e r_e \int_e u_e^2 ds = \sum_e \int_e c_e |\partial_s \bar{p}|^2 ds = \int_{\bar{\mathcal{N}}} c_e |\partial_s \bar{p}|^2 ds.$$

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*.

16. 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 représentation plate et rectiligne du réseau, mais en restant différentes pour \mathcal{N} (la distance entre leurs milieux serait par exemple r).

2.8 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. 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 Kirchhoff sur les sommets intérieurs.

On note \vec{e} la mesure vectorielle associée à l'arête e . Plus précisément, pour tout

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

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

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

Proposition 2.33. La mesure vectorielle G définie par

$$G = \sum_{e \in E} u_e \vec{e} \tag{2.22}$$

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

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

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

$$\begin{aligned} \langle \nabla \cdot G, \varphi \rangle &= -\langle G, \nabla \varphi \rangle = - \sum_{e \in E} u_e \langle \vec{e}, \nabla \varphi \rangle = - \sum_{e \in E} u_e \int_x^y n_e \cdot \nabla \varphi \\ &= - \sum_{e \in E} u_e \int_x^y \partial \varphi / \partial s \, ds = - \sum_{e \in E} u_e (\varphi(y) - \varphi(x)) = \sum_{x \in V} \varphi(x) \sum_{y \sim x} u_{xy} \\ &= - \sum_{x \in V} (\partial u)_x \varphi(x) = - \sum_{x \in \Gamma} (\partial u)_x \langle \delta_x, \varphi \rangle, \end{aligned}$$

d'où la propriété annoncée. \square

Remarque 2.34. 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)

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

on peut alors écrire

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

2.9 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 ouvertes¹⁷. 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 segments¹⁸ reliant ces points), tels que $\text{supp}(\mu_0) \cup \text{supp}(\mu_1) = \Gamma$. 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 (2.22) (voir section 2.8). On dira que u est admissible, ce qu'on écrira $u \in \Pi_{\mu_0, \mu_1}$, si

$$\nabla \cdot G_u = \mu_0 - \mu_1, \quad (2.23)$$

au sens de la proposition 2.33.

Remarque 2.35. 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 défini dans le cadre du transport optimal (voir chapitre 14). 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. 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 (2.23) 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

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

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 fils¹⁹ en nombre croissant (en parallèle), on fait diminuer la résistance, et donc la puissance dissipée, l'infimum est ainsi nul, et n'est pas atteint²⁰.

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). Pour le cas $\alpha \in]0, 1[$, le plus riche, on se reportera à l'ouvrage “Optimal Transportation Networks”²¹.

17. 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.

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

19. 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.

20. 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.

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

2.10 Réseaux infinis

Nous donnons ici quelques éléments sur l'étude de réseaux infinis, en prolongement direct de ce qui a été vu précédemment. 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

$$H^1 = \left\{ q \in \mathbb{R}^V, q_o = 0, \sum_e c_{xy} |q_y - q_x|^2 < +\infty \right\},$$

qui est le pendant discret de l'espace de Sobolev associé à un domaine de l'espace euclidien.

Proposition 2.36. L'espace H^1 défini ci-dessus est un espace de Hilbert pour le produit scalaire

$$(p, q) \mapsto \sum_e c_{xy} (p_y - p_x)(q_y - q_x).$$

Démonstration. Remarquons en premier lieu que la condition définissant H^1 assure la convergence absolue de la série définissant le produit de dualité. Ce produit est symétrique, et

$$\sum_e c_{xy} |q_y - q_x|^2 = 0$$

implique $q \equiv 0$, du fait que $q_o = 0$, par connexité du réseau. Considérons maintenant l'application

$$T : q \in H^1 \mapsto u = Tq \in \ell^2(E), \quad u_e = \sqrt{c_{xy}}(q_y - q_x).$$

La condition définissant H^1 assure bien l'appartenance de $u = Tq$ à $\ell^2(E)$. Considérons maintenant une suite p^n de Cauchy dans H^1 . L'argument usuel (inégalité de Poincaré discrète appliquée à la quantité de Cauchy) assure que (p_x^n) est de Cauchy, donc convergente vers un certain p_x^∞ , pour tout $x \in V$. Par ailleurs la suite Tp^n est de Cauchy dans ℓ^2 , elle converge donc vers $u^\infty \in \ell^2(E)$. Comme cette convergence implique la convergence terme à terme, la limite u_{xy}^∞ de $\sqrt{c_{xy}}(p_y^n - p_x^n)$ s'écrit $\sqrt{c_{xy}}(p_y^\infty - p_x^\infty)$. On a donc bien convergence de p^n vers p^∞ pour la norme définie sur H^1 . \square

On définit $H_0^1 = \overline{D}$ comme l'adhérence de l'espace D des champs de \mathbb{R}^V à support fini.

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 , résistance du sous-réseau des points à distance²² 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 fondamental²³ 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 2.37. 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 (2.17) est récurrente.

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

23. Pour la démonstration, voir par exemple :

P. M. Soardi, Potential Theory on Infinite Networks, Springer-Verlag Berlin and Heidelberg 1994.

2.11 Remarques diverses

2.11.1 Réseau résistifs comme espace métrique mesuré

Il est apparu dans ce chapitre qu'un réseau résistif connexe était canoniquement structuré comme espace métrique mesuré. Comme indiqué dans la section 6.2, la mesure "naturelle" est donnée par la collection des

$$C_x = \sum_{y \sim x} c_{xy}$$

où l'on a préalablement normalisé les conductances de façon à ce que la somme les C_x soit 1, ce qui permet d'avoir une mesure de probabilité qui charge tous les points. On notera que les points "lourds" sont ceux qui appartiennent à des arêtes de forte conductance, donc de faible résistance.

Nous avons par ailleurs (section 2.7, page 57) vu que la métrique canoniquement associée au réseau est basée sur le fait que les longueurs des arêtes sont les *résistances*, la métrique induite sur V étant celle des plus courts chemins (voir définition 10.21, page 227), la longueur d'un chemin étant la somme des longueurs des arêtes.

On notera que cette double structuration peut sembler contre intuitive si l'on pense aux domaines de \mathbb{R}^d , munis de la métrique euclidienne et de la mesure de Lebesgue. La mesure de Lebesgue est construite sur le fait que le volume d'un parallélégramme est défini comme le produit des d longueurs qui définissent sa forme, des grandes longueurs correspondant donc à des grandes mesures. Dans le cas des réseaux résistifs, la situation est différentes, puisque de grandes longueurs (= résistances) correspondent à des petites conductances, donc à des petits *volumes*.

2.12 Exercices

Exercice 2.4. 1) Préciser en quoi l'hypothèse de connexité du théorème 2.5 peut être affaiblie. De façon générale (sans aucune hypothèse sur la connexité du réseau), que peut-on dire de l'ensemble des solutions du problème de Dirichlet (2.2) ?

2) Quelles hypothèses doit-on faire pour que cela ait du sens de définir la résistance équivalente entre o et Γ ?

Exercice 2.5. (Résistances en série et en parallèle)

1) On considère un réseau résistif linéaire représenté par un graphe $G = (V, E)$, avec $V = \{0, \dots, N\}$, de connectivité

$$0 \longleftrightarrow 1 \longleftrightarrow \dots \longleftrightarrow N-1 \longleftrightarrow N,$$

et de résistance $r_{0,1}, r_{1,2}, \dots, r_{N-1,N} > 0$.

On exerce une pression nulle en 0, et une pression égale à P sur $\Gamma = \{N\}$. Préciser la solution du problème de Dirichlet associé, que la résistance équivalente du réseau, et la puissance dissipée.

2) On considère maintenant un réseau à 2 points (donc aucun point intérieur), reliés par N arêtes²⁴ de résistances r_1, \dots, r_N . Préciser les flux qui se distribuent dans les différentes arêtes, et décrire la résistance équivalente en fonction des résistances individuelles.

3) Décrire le comportement de la résistance équivalente et de la puissance dissipée (pour une valeur de P fixée), dans les cas série (1) et parallèle (2), lorsque l'une des résistances tend vers l'infini, et de même lorsque l'une des résistances tend vers 0.

Exercice 2.6. On considère un réseau résistif linéaire comme au début de l'exercice précédent. Écrire précisément la relation de dualité entre les opérateurs ∂ et ∂^* dans ce cas particulier (voir proposition 2.12), et préciser en quoi on peut l'interpréter comme une formule d'intégration par parties discrète.

Exercice 2.7. (Condensation)

On souhaite ici définir une procédure de condensation qui, à partir d'un réseau résistif (V, E, r) et un ensemble de sommets de V (non nécessairement connectés), construit un nouveau réseau pour lequel les sommets ont été fusionnés en un seul.

1) Préciser les règles de construction du nouveau réseau en termes de connectique (définition du nouvel ensemble d'arêtes) et de résistances compatible avec les lois de Poiseuille et de Krichhoff, lorsque l'ensemble des sommets fusionnés ne contient que 2 points.

2) Décrire la procédure dans le cas général.

Exercice 2.8. (Cycles)

On considère un réseau (graphe non orienté) connexe $\mathcal{N} = (V, E)$, sans boucle ((x, x) n'est élément de E pour aucun $x \in V$), avec $N = \sharp(V) \geq 2$. On rappelle que ∂ est l'opérateur

$$\partial : u \in \mathbb{R}^E \longmapsto \partial u \in \mathbb{R}^V, (\partial u)_x = \sum_{y \sim x} u_{yx}.$$

Si $W \subset V$ est un sous-ensemble de sommets, on note

$$C_W = \{u, \partial u(x) = 0 \quad \forall x \in W\}.$$

1) Montrer que l'application $W \longmapsto C_W$ est décroissante pour l'inclusion, c'est à dire que

$$W \subset W' \implies C_{W'} \subset C_W.$$

24. Cette situation a été a priori exclue, mais on peut la faire rentrer dans le cadre de la définition 2.1 en considérant que chaque arête contient en fait un sommet en son milieu.

- 2) On considère ici la dépendance, pour $W \subset V$ donné, de C_W vis à vis de l'ensemble des arêtes E . Montrer que C_W est croissant (toujours pour l'inclusion) par rapport à E .
- 3) Préciser la dimension de C_\emptyset et montrer que, si le réseau \mathcal{N} est acyclique, $C_V = \ker \partial$ est réduit au vecteur nul.
- 4) Donner des exemples de réseaux connexes pour lesquels C_V n'est pas réduit au vecteur nul.
- 5) On suppose le graphe complet (i.e. $(x, y) \in E$ pour tous $x \neq y$). Donner la dimension de C_V en fonction de N .
- 6) On considère un réseau dont la topologie correspond à un arbre dyadique à N générations, enraciné sur o , avec Γ l'ensemble des 2^N feuilles. Quelle est la dimension de $C_{V \setminus (\{o\} \cup \Gamma)}$?
- 7) Plus généralement, si l'on considère un arbre (graphe connexe acyclique), et l'on définit la frontière Γ comme l'ensemble des sommets qui sont de degré 1. Quelle est la dimension de $C_{V \setminus \Gamma}$?
- 8) (description $C_V = \ker \partial$ dans le cas général)
- a) Soit G un arbre. Montrer que pour tous sommets x et y distincts, il existe un unique chemin simple qui les relie.
- b) Soit G un graphe non orienté connexe, et A un arbre couvrant pour G (i.e. un arbre qui contient tous les sommets de G). On appelle corde une arête de G qui n'est pas dans A . Montrer que l'on peut associer à chaque corde un cycle sur G , au sens algébrique, i.e. un élément non trivial du noyau de $\ker \partial$ (on pourra considérer celui dont les flux sont unitaires, avec un sens de rotation arbitraire).
- c) Montrer que les cycles ainsi construits forment une base de $C_V = \ker \partial$.

Exercice 2.9. 1) On considère un arbre résistif (réseau résistif connexe sans cycle), et l'on se donne un champ de flux $u \in \mathbb{R}^E$. Montrer qu'il existe un champ de pression $p \in \mathbb{R}^V$ tel que $u = -c\partial^* p$, i.e. $u(x, y) = c(x, y)(p(x) - p(y))$ pour tout $(x, y) \in E$. Ce champ est-il unique ?

- 2) Montrer qu'un tel champ de pression n'existe pas en général (pour un réseau quelconque).
- 3) Dans le cas d'un réseau général (avec possible non existence d'un champ de pression dont le flux est un gradient) on peut chercher les champs de pressions qui s'en approchent le plus. Faire l'analyse du problème d'optimisation correspondant, qui consiste à minimiser

$$J(p) = \frac{1}{2} |u + c\partial^* p|_r^2,$$

où $|\cdot|_r^2$ est la puissance dissipée.

Exercice 2.10. (Expression variationnelle de la résistance équivalente)

On considère un réseau enraciné (V, E, r, o, Γ) connexe. On note Λ_1 l'ensemble des champs de flux de \mathbb{R}^E conservatifs sur \dot{V} , et tels que le flux entre Γ et o est unitaire, i.e.

$$(\partial u)_o = \sum_{x \sim o} u_{xo} = - \sum_{y \in \Gamma} (\partial u)_y = \sum_{y \in \Gamma} \sum_{x \sim y} u_{yx} = 1.$$

Montrer que résistance équivalente (entre o et Γ) s'exprime

$$R = \min_{u \in \Lambda_1} \sum_e r_e u_e^2. \quad (2.24)$$

Exercice 2.11. (Loi de Poiseuille / Ohm non linéaire)

On considère un graphe non orienté connexe enraciné avec frontière $G = (V, E, o, \Gamma)$, et l'on suppose que la loi de Poiseuille est remplacée par une loi non linéaire

$$u(x, y) = \varphi(p(x) - p(y)),$$

où φ est une fonction continûment différentiable, impaire, de dérivée strictement positive.

- 1) Écrire le problème de type Dirichlet non linéaire associé.
- 2) Énoncer et démontrer le principe du maximum dans ce cas.
- 3) a) Écrire le problème de Dirichlet correspondant à la situation où l'on garde la racine à la pression nulle, et où une pression $P > 0$ est imposée en tous les points de Γ .
- b) Soit $x \in \mathring{V}$. On se donne une collection de valeurs $(p_y)_{y \sim x}$. Montrer qu'il existe un unique réel α tel que

$$\sum_{y \sim x} \varphi(\alpha - p_y) = 0.$$

Montrer que q est dans l'enveloppe convexe des p_y et que l'application qui à (p_y) associe α est continue.

- c) On introduit $m = \min(0, \min P)$ et $M = \max(0, \max P)$, et l'ensemble

$$\Lambda = \left\{ q \in \mathbb{R}^V, q(0) = 0, q(x) = P(x) \quad \forall x \in \Gamma, q(x) \in [m, M] \quad \forall x \in \mathring{V} \right\}$$

Pour tout q dans Λ , on définit $p = F(q) \in \mathbb{R}^V$ comme suit : p vérifie les conditions aux limites et, pour tout $x \in \mathring{V}$, p_x est l'unique réel qui vérifie

$$\sum_{y \sim x} \varphi(p_x - q_y) = 0.$$

Montrer que p est dans Λ .

- d) Déduire de ce qui précède et du théorème de Brouwer (théorème 19.8, page 375) que le problème de Dirichlet non linéaire admet une solution.

Chapitre 3

Le poumon humain vu comme arbre résistif

Sommaire

3.1	Vue d'ensemble de l'appareil respiratoire humain	65
3.2	Modèle tuyau-ballon	66
3.3	Le poumon comme arbre résistif	69
3.4	Numérotation dyadique	70
3.5	Arbre résistif non symétrique	72
3.6	Arbre optimal ?	76
3.7	Vers un poumon infini	77
3.8	Particules et dépôt	78
3.8.1	Sédimentation.	78
3.8.2	Particule inertielle vs. traceur passif	79

3.1 Vue d'ensemble de l'appareil respiratoire humain

La fonction principale des poumons est d'assurer les échanges gazeux entre l'air extérieur et le sang : passage de dioxygène de l'air extérieur vers le sang, et évacuation du dioxyde de carbone dans l'autre sens.

Ces échanges se font par diffusion passive au travers d'une fine membrane (dite alvéolo-capillaire) qui constitue la frontière d'une collection d'un très grand nombre d'alvéoles (de l'ordre de 300 millions). Chacune de ces petites boules (diamètre de l'ordre de 0.2 mm) dispose d'une ouverture vers l'arbre bronchique qui assure le renouvellement régulier de l'air (arrivée d'air chargé en O_2 , et évacuation de l'air chargé de CO_2). Cet arbre bronchique présente une structure dyadique : la trachée, directement connectée en amont aux voies aériennes supérieures (gorge, cavité nasale, et bouche) se sépare en aval en deux sous branches, qui elles même se séparent en deux, etc ... Le nombre de bifurcations successives est de l'ordre de 23 pour un adulte. Les 16 premières¹ générations sont purement conductrices. Au delà, les surfaces des bronches sont recouvertes d'alvéoles sus-mentionnées qui assurent les échanges gazeux. L'unité respiratoire élémentaire, qui topologiquement correspond à un sous arbre enraciné en un point de la 16-ième génération, est appelé *acinus*. Les acinus (ou *acini*) sont donc au nombre de 2^{16} , dans l'hypothèse d'une naissance à la 16-ème génération. L'ensemble arbre + alvéoles est contenu dans l'espace délimité par la cage thoracique. Le muscle appelé diaphragme, situé entre les poumons et les muscles abdominaux, induit en se contractant un mouvement des deux parties

1. Ce nombre de 16 peut varier légèrement d'une personne à l'autre, nous le fixerons à 16 pour simplifier la présentation.

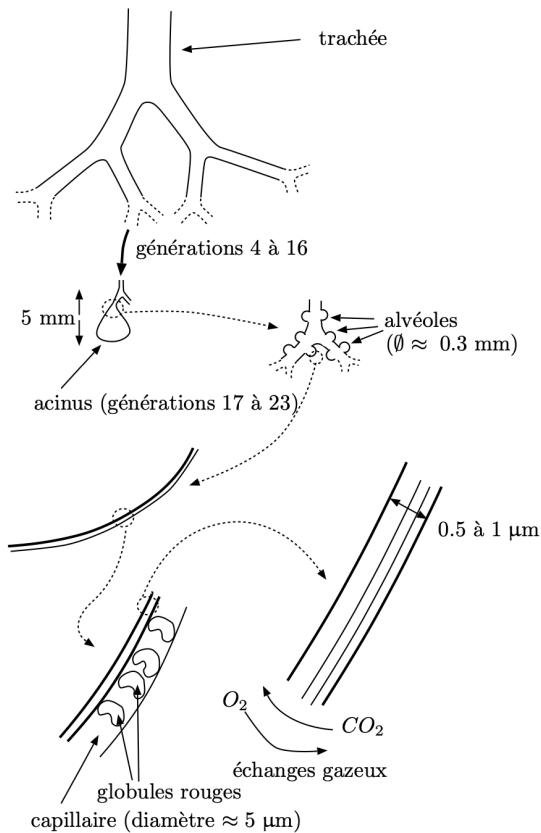


FIGURE 3.1 – Vue d'ensemble

symétriques de la cage thoracique, qui entraîne une augmentation du volume de la cavité thoracique. La matière organique contenue dans cette cavité (*paremchyme*) étant incompressible, cette augmentation de volume est essentiellement portée par les alvéoles, qui en augmentant de taille créent un appel d'air au travers de l'arbre bronchique (inspiration). Le retour vers la position d'équilibre (expiration) se fait spontanément, grâce au caractère élastique du système dans son ensemble.

3.2 Modèle tuyau-ballon

Le modèle le plus simple du système ventilatoire est fait de deux ingrédients :

1. Un *ballon* qui représente le poumon dans son ensemble, et dont l'intérieur représentent l'ensemble des alvéoles du poumon réel. Ce ballon permet d'encoder le caractère élastique du système : on supposera ses variations de volume proportionnelles au saut de pression entre l'intérieur et l'extérieur. Si l'on note P_a (a pour alvéole) cette pression interne, P la pression (supposée uniforme à l'extérieur du ballon) externe, V le volume courant du ballon, et V_0 le volume à l'équilibre, on écrira

$$P_a - P = E(V - V_0),$$

où le paramètre $E > 0$ est appelé *élastance*² du ballon.

2. Cette élastance joue le rôle de la constante de raideur pour un ressort linéaire. Noter que les unités s'en distingue : la raideur quantifie la proportionnalité entre un déplacement et une force, en Nm^{-1} , alors que l'élastance s'exprime en unité de pression par unité de volume. Dans le cas du poumon, pour des raisons historiques et de commodité, on mesure les pressions en centimètres d'eau (à la pression atmosphérique), et les volumes en litres, de telle sorte que l'élastance s'exprime en $\text{cmH}_2\text{O L}^{-1}$.

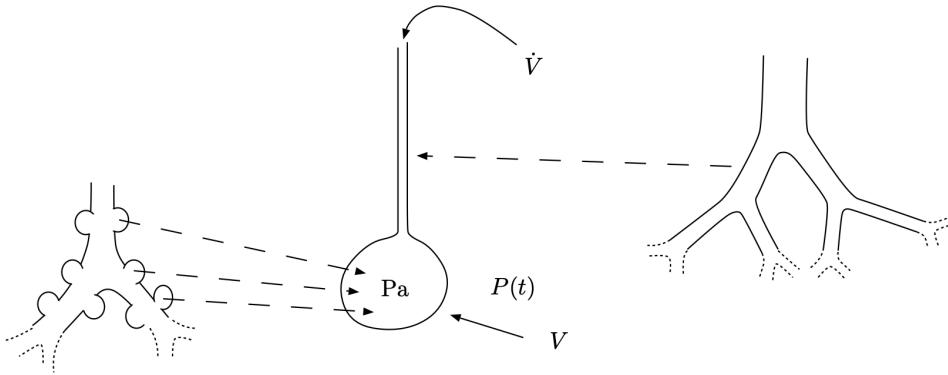


FIGURE 3.2 – Modèle tuyau-ballon

2. Un *tuyau* qui relie l'intérieur du ballon au monde extérieur. Cet élément va permettre d'encoder la résistance du système à l'écoulement. Le modèle le plus simple consiste à considérer que le flux d'air au travers du tuyau est proportionnel à la différence de pression à ses extrémités : valeur fixée à 0 pour le monde extérieur, et P_a pour l'intérieur du ballon. Le système étant supposé étanche, le débit d'air est égal à la dérivée du volume du ballon. Cette relation prend la forme d'un loi de type Ohm (ou Poiseuille) :

$$0 - P_a = \frac{dV}{dt},$$

où R est la résistance du tuyau. Elle s'exprime dans le contexte pneumologique en $\text{cmH}_2\text{O s L}^{-1}$.

En éliminant P_a on obtient l'équation

$$R \frac{dV}{dt} + E(V - V_0) = -P(t), \quad (3.1)$$

$$\frac{dV}{dt} + \frac{E}{R}(V - V_0) = -\frac{1}{R}P(t) \quad (3.2)$$

où la pression extérieure $P(t)$ peut ici être vue comme un contrôle du système à un degré de liberté (le volume V).

Remarque 3.1. (Modélisation et géométrie)

Les deux ingrédients décrits ci-dessus doivent être pensés comme des unités fonctionnelles encodant un certain type de phénomène, plus que comme des briques simplifiant des zones géométriques bien déterminées. Le ballon par exemple exprime le caractère élastique du dispositif dans son ensemble, et ce caractère élastique, quantifié par l'unique paramètre d'élastance E , synthétise de multiples ingrédients : présence de fibres élastique (élastine) au sein du parenchyme, forces de rappel élastiques pour les deux composantes de la cage thoracique, forces de tension surfacique au niveau des alvéoles, caractère élastiques de certaines bronches. Le tuyau, de son côté, semble représenter de façon simplifiée l'arbre bronchique, la résistance quantifiant alors la dissipation visqueuse au sein du fluide transporté dans des conduits étroits. De fait, l'essentiel de la dissipation visqueuse (quantifiée par le premier terme dans le bilan ci-dessous) a lieu au sein du fluide, mais pas seulement. La déformation de la partie structurelle du poumon s'accompagne elle aussi de dissipation, et il est en général considéré que 20 % de la résistance correspond à un dissipation au sein du parenchyme. Le tuyau, en tant qu'entité fonctionnelle, encode donc des phénomènes qui se passent à l'extérieur de la zone qu'il semble représenter au premier abord.

Bilan d'énergie. On obtient un bilan d'énergie en multipliant par la dérivée du volume :

$$R\dot{V}^2 + \frac{d}{dt}E \frac{(V - V_0)^2}{2} = -P\dot{V}. \quad (3.3)$$

Solution exacte. Si l'on suppose pour simplifier que la condition initiale est prise égale au volume au repos, la solution exacte s'écrit

$$V(t) = V_0 - \frac{1}{R} \int_0^t P(s) e^{-\lambda(t-s)} ds, \quad (3.4)$$

avec $\lambda = E/R$.

Forçage périodique. Lorsque le forçage est périodique, i.e. $P(\cdot)$ est une fonction T -périodique, la solution converge vers la solution périodique définie sur $[0, T]$ par

$$V_\infty(t) = V_0 + e^{-\lambda t} W - \frac{1}{R} \int_0^t P(s) e^{-\lambda(t-s)} ds, \quad (3.5)$$

with

$$W = -\frac{1}{R} \frac{1}{1 - e^{-\lambda T}} \int_0^T P(s) e^{-\lambda(T-s)} ds, \quad \lambda = \frac{E}{R}. \quad (3.6)$$

Dans le cas d'un forçage périodique constant par morceaux :

$$P(t) = \begin{cases} P_{insp} < 0 & \text{in } [0, T_{insp}[\\ P_{exp} \geq 0 & \text{in } [T_{insp}, T[, \end{cases} \quad (3.7)$$

on peut calculer le Volume courant (*Tidal Volume en anglais*)

$$\begin{aligned} V_T &= \frac{1}{E} \frac{(1 - e^{-\lambda T_{insp}})(1 - e^{-\lambda(T - T_{insp})})}{1 - e^{-\lambda T}} (P_{exp} - P_{insp}) \\ &= \frac{1}{E} \Lambda(T, T_{insp}, \lambda) (P_{exp} - P_{insp}) \end{aligned} \quad (3.8)$$

où

$$\Lambda(T, T_{insp}, \lambda) = \frac{(1 - e^{-\lambda T_{insp}})(1 - e^{-\lambda(T - T_{insp})})}{1 - e^{-\lambda T}} \quad (3.9)$$

est une fonction sans dimension des paramètres T , T_{insp} , et $\lambda = E/R$. dans la situation standard, T , T_{insp} , et $T - T_{insp}$ sont significativement plus grands que $\tau = 1/\lambda \approx 0.4$ s, de telle sorte que $\Lambda(T, T_{insp}, \lambda) \approx 1$. En ventilation normale, avec $P_{exp} = 0$ (expiration passive), on retrouve $V_T \approx -P_{insp}/E$, qui correspond à l'équilibre statique associé à $-P_{insp}$.

Aspects énergétiques La puissance des forces extérieures s'écrit

$$\mathcal{P}(t) = -P(t) \dot{V}(t).$$

Dans la situation standard (3.7) cette puissance peut être estimée sur les intervalles $(0, T_{insp})$ et (T_{insp}, T) , en notant que le travail infinitésimal associé à une variation de volume dV s'écrit

$$dW = -P dV,$$

de telle sorte que le travail total pendant l'inspiration s'écrit

$$W_{insp} = -P_{insp} V_T.$$

De la même manière, durant l'expiration (le volume subit une variation de $-V_T$), on a $W_{exp} = +P_{exp} V_T$. Finalement, l'énergie fournie au système est

$$W = \int_0^T \mathcal{P}(s) ds = W_{insp} + W_{exp} = \frac{1}{E} \Lambda(T, T_{insp}, \lambda) (P_{insp} - P_{exp})^2, \quad (3.10)$$

où $\Lambda(T, T_{insp}, \lambda)$ est la fonction donnée par (3.9)).

Dans la situation standard, $\Lambda(T, T_{insp}, \lambda)$ est proche de 1, de telle sorte que, si l'expiration est passive, $W \approx P_{insp}^2/E$, qui est le double de l'énergie potentielle de l'équilibre statique associé à la pression $-P_{insp}$. Le scénario est donc le suivant : on injecte une certaine quantité d'énergie pendant l'inspiration, une moitié de cette énergie est stockée sous forme potentielle, l'autre instantanément dissipée par effet visqueux. L'énergie stockée sous forme potentielle permet d'assurer l'expiration passive, pendant laquelle elle est dissipée par effet visqueux.

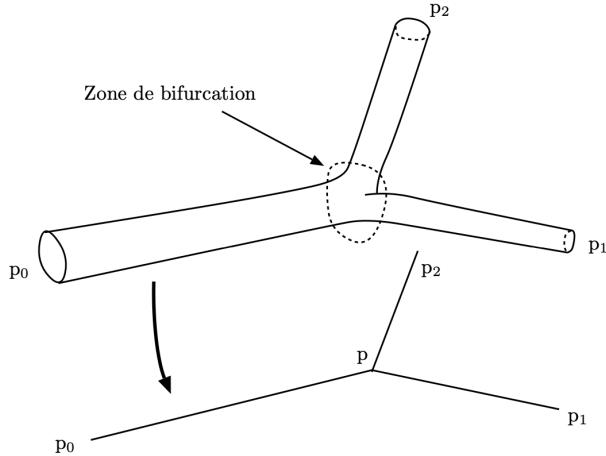


FIGURE 3.3 – Écoulement de Stokes dans un réseau

3.3 Le poumon comme arbre résistif

On cherche ici à modéliser l'écoulement de l'air dans l'arbre bronchique de façon simplifiée, en s'appuyant sur une description en réseau. Le point de départ de cette approche est l'écoulement de Poiseuille, solution analytique des équations de Stokes dans un tuyau de section circulaire. La figure 3.3 (haut) représente un réseau élémentaire impliquant 3 tubes. Si l'on suppose que les longueurs des tubes sont significativement plus grands que leurs diamètres respectifs, on peut s'attendre à ce que les variations de pression au niveau de la zone de bifurcation (dont la taille est de l'ordre des diamètres) soient négligeables devant des variations de pression le long des tubes. Ces considérations conduisent à décrire le réseau réel plongé dans l'espace à 3 dimensions par un réseau symbolique à 4 points et 3 arêtes, la zone de bifurcation ayant été réduite en un sommet du réseau, en lequel une pression ponctuelle est définie. Cette approximation peut être justifiée rigoureusement³ par des développements asymptotiques rigoureux⁴.

Notons u_i , $i = 0, 1, 2$ les flux au travers des conduits (flux comptés positivement lorsqu'ils sortent du réseau), et r_i , $i = 0, 1, 2$ les résistances des 3 tubes impliqués. La loi de Poiseuille s'écrit

$$p - p_0 = r_0 u_0, \quad p - p_1 = r_1 u_1, \quad p - p_2 = r_2 u_2,$$

et la conservation du volume (loi de Kirchhoff) impose

$$u_0 + u_1 + u_2 = 0.$$

Les flux peuvent être éliminés, ce qui conduit à

$$\left(\frac{1}{r_0} + \frac{1}{r_1} + \frac{1}{r_2} \right) p - \frac{p_0}{r_0} - \frac{p_1}{r_1} - \frac{p_2}{r_2} = 0.$$

4. On se reportera par exemple à

E. Marusić-Paloka, Incompressible newtonian flow through thin pipes Proceedings of the second conference on Applied Mathematics and Scientific Computing, held June 4-9, 2001 in Dubrovnik, Croatia, Springer, 2003.

Noter que cette approximation n'est rigoureusement justifiée que pour des rapports d'aspects asymptotiquement infinis, ou tout du moins significativement supérieurs à ceux que l'on observe dans le cas du poumon (de l'ordre de 3, qu'il faut beaucoup de bonne volonté pour considérer comme infini). Néanmoins, pour des rapports d'aspect d'ordre 1, même si l'estimation des résistances effectives des branches ne peut pas être exprimée par une formule explicite, la description d'un réseau complexe par un réseau résistif reste pertinente du fait de la linéarité des équations de Stokes impliquées. Par ailleurs, si l'on estime les résistances des tubes en supprimant simplement la zone de la bifurcation, on peut montrer rigoureusement que la résistance effective du réseau résistif associé est inférieure à la résistance que l'on estimerait par résolution complète des équations de Stokes.

Cet innocente équation, où l'on considère que p_1 , p_2 et p_3 sont imposés, peut être vue comme une forme (très) rudimentaire de problème de Laplace discret avec conditions de Dirichlet. Si l'on prescrit les flux (i.e. les différences $p_0 - p_i$), il s'agit alors d'un problème de Poisson avec conditions de Neuman. Le cadre général traitant de ces problèmes sur un réseau quelconque est développé dans le chapitre 2, page 37.

La situation est particulièrement simple si l'on considère l'arbre bronchique comme un arbre dyadique 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 N . 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'illustration, la figure 3.4 (gauche) représente un arbre à 4 générations.

On suppose ici l'arbre *symétrique*, i.e. tel que les propriétés géométriques des bronches, et donc leurs résistances, sont uniformes sur chaque génération. Si l'on considère le problème de Dirichlet permettant de définir la résistance équivalente, la pression sera, par symétrie du problème, constante à chaque génération. On peut ainsi remplacer chaque génération de bronche par un bouquet de bronches de même résistance en parallèle. Ces bouquets peuvent être identifiés à une unique bronche virtuelle de résistance équivalente $\bar{r}_n = r_n/2^n$. Ces différents étages, sont en série, de telle sorte que la résistance globale vaut

$$\bar{R} = \sum_{n=0}^N \bar{r}_n = \sum_{n=0}^N \frac{r_n}{2^n}. \quad (3.11)$$

Plus précisément, si l'on considère que les bronches d'une même génération n ont la même longueur ℓ_n et le même diamètre d_n , la loi de Poiseuille permet de préciser l'expression (3.11) :

$$\bar{R} = C \sum_{n=0}^N \frac{1}{2^n} \frac{\ell_n}{d_n^4}. \quad (3.12)$$

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

$$\bar{R} = r_0 \sum_{k=0}^N \frac{1}{2^k} \frac{1}{\lambda^{3k}}. \quad (3.13)$$

Remarque 3.2. Remarquer que cette série diverge dès que λ est inférieur à $2^{-1/3}$. Selon les données expérimentales, λ est situé autour de $0.85 > 2^{-1/3} (\approx 0.79)$, 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 sorte que le poumon infini extrapolé remplit (très largement, d'une certaine manière, du fait de l'inégalité stricte) l'espace euclidien.

3.4 Numérotation dyadique

On s'intéresse dans cette section à une question qui peut sembler mineure, mais qui a des conséquences importantes sur la manière dont on peut concevoir, lorsque l'on fait tendre le nombre de générations vers l'infini, l'ensemble des “feuilles” (on parle alors d'ensemble des bouts). Pour un arbre fini, la problématique est la suivante : la numérotation canonique 0, 1, ... des feuilles d'un arbre dyadique à N génération ne respecte pas la structure de l'arbre, au sens où $j - i$ ne donne aucune indication sur la proximité de i et de j vis-à-vis de l'arbre c'est à dire, si l'on voit cet arbre comme un arbre généalogique, l'ancienneté de leur plus proche ancêtre commun. Par exemple pour l'arbre à 4 générations représenté sur la figure 3.4, la feuille 0 est très proche (sœur) de la feuille 1), alors que la feuille 7 est très éloignée de la feuille 8, bien que dans les deux cas la différence soit égale à 1.

La notion de valuation dyadique permet de construire une distance plus conforme respectueuse de la structure de l'arbre. En premier lieu, on identifie Γ , ensemble des feuilles, à $\mathbb{Z}/2^N\mathbb{Z}$. En fait, pour la

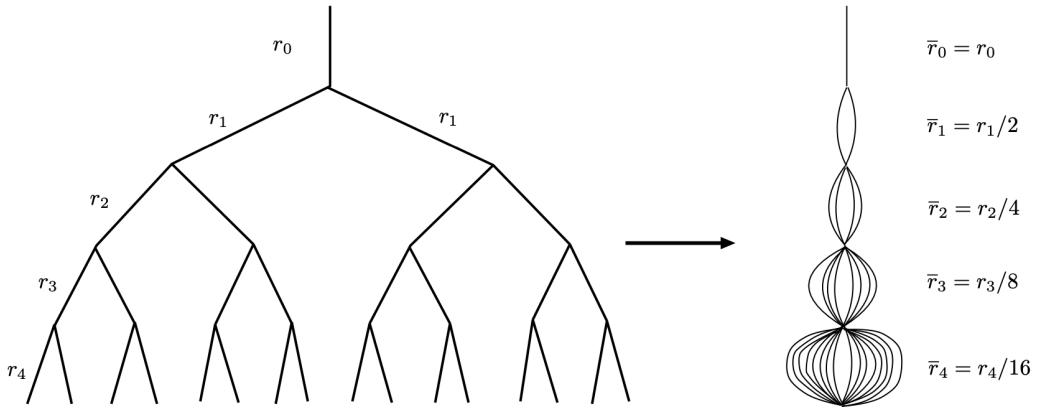


FIGURE 3.4 – Arbre dyadique régulier

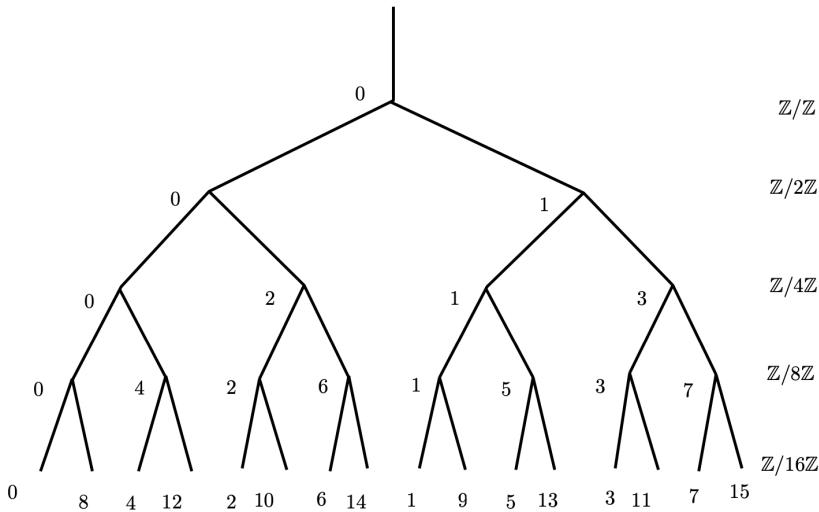


FIGURE 3.5 – Numérotation dyadique

construction de la distance, nous pourrions nous contenter de considérer l'ensemble $\{0, 1, \dots, 2^N - 1\}$ sans structure particulière, mais le choix fait donne un cadre à l'approche qui consiste à faire tendre N vers $+\infty$.

Pour tout $a \in \Gamma = \mathbb{Z}/2^N\mathbb{Z}$, on écrit $a = 2^k a'$, où a' est impair. On appelle k la valuation dyadique de a , et l'on définit $|a|_2$ comme 2^{-k} . Pour tous a et b dans Γ , on définit alors leur distance comme

$$d(a, b) = |b - a|_2.$$

On numérote alors les feuilles de l'arbre comme représenté sur la figure 19.5. Cette numérotation peut se construire par récurrence à partir de la racine. Si une feuille à la génération n a un indice $a \in \mathbb{Z}/2^n\mathbb{Z}$, ses deux filles ont des indices $a \in \mathbb{Z}/2^{n+1}\mathbb{Z}$ (gauche) et $a + 2^n \in \mathbb{Z}/2^{n+1}\mathbb{Z}$ (droite).

On vérifie immédiatement que, pour ce choix d'indexation, la distance dyadique sur Γ_N (l'ensemble des feuilles de l'arbre à N générations) est compatible avec la structure de l'arbre. Pour le cas $N = 4$ par exemple, représenté sur la figure 19.5. Pour toute feuille a du lobe de gauche et toute feuille b du lobe de droite, $b - a$ est impair, donc $|b - a|_2 = 1$, qui est le diamètre de Γ_n pour tout n . Pour toute paire $\{a, b\}$ de feuilles-sœurs, la différence est divisible par $8 = 2^{n-1}$, la distance est donc $1/8$, qui

est la plus petite distance possible entre deux feuilles, i.e. la *granularité* de l'espace. Cette granularité diminue bien sûr quand on augmente le nombre de génération.

Remarque 3.3. La distance ainsi construite sur $\Gamma_n = \mathbb{Z}/2^n\mathbb{Z}$ a la propriété d'être *ultramétrique*, c'est à dire qu'elle jouit d'une inégalité triangulaire renforcée :

$$d(a, b) \leq \max(d(a, c), d(b, c)) \quad \forall a, b, c.$$

L'une des conséquences, qui invalide la représentation usuelle des distances euclidiennes entre points du plan, est que tout élément d'une boule est aussi centre de cette boule. Cette propriété s'étend à la version infinie Γ_∞ présentée ci-dessous.

Vers un nombre de générations infini

L'approche proposée précédemment permet d'identifier l'ensemble des bouts de l'arbre dyadique infini à un objet noté \mathbb{Z}_2 , qui est l'ensemble des entiers dyadiques. D'un point de vue très abstrait, on peut voir cet espace comme ce qu'on appelle la limite projective du système

$$(\mathbb{Z}/2^n\mathbb{Z}, \varphi_n^m),$$

où $\mathbb{Z}/2^n\mathbb{Z}$ est l'ensemble des feuilles de l'arbre à n générations, et φ_n^m la surjection canonique de $\mathbb{Z}/2^m\mathbb{Z}$ dans $\mathbb{Z}/2^n\mathbb{Z}$, pour $m \geq n$. Plus précisément, on identifie l'ensemble limite à l'ensemble

$$\Gamma_\infty = \varprojlim (\mathbb{Z}/2^n\mathbb{Z}, \varphi_n^m) = \{(z_n)_{n \in \mathbb{N}} \in \prod (\mathbb{Z}/2^n\mathbb{Z}), \varphi_n^{n+1}(z_{n+1}) = z_n \quad \forall n \geq 0\}$$

des suites (z_0, z_1, \dots) , avec $z_n \in \Gamma_n$ pour tout n , consistantes avec le système de surjection (φ_n^m) . Cet ensemble limite projectif est ce que l'on appelle l'ensemble des entiers dyadiques, noté \mathbb{Z}_2 . On peut représenter tout élément de cet espace comme une suite infinie de *bits* : toute suite $(z_n)_{n \in \mathbb{N}} \in \mathbb{Z}_2$ est associée univoquement à une suite $(a_n)_{n \geq 0}$ de 0 ou 1, telle que

$$z_m = \sum_{n=0}^{m-1} a_n 2^n,$$

ce qui justifie que l'on note un *bout* (i.e. chemin vers l'infini) comme un élément de \mathbb{Z}_2 , c'est à dire comme une suite infinie de 0 et de 1, écrite traditionnellement de la droite vers la gauche : $q = \dots a_n \dots a_1 a_0 \in \Gamma_\infty = \mathbb{Z}_2$.

3.5 Arbre résistif non symétrique

On considère ici un arbre dyadique à N générations⁵, with root o , which we suppose is set to pressure 0. On note (x_n^k) les sommets, et (e_n^k) les arêtes⁶, with $0 \leq n \leq N$, $0 \leq k < 2^n$. On note p_n^k la pression au sommet x_n^k , par r_n^k la résistance de e_n^k , et par u_n^k le flux au travers de e_n^k (see Fig. 3.6).

Poiseuille's law writes

$$p_n^k - p_{n+1}^{2k} = r_{n+1}^{2k} u_{n+1}^{2k}, \quad p_n^k - p_{n+1}^{2k+1} = r_{n+1}^{2k+1} u_{n+1}^{2k+1} \quad 0 \leq k < 2^n,$$

et la loi des noeuds

$$u_n^k - u_{n+1}^{2k} - u_{n+1}^{2k+1} = 0.$$

5. Nous suivrons la convention suivante : un arbre à N génération a en fait $N + 1$ étages, qui comprend l'étage 0 constitué de la simple trachée. Un arbre à 1 génération est ainsi constitué de 4 sommets, 3 arêtes, et présente une bifurcation entre les étages 0 et 1.

6. On choisit d'orienter les arêtes de la racine vers les feuilles, i.e.

$$e_{n+1}^{2k} = [x_n^k, x_{n+1}^{2k}], \quad e_{n+1}^{2k+1} = [x_n^k, x_{n+1}^{2k+1}].$$

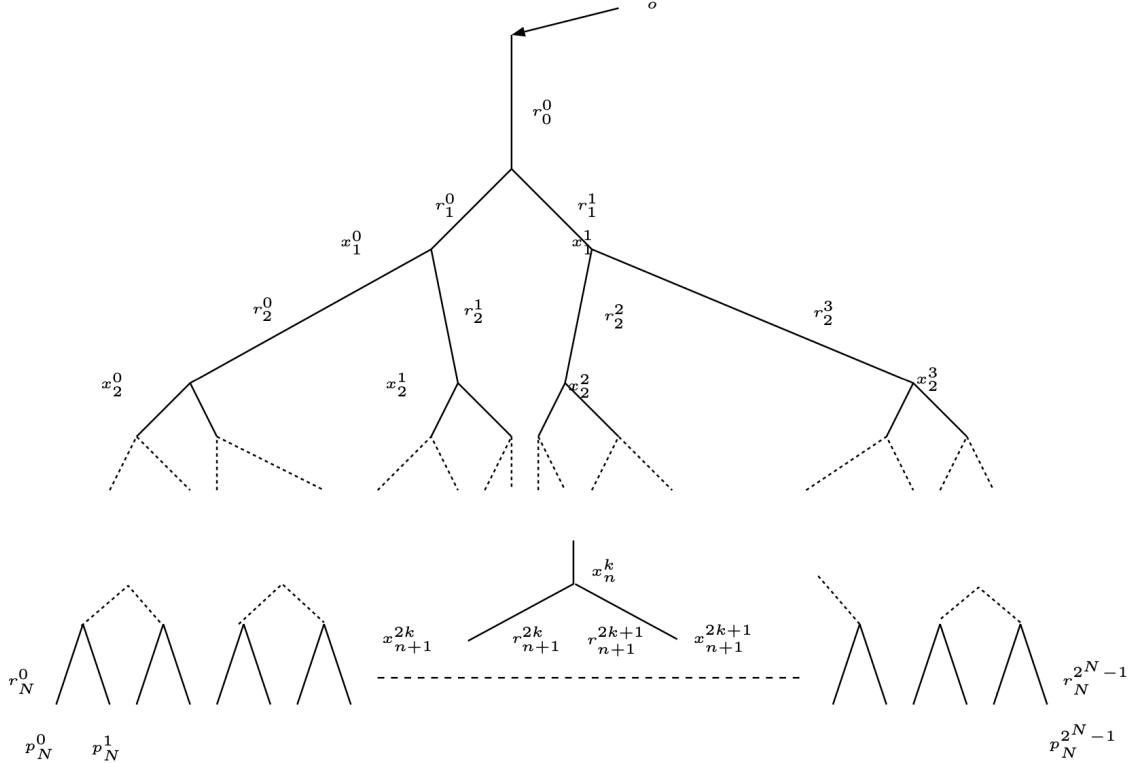


FIGURE 3.6 – N -generation resistive tree

External pressure (at root o) being set to 0, the generalized Poiseuille's law across the tree takes the form of a linear relation between pressures $p = (p_N^k)_{0 \leq k < 2^N}$ and fluxes $u = (u_N^k)_{0 \leq k < 2^N}$

$$0 - p = Ru,$$

and we aim at expressing how matrix R depends on the resistances.

Matrice de résistance On considère que les flux sur la frontière u_N^k , $0 \leq k < 2^N$ sont connus. Let us determine pressures along the path from 0 to x_N^0 . La loi de Poiseuille sur la trachée (qui connecte o et x_0^0) s'écrit

$$0 - p_0^0 = r_0^0 u_0^0,$$

où, par conservation,

$$u_0^0 = \sum_{k=0}^{2^N-1} u_N^k.$$

A l'étage suivant, on a

$$p_0^0 - p_1^0 = r_1^0 u_1^0,$$

où, comme précédemment, u_1^0 peut être calculé en sommant les flux sur la première moitié des feuilles. On obtient ainsi p_2^0, p_3^0, \dots , et finalement

$$\begin{aligned} p_N^0 &= -r_0^0 u_0^0 - r_1^0 u_1^0 - \cdots - r_N^0 u_N^0 \\ &= -\sum_{n=0}^N r_n^0 \left(\sum_{k=0}^{2^{(N-n)}-1} u_N^k \right). \end{aligned}$$

L'approche peut être généralisée sur chaque chemin reliant o à n'importe quel point de la frontière x_N^k :

$$[o, x_0^0, x_1^{k_1}, \dots, x_n^{k_n}, \dots, x_N^k],$$

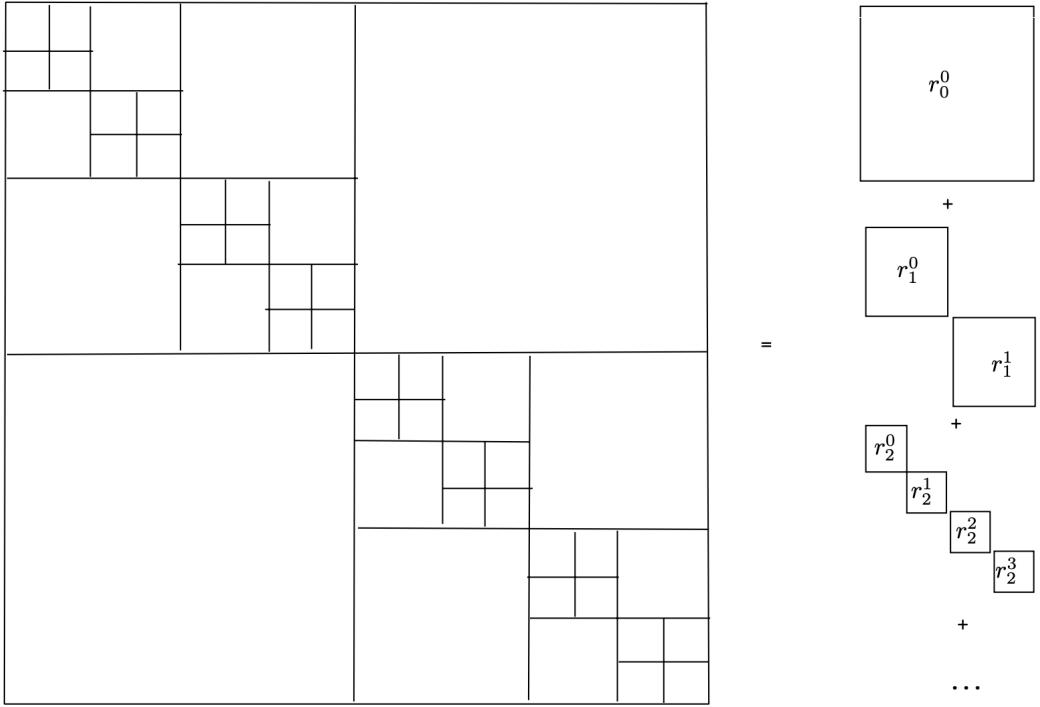


FIGURE 3.7 – Construction de la matrice R

which allows to express $p = (p_N^k)_{0 \leq k < 2^N}$ with respect to fluxes as

$$0 - p = Ru,$$

where R is the resistance matrix, which is expressed below. Let J_n be the $2^n \times 2^n$ matrix with all entries equal to 1 (one-rank matrix), R writes

$$\begin{aligned}
R &= r_0^0 J_N + \begin{pmatrix} r_1^0 J_{N-1} & 0 \\ 0 & r_1^1 J_{N-1} \end{pmatrix} + \begin{pmatrix} r_2^0 J_{N-2} & 0 & 0 & 0 \\ 0 & r_2^1 J_{N-2} & 0 & 0 \\ 0 & 0 & r_2^2 J_{N-2} & 0 \\ 0 & 0 & 0 & r_2^3 J_{N-2} \end{pmatrix} + \dots \quad (3.14) \\
&\quad + \begin{pmatrix} r_N^0 & 0 & \dots & \dots & 0 \\ 0 & r_N^1 & 0 & \dots & \vdots \\ \vdots & \ddots & & & \vdots \\ \vdots & & \ddots & & \vdots \\ 0 & \dots & \dots & 0 & r_N^{2^N-1} \end{pmatrix}.
\end{aligned}$$

Neumann and Dirichlet problems We have actually presented a way to solve the so called *Neumann* problem for a justification of this term) : considering a N -generation dyadic tree, we assumed the fluxes on the last generation were given, and we described a way to built the pressure field all over the tree.

Proposition 3.4. (Solution to the Neumann problem on a dyadic tree)

We consider a dyadic tree (see Fig. 3.6, page 73) characterized by its resistances

$$r_n^k, \quad 0 \leq n \leq N, \quad 0 \leq k \leq 2^n - 1,$$

We assume that the fluxes at the last generation are prescribed, and that the root o is set to pressure 0. We denote by \bar{u}_n^k the flux going through edge e_n^k , which is by mass conservation,

$$\bar{u}_n^k = \sum_{\ell=k}^{(k+1)2^{N-n}-1} u_n^\ell.$$

Then the collection of pressures at generation n , $p_n = (p_n^k)_k$ can be obtained recursively as follows

$$p_0^0 = -r_0^0 \bar{u}_0^0, \quad p_1^0 = p_0^0 - r_1^0 \bar{u}_1^0, \quad p_1^1 = p_0^1 - r_1^1 \bar{u}_1^1,$$

and, when the values at generation n are known,

$$p_{n+1}^{2k} = p_n^k - r_{n+1}^{2k} \bar{u}_{n+1}^{2k}, \quad p_{n+1}^{2k+1} = p_n^k - r_{n+1}^{2k+1} \bar{u}_{n+1}^{2k+1},$$

for $k = 0, \dots, 2^n - 1$.

Démonstration. This expression is a straightforward application of Poiseuille's law, as described previously, starting from the first edge (which contains the root), and then applied recursively throughout the generations. \square

Remarque 3.5. The problem that we have considered can be seen as a mixed Dirichlet to Neumann problem, as the pressure at the root is prescribed, whereas fluxes are prescribed at other boundary vertices. Actually, the flux through the root is also fixed by mass conservation (it is the sum of all fluxes of the last generation). As in the PDE (Partial Differential Equation) context, the Neumann problem is actually ill posed in the following sense : it requires a condition on the fluxes which are prescribed (i.e. global mass conservation) for a solution to exist, and, whenever this condition is met, the solution (in terms of a pressure field defined on vertices) is defined up to a constant. In the previous proposition, we have simply dropped the Neumann (flux) condition at the root to relax the condition, which is then automatically met, and we have prescribed the pressure at that note to select the solution, and thereby obtain uniqueness.

The fact that the Neumann problem is trivial (it does not require the solution to a linear system) reflects the hierarchical structure of the tree : whenever the fluxes at the boundary are known, then the flux through an edge is obtained straightforwardly by writing mass conservation on the subtree stemming from this edge. On the contrary, the Dirichlet problem, which consists in computing pressures within the tree from prescribed value on the boundary, is a nontrivial problem, except in the case where the tree is symmetric, and the pressure field on the last generation is uniform. In the latter situation, symmetries of the problem allow to identify all vertices of a given generation, which leads to a trivial linear network : resistances in series, with prescribed pressures at both ends. But as soon as the pressure field is not uniform (even if the tree is regular), the problem is non trivial and requires the solution of a large linear system.

Equivalent resistance The equivalent resistance of the tree is defined as follows : one prescribes pressure 0 at the root, and a constant pressure P at the leafs. The flow rate q through the root linearly depends on $0 - P$, and the equivalent resistance is defined as the proportionality constant

$$0 - P = \bar{R}q.$$

As we mentioned previously, in case of a symmetric dyadic tree (uniform resistance r_n within generation n), \bar{R} can be computed straightforwardly as

$$\bar{R} = \sum_{n=0}^N \frac{r_n}{2^n}.$$

In the general case, the equivalent resistance can be obtained by prescribing a pressure -1 at all ends of the tree, keeping 0 at root o , and computing the total flux. Let us introduce

$$\vec{e} = (1, 1, \dots, 1)^T \in \mathbb{R}^{2^N}.$$

The prescribed pressure field is $-\vec{e}$. The total flux through the tree is $u \cdot \vec{e}$, where $u = (u_N^k)_{0 \leq k < 2^N}$ is the vector of fluxes through the leafs. As a consequence

$$\bar{R} = (\vec{e} \cdot R^{-1} \vec{e})^{-1}. \quad (3.15)$$

Note that the computation of the equivalent resistance in the general case necessitates the solution to a full $2^N \times 2^N$ linear system. This system can be checked to be ill-conditioned for a tree built according to physiological observations. For a geometric tree, the *condition number* can be proved to behave like the reciprocal of \bar{r}_n/\bar{R} (where \bar{r}_n is the equivalent resistance of the n -th generation, see Eq. (3.11)), which is the relative contribution of the last generation to the equivalent resistance. As the series $\sum \bar{r}_n$ has a convergent behavior for the physiological lung, this contribution goes to 0, and therefore the condition number grows geometrically with the number of generations. If one considers non-symmetric trees that are perturbations of the actual geometric tree (so that the condition number can be expected to be close to that of the geometric tree), the ill-conditioned character of the matrix R can make it difficult to solve the linear system involved in the expression (3.15) of the effective resistance.

3.6 Arbre optimal ?

Nous explorons dans cette section les propriétés d'optimalité de l'arbre bronchique. Le problème d'optimisation est formulé de la manière suivante : on considère un arbre résistif à N générations. On suppose que les bronches ont toutes la même forme, et toutes la même taille au sein d'une génération. On note v_n le volume d'une bronche à la génération n . Du fait que la résistance d'un tuyau est proportionnelle à la longueur divisée par la puissance 4 du diamètre, l'homogénéité vis à vis de la taille est de -3 , la résistance est donc inversement proportionnelle au volume. On prendra une constante de proportionnalité égale à 1 pour simplifier : la résistance d'une bronche à la génération n est donc $1/v_n$, de telle sorte que la résistance globale vaut (équation (3.11))

$$R(v) = R(v_0, v_1, \dots, v_N) = \sum_{n=0}^N \frac{r_n}{2^n} = \sum_{n=0}^N \frac{1}{2^n v_n}.$$

Le volume occupé par l'arbre (volume totale des bronches) est simplement

$$V(v) = \sum_{n=0}^N 2^n v_n.$$

Proposition 3.6. Il existe une unique collection $v \in]0, +\infty[^{N+1}$ de volume qui minimise la résistance sous contrainte de volume fixé $V(v) \leq V_{max}$. Pour cet optimum, les v_n sont en progression géométrique de rapport $1/2$.

Démonstration. Notons en premier lieu que, la résistance tendant vers $+\infty$ quand l'un des volumes tend vers 0, on peut se ramener à un problème de minimisation sur $[\eta, +\infty[^{N+1}$, avec $\eta > 0$. Comme chaque volume est borné par V_{max} d'après la contrainte, on se ramène à la minimisation d'une fonction continue sur $[\eta, V_{max}]^{N+1} \cup K$ (où K est l'ensemble des $v \in \mathbb{R}_+^{N+1}$ qui vérifient la contrainte), qui est compact, la résistance atteint son minimum en un point v . Comme la résistance est une fonction strictement convexe, et que le domaine admissible est convexe, ce minimiseur v est unique. Si la contrainte n'était pas saturée en ce minimiseur, on pourrait augmenter l'un quelconque des volumes sans la violer, en diminuant la valeur de la résistance. On a donc $V(v) = V_{max}$. Le vecteur v est donc aussi minimiseur de la résistance R sur l'ensemble des champs de $v \in]0, +\infty[^{N+1}$ qui vérifient la contrainte d'égalité $V(v) = V_{max}$. Il s'agit d'une contrainte affine, que l'on peut écrire sous la forme $w \cdot v = V_{max}$. On a donc d'après la proposition 13.12, page 263, existence de $\lambda \in \mathbb{R}$ tel que

$$\nabla R + \lambda w = 0,$$

d'où

$$-\frac{1}{2^n v_n^2} + \lambda 2^n = 0,$$

d'où l'on déduit $v_n = c 2^{-n}$.

On peut proposer une autre démonstration qui ne passe pas par les multiplicateurs de Lagrange. On écrit la résistance globale R et le volume V comme fonctions de la collection de longueurs

$$\ell = (\ell_0, \ell_1, \dots, \ell_N) \in]0, +\infty[^{N+1},$$

en prenant pour simplifier une constante de proportionnalité pour la résistance égale à 1 :

$$R(\ell) = \sum_{n=0}^{N+1} \frac{1}{2^n \ell_n^3} , \quad V(\ell) = \sum_{n=0}^{N+1} 2^n \ell_n^3.$$

Le problème consiste donc à minimiser $R(\ell)$ sous la contrainte $V(\ell) \leq V_{max}$. Comme précédemment, la contrainte de volume est nécessairement saturée (sinon on peut diminuer la résistance sans violer la contrainte). On a

$$N+1 = \sum_{n=0}^N 1 = \sum_{n=0}^N \frac{1}{\sqrt{2^n \ell_n^3}} \sqrt{2^n \ell_n^3} \leq \left(\sum_{n=0}^N \frac{1}{2^n \ell_n^3} \right)^{1/2} \left(\sum_{n=0}^N 2^n \ell_n^3 \right)^{1/2} = R^{1/2} V^{1/2},$$

On a donc

$$R \geq \frac{(N+1)^2}{V},$$

Si $2^n \ell_n^3$ est constant égal à α , on a pour ce cas particulier

$$V = (N+1)\alpha, \quad R = \frac{N+1}{\alpha} = \frac{(N+1)^2}{V},$$

qui réalise donc l'égalité. La résistance est donc minimale pour $2^n \ell_n^3$ constant, c'est à dire

$$\ell_n = c \lambda^n, \quad \lambda = 2^{-1/3},$$

qui conduit donc à la même conclusion, du fait que le volume v_n est proportionnel à ℓ_n^3 . \square

Remarque 3.7. Les volumes diminuant de moitié d'un génération à l'autre, le facteur de réduction des distances est de $1/2^{1/3} \approx 0.79$. On pourra rapprocher ce nombre du facteur effectivement mesuré en pratique, au moins sur les générations intermédiaires (entre la génération 4 et la génération 15), qui est autour de 0.85. On se reportera à Mauroy et al.⁷ pour plus de détails sur ces questions.

3.7 Vers un poumon infini

On considère ici un "poumon infini", c'est-à-dire un arbre dyadique possédant une infinité de générations. On se place dans le cadre du chapitre ??, étendu au cas d'un nombre infini de sommet. La racine de l'arbre est noté o , en revanche, l'ensemble des sommets qui jouaient le rôle de frontière pour les réseaux finis est maintenant vide. Le problème consiste précisément à explorer la possibilité que du fluide puisse traverser l'arbre en rentrant (ou sortant) par l'infini.

Le cadre que nous définissons ci-dessous s'appliquant à un réseau infini quelconque, nous considérons pour l'instant un tel réseau enraciné $\mathcal{N} = (V, E, r, o)$, sans hypothèse de structure. On suppose que V est infini, que le réseau est connexe, et que chaque sommet appartient à un nombre fini de côtés.

La puissance dissipée au sein du réseau par circulation d'un champ de flux sur les côtés conduit à la définition de l'espace d'énergie pour les flux :

$$L^2(\mathcal{N}) = \left\{ u \in \mathbb{R}^E, \quad \sum_e r(e) |u(e)|^2 < +\infty \right\}.$$

⁷. B. Mauroy, M. Filoche, E. R. Weibel, & B. Sapoval, An optimal bronchial tree may be dangerous, Nature volume 427, pages633–636 (2004). <https://www.nature.com/articles/nature02287>

On considère que ces flux sont induits (loi de Poiseuille) par des sauts de pressions aux extrémités des arêtes, l'espace naturel en pressions est donc

$$H^1(\mathcal{N}) = \left\{ p \in \mathbb{R}^V, p(o) = 0, |p|_1^2 = \sum_e c(e) |p(y) - p(x)|^2 < +\infty \right\}.$$

Il s'agit d'espaces de Hilbert séparables : $L^2(\mathcal{N})$ est un espace de type ℓ^2 à poids, et $c\partial^*$ envoie isométriquement H^1 vers L^2 . On définit H_0^1 comme l'adhérence de $D(\mathcal{N})$, sous-espace des champs de pression nuls sauf en un nombre fini de sommets. Ces définitions élémentaires permettent d'exprimer précisément la version abstraite du problème de définition d'un espace de trace l'infini : l'espace quotient H^1/H_0^1 est-il trivial ou pas ?

Théorème 3.8. Soit $\mathcal{N} = (V, E, r, o)$ un réseau infini connexe, H^1 et H_0^1 les espaces de Sobolev associés. On a

$$\bar{R}(\mathcal{N}) = +\infty \iff H^1/H_0^1 = \{0\}.$$

3.8 Particules et dépôt

Nous nous intéressons ici au destin de petites particules inhalées lors du processus de ventilation. Nous nous attacherons en particulier à concevoir des outils méthodologiques permettant de déterminer si ces particules se déposent à l'intérieur du poumon, voire sur la surface des alvéoles, lors du cycle ventilatoire. Nous nous limiterons à des particules de tailles suffisamment petite pour que l'écoulement de l'air autour de chacune d'elle puisse être décrit par des équations de Stokes, c'est à dire que le nombre de Reynolds (voir définition 9.12, page 212) particulaire (associé à la vitesse de la particule et de l'air environnant) soit petit devant 1. Par ailleurs nous supposerons que ces particules sont des gouttes d'un liquide d'une densité proche de celle de l'eau, et de taille là encore suffisamment petite pour que la tension surfacique⁸ préserve une forme sphérique.

Selon les hypothèses faites ci-dessus, l'interaction dynamique entre la goutte et le fluide environnant est dominé par la loi dite de Faxén

$$F = 6\pi\mu a (U_f - U_p), \quad (3.16)$$

où U_p est la vitesse de la particule, et U_f la vitesse du fluide environnant⁹.

3.8.1 Sédimentation.

Considérons une particule de densité ρ , soumise à l'action de son propre poids. On considère ρ très supérieur à la densité du gaz environnant, de telle sorte que la poussée d'archimète est négligeable. En régime stationnaire, dans l'hypothèse où la force d'interaction est bien donnée par (3.16), la particule

8. Ce que l'on appelle tension surfacique résulte de forces internes de cohésion au sein d'un fluide, dont la résultante est nulle au cœur du domaine occupé par ce fluide, mais non nulle sur la frontière lorsque cette dernière est courbe. L'effet résultant peut être décrit par un saut de pression au travers de la surface, proportionnel à la courbure moyenne locale, qui tend à régulariser les surfaces. On peut vérifier que cette tension surfacique (i.e. ce saut de pression) agit dans la direction de l'opposé du gradient de la fonctionnelle aire, elle tend donc à minimiser cette aire. Pour une goutte d'un volume de liquide donné, ce phénomène, hors de toute autre sollicitation, tend à donner à la goutte une forme sphérique (qui minimise l'aire à volume imposé). L'effet étant proportionnel à la courbure, il est très significatif pour des diamètres petits, au point de figer essentiellement la goutte sous forme sphérique pour des diamètres qui tendent vers 0, toutes choses égales par ailleurs.

9. Cette notion peut sembler ambiguë, puisque la vitesse du fluide s'identifie localement à celle de la particule. Il faut avoir en tête cette particule comme baignant dans un volume mésoscopique de fluide, de taille significativement plus grande que le diamètre de la particule, mais qui reste petit devant la taille du domaine d'intérêt dans sa globalité. La particule modifie localement la vitesse du fluide, qui sinon serait considérée comme constante sur ce volume élémentaire, et c'est cette valeur constante avant modification, c'est à dire la vitesse du fluide loin de la particule (relativement à la taille de cette dernière) qui est considérée comme la vitesse du fluide. Nous avons implicitement supposé que la particule était seule dans son voisinage mésoscopique, c'est en effet une hypothèse qui conditionne la validité de l'approche, qui ne s'applique pas aux fortes densités de particules.

chute à vitesse constante. Cette vitesse v_s équilibre les forces visqueuses et le poids :

$$\frac{4}{3}\pi a^3 \rho_\ell g = 6\pi\mu a v_s \quad \text{d'où} \quad v_s = \frac{2}{9} \frac{\rho_\ell g a^2}{\mu},$$

où $\mu = 2 \times 10^{-5}$ Pa s est la viscosité de l'air, $\rho_\ell = 1.2 \times 10^3$ kg m⁻³ la densité du constituant des particules (dans l'hypothèse où il s'agit d'un liquide proche de l'eau), et a le rayon. Pour une particule de diamètre 1 μm, on trouve par exemple

$$v_s \approx 3.25 \times 10^{-5} \text{ ms}^{-1} \approx 30 \text{ μms}^{-1}.$$

Noter que, le diamètre d'une alvéole étant de l'ordre de 200 μm, il faut au maximum 6 secondes à une telle particule pour se déposer, quelle que soit sa position initiale. Une suspension de ces particules dans les alvéoles se sera donc entièrement déposée au bout de ce temps, qui est de l'ordre de la durée du cycle respiratoire.

Une telle particule dans l'air libre peut en revanche être considérée comme en suspension (il lui faut 9 heures pour tomber d'une hauteur de 1 m).

Diffusion. Une particule dans un gaz (on suppose la taille de la particule significativement plus grande que la distance intermoléculaire) subit des chocs avec les molécules du gaz environnant. Ces chocs répétés dans toutes les directions induisent sur la particule un comportement de nature diffusive, quantifié par la loi de Stokes-Einstein :

$$D = \frac{k_B T}{6\pi\mu a}$$

où $k_B = 1.4 \times 10^{-23}$ est la constante de Boltzman, $T \approx 300 K$ la température (en Kelvin), μ la viscosité du fluide environnant, et a le rayon de la particule.

Dans le cas de l'inhalation de particules lors de la ventilation, il peut être intéressant d'estimer la distance typique parcourue par une particule qui diffuse selon la loi ci-dessus, pendant un temps de l'ordre de celui du cycle respiratoire, $t = 5$ s. Pour une particule de rayon a initialement située en un point de l'espace, sa probabilité de présence suivra après le temps t une loi Gaussienne d'écart-type $\sigma = \sqrt{2Dt}$ (voir (4.6), page 91). Pour un diamètre de 1 μm, on trouve $D \approx 2.2 \times 10^{-11} \text{ m}^2 \text{s}^{-1}$, soit $\sigma \approx 15 \text{ μm}$. Pour une particule dix fois plus petite, on trouve σ proche de 50 μm, qui est de l'ordre du rayon (la moitié plus précisément) de l'alvéole. Une suspension de telles particules initialement disposées uniformément dans l'alvéole se sera donc déposée en grande partie sur le bord par diffusion.

On peut estimer le temps caractéristique de dépôt par diffusion de façon un peu plus précise en utilisant une représentation spectrale de la solution de l'équation de la chaleur dans une sphère de rayon a .

3.8.2 Particule inertielle vs. traceur passif

Considérons une particule sphérique de densité ρ , de rayon a , lancée initialement à la vitesse U dans un fluide au repos. L'équation du mouvement s'écrit

$$\frac{4}{3}\pi a^3 \rho \dot{U} = -6\pi\mu a U,$$

de telle sorte que la vitesse de la particule va s'amortir exponentiellement avec un temps caractéristique

$$\tau_a = \frac{2}{9} \frac{\rho a^2}{\mu}.$$

De façon plus générale, τ est le temps caractéristique mis par une particule au sein d'un fluide en mouvement pour acquérir la vitesse du fluide dans son voisinage. Considérons maintenant une telle particule transportée par un fluide visqueux s'écoulant autour d'un obstacle de taille caractéristique L . La question de savoir si la particule va heurter l'obstacle ou au contraire suivre les lignes de courant

qui contournent cet obstacle peut se formuler en termes de temps caractéristique : le temps τ_a est-il très inférieur au temps mis par un élément de fluide pour contourner l'obstacle, auquel cas la particules va en effet suivre une ligne de courant et donc contourner l'obstacle, ou au contraire très supérieur, auquel cas la particule va heurter l'obstacle en suivant sa trajectoire balistique ? Le temps mis pour contourner l'obstacle est de l'ordre de L/U . Le rapport des deux nombres est donc un nombre sans dimension appelé

Definition 3.9. (Nombre de Stokes)

On considère une particule de densité ρ et de rayon a dans un fluide visqueux de viscosité μ . On définit le nombre de Stokes comme

$$\text{St} = \frac{2}{9} \frac{\rho a^2 U}{L\mu},$$

où U est la vitesse caractéristique du fluide, et L une taille caractéristique du phénomène considéré, qui correspond à la distance typique entre deux points en lesquels les vitesses du fluide sont significativement différentes.

La figure 3.8 représente les trajectoires de particules transportées dans une bifurcation du type de celles que l'on rencontre dans l'arbre bronchique. Les différentes trajectoires se distinguent par le nombre de Stokes (décroissant de gauche à droite). Pour un nombre de Stokes important (100), la particule continue sa trajectoire inertielle en ligne droite et impacte la frontière au niveau de la bifurcation. Pour des nombres de Stokes plus petits, la trajectoire est infléchie, mais pas suffisamment pour éviter l'impact. En dessous d'un nombre de Stokes autour de 1.2, la particule suit suffisamment le fluide pour éviter l'impact. Pour une gouttelette d'un fluide d'une densité proche de celle de l'eau, si l'on considère (inspiration au repos) la vitesse de l'ordre de 1 ms^{-1} , une taille caractéristique de $L = 2 \text{ cm}$, le diamètre correspondant à un nombre de Stokes unitaire est

$$2a = 2\sqrt{\frac{9L\mu}{2\rho U}} \approx 35 \text{ } \mu\text{m}.$$

Les particules de cette taille ou plus grosses auront donc tendance impacter la paroi des bronches dès la première génération. Les particules plus petites vont continuer leur route. On peut vérifier en estimant le nombre de Stokes local au niveau de chaque bifurcation qu'une particule qui a passé la première étape (i.e. n'a pas impacté) doit passer les suivantes. En effet, si la taille caractéristique diminue au fil des générations, la vitesse caractéristique aussi diminue, plus rapidement, de telle sorte que le nombre de Stokes, pour une taille de particule donnée, décroît au fil des générations.

La transition, pilotée par le nombre de Stokes, entre particule inertielle insensible au fluide ($\text{St} \rightarrow +\infty$), et à l'opposé traceur passif ($\text{St} \rightarrow 0$, la particule suit passivement le mouvement du fluide environnant) peut s'exprimer mathématiquement de la façon suivante :

Proposition 3.10. Soit $x \mapsto U(x) \in \mathbb{R}^d$ un champ donné (vitesse du fluide environnant), supposé borné et Lipschitzien sur \mathbb{R}^d , et $t \mapsto x_\tau(t)$ l'unique solution de l'équation différentielle

$$\ddot{x} = \frac{1}{\tau} (U(x) - \dot{x}),$$

sur $[0, T]$, pour les conditions initiales $x(0) = x^0$, $\dot{x}(0) = u^0$. Quand τ tend vers 0, x_τ converge uniformément vers $t \mapsto \xi(t)$ in $[0, T]$, et $u_\tau = \dot{x}_\tau$ converge uniformément vers $u = \dot{\xi}$ sur tout $[\eta, T]$, avec $\eta > 0$, où ξ est la solution “traceur passif”, i.e. la solution sur $[0, T]$ de l'équation d'ordre 1

$$\dot{\xi} = U(\xi), \quad \xi(0) = x^0.$$

Démonstration. On introduit

$$\varphi_\tau(t) = \frac{1}{2} |\dot{x}_\tau - U(x_\tau)|^2.$$

Sa dérivée en temps est

$$\dot{\varphi}_\tau = (\ddot{x}_\tau - \nabla U(x_\tau) \cdot \dot{x}_\tau) \cdot (\dot{x}_\tau - U(x_\tau)) = -\frac{2}{\tau} \varphi_\tau - (\nabla U \cdot \dot{x}_\tau) \cdot (\dot{x}_\tau - U(x_\tau)).$$

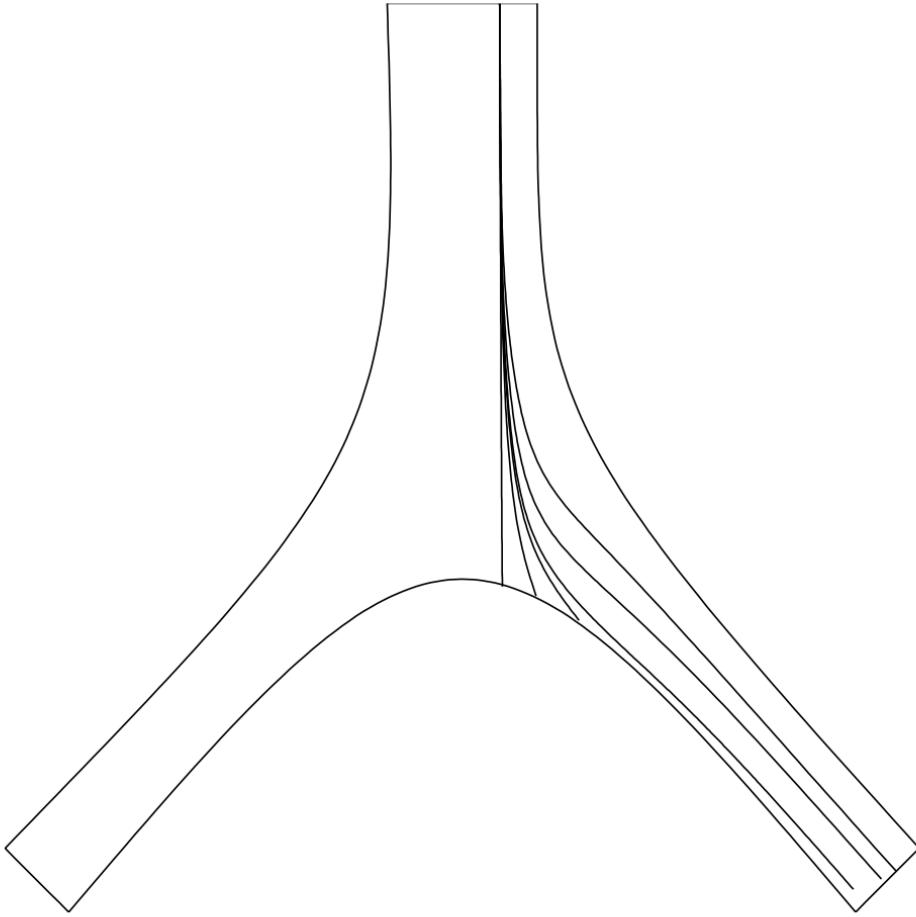


FIGURE 3.8 – Trajectoires de particules, pour $\text{St} = 100, 5, 2, 1.2, 1$, et 0.01 (de gauche à droite)

Par définition de φ_τ , on a

$$|\dot{x}_\tau| \leq \sqrt{2\varphi_\tau} + |U|,$$

qui entraîne

$$\begin{aligned} -(\nabla U \cdot \dot{x}_\tau) \cdot (\dot{x}_\tau - U(x_\tau)) &\leq |\nabla U| \left(\sqrt{2\varphi_\tau} + |U| \right) \sqrt{2\varphi_\tau} \\ &\leq \frac{1}{2} \left(|\nabla U|^2 \left(\sqrt{2\varphi_\tau} + |U| \right)^2 + 2\varphi_\tau \right) \\ &\leq |\nabla U|^2 \left(2\varphi_\tau + |U|^2 \right) + \varphi_\tau. \end{aligned}$$

On obtient finalement

$$\dot{\varphi}_\tau \leq \left(2 \|\nabla U\|_\infty^2 + 1 - \frac{2}{\tau} \right) \varphi_\tau + \|\nabla U\|_\infty^2 \|U\|_\infty^2.$$

Pour τ assez petit, le facteur devant φ_τ est plus petit que $-1/\tau$, de telle sorte que

$$\dot{\varphi}_\tau \leq -\frac{1}{\tau} \varphi_\tau + C.$$

La solution g de l'équation ci-dessus (en remplaçant l'inégalité par une égalité, et avec $g(0) = \varphi_\tau(0)$) est

$$g(t) = \varphi_\tau(0) e^{-t/2\tau} + C\tau(1 - e^{-t/\tau}).$$

Posant $\psi_\tau = \varphi_\tau - g$, on a $\dot{\psi}_\tau \leq -\psi_\tau/\tau$, d'où $\psi_\tau \leq 0$ pour tout temps. On déduit ainsi

$$0 \leq \varphi_\tau(t) \leq \varphi_\tau(0)e^{-t/\tau} + C\tau(1 - e^{-t/\tau}).$$

On a donc, pour tout $\eta > 0$, convergence uniforme sur $[\eta, T]$ de φ_τ vers 0, i.e. convergence uniforme de \dot{x}_τ vers $U(x_\tau)$.

Il reste à montrer la convergence uniforme de x_τ vers ξ sur $[0, T]$. Noter en premier lieu que, d'après l'inégalité sur φ_τ ci-dessus, et du fait que $\varphi_\tau(0)$ ne dépend pas de τ , $\varphi_\tau(t)$ est borné indépendamment de τ (que l'on peut prendre ≤ 1 , puisque sa vocation est de tendre vers 0).

On écrit maintenant

$$x_\tau(t) - \xi(t) = \int_0^t (\dot{x}_\tau - \dot{\xi}) = \int_0^t (\dot{x}_\tau - U(x_\tau)) + \int_0^t (U(x_\tau) - U(\xi)).$$

D'après la convergence uniforme de \dot{x}_τ vers $U(x_\tau)$ sur tout intervalle $[\eta, T]$, $\eta > 0$, et du fait que $|\dot{x}_\tau - U(x_\tau)|$ est borné, la première intégrale ci-dessus tend vers 0 quand τ tend vers 0, uniformément en $t \in]0, T]$. On a donc

$$|x_\tau(t) - \xi(t)| \leq C_\tau + \|\nabla U\| \int_0^t |x_\tau - \xi|,$$

d'où, d'après le lemme de Gronwall (proposition 11.24, page 242),

$$|x_\tau(t) - \xi(t)| \leq C_\tau e^{t\|\nabla U\|},$$

où C_τ tends vers 0 avec τ . On a donc bien convergence uniforme de x_τ , la trajectoire de la particule inertielle transportée par le fluide environnant, vers ξ , trajectoire du traceur passif, quand le temps caractéristique τ tends vers 0. \square

Chapitre 4

Lois de conservation, transport et diffusion

Sommaire

4.1	Vecteur flux, équation de conservation	84
4.2	Transport	85
4.3	Diffusion	89
4.3.1	Loi de Fick, équation de la chaleur	89
4.3.2	Cadre mathématique pour le problème de Poisson et l'équation de la chaleur	92
4.3.3	Interprétations stochastique du Laplacien	94
4.4	Transport - diffusion	95
4.5	Exercices	97

Ce chapitre porte sur l'élaboration et l'analyse (succincte ici) d'équations aux dérivées partielles qui expriment la conservation d'une certaine substance. Le cœur de l'approche consiste à exprimer de façon *eulérienne* une réalité essentiellement *lagrangienne* (voir section 1.1, page 8).

Bien qu'il s'agisse d'équations exprimant des principes physiques très simples, cette élaboration est assez délicate, notamment pour les raisons suivantes :

1. Les équations construites dans ce chapitre reposent sur une description de la matière par un continuum, décrit par une densité locale. Dans la réalité, on ne peut définir une notion de densité locale que sur les zones de taille très significativement à la tailles des "entités" impliquées (molécules, cellules, personnes, véhicules, voire même planètes si l'on se place à une échelle suffisamment grande). La démarche basée sur des objets géométriques (comme des petits disques au travers desquels on mesure le flux, ou de petits volumes dans lesquels on intègre la masse totale) n'a de sens que si la taille de ses objets ne descend pas en dessous d'un certain seuil, qui dépend de l'application considérée.
2. On peut établir des propriétés différentes aux solutions de l'équation de transport, basées sur la notion de trajectoires, qui sont bien définies si le champ est régulier (e.g. Lipschitz). S'il ne l'est pas, l'analyse de cette équation est beaucoup plus délicate. Mais on peut s'interroger sur le sens qu'a cette équation lorsque les trajectoires ne sont pas définies, puisqu'elle n'a été construite que pour précisément exprimer au niveau macroscopique le transport de "particules".
3. Toujours concernant l'équation de transport : le phénomène le plus simple que l'on puisse décrire concerne la cinématique d'un point matériel. Exprimé de façon Lagrangienne, cela consiste à suivre la trajectoire d'un point dans l'espace, selon une vitesse qui est simplement la dérivée de la position. Si l'on cherche à exprimer de transport de façon *eulérienne*, par une équation aux dérivées partielles basée sur une variable d'espace fixe, la particule est décrite par un objet extrêmement singulier, une mesure atomique (masse de Dirac) qui se déplace dans l'espace,

et ce le couple Dirac-vitesse associée ne peut être solution d'une EDP que dans un sens très généralisé, appelé sens faible.

4.1 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 4.1. (Vecteur flux)

Soit x un point du domaine occupé par la substance, 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 $Q(\varepsilon, n)$ la quantité de substance qui traverse D_ε par unité de temps, comptée positivement dans le sens n . Si, pour tout n , la quantité $Q(\varepsilon, n)/\varepsilon$ tende vers une limite quand ε tend vers 0, et que cette limite est linéaire par rapport à n , i.e. s'écrit $J \cdot n$, on appelle $J = J(x)$ le vecteur flux en x .

On se reportera à la section 4.4 pour des commentaires critiques sur le sens de cette définition.

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, qui peut s'écrire comme l'intégrale en volume de la divergence d'après le théorème d'Ostrogradsky (ou théorème de la divergence) :

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

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

$$\frac{\partial \rho}{\partial t} + \nabla \cdot J = 0. \quad (4.1)$$

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

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

Remarque 4.2. On peut, d'une certaine manière, rendre statique le problème d'évolution en le considérant comme un problème posé sur l'espace-temps. Une solution de l'équation de conservation peut alors se voir comme une densité $\rho(x, t)$ telle que le champ $F = (\rho, J)$ est à divergence nulle en espace-temps :

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

On peut écrire le théorème d'Ostrogradsky sur un volume cylindrique de l'espace-temps $C = \Omega \times]0, T[$:

$$0 = \int_C \nabla_{t,x} \cdot F = \int_{\partial C} F \cdot n = \int_0^T \int_{\partial\Omega} \nabla \cdot J - \int_\Omega \rho(x, 0) dx + \int_\Omega \rho(x, T) dx,$$

d'où, en écrivant $M(t)$ la quantité totale de matière dans Ω au temps t ,

$$M(t) = M(0) - \int_0^T \int_{\partial\Omega} \nabla \cdot J.$$

Dans le cas d'un transport à vitesse $u(x, t)$ (détailé ci-après), le flux s'écrit ρu . On peut alors voir F comme $\rho \times (1, u)$, où 1 est la vitesse (de 1 seconde par seconde) selon l'axe des temps.

4.2 Transport

Cette section traite de l'écriture sous forme d'équations aux dérivées partielles de phénomènes de transport. Dans le contexte de transport conservatif d'une variable extensive (de type masse, nombre de particules, ou volume pour un fluide incompressible), on parle parfois d'équation de continuité.

Modèle 4.3. (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

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho u) = f. \quad (4.2)$$

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

Remarque 4.4. Dans le cas où le champ convectant est à divergence nulle, l'équation s'écrit

$$\frac{\partial \rho}{\partial t} + u \cdot \nabla \rho = 0, \quad (4.3)$$

c'est cette dernière équation qui est le plus couramment appelée équation de transport. On prendra garde cependant au fait qu'elle correspond (dans le cas où le champ n'est pas à divergence nulle) au transport d'une quantité essentiellement intensive (voir remarque (1.4)). 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 à l'entité transportée, un label, une information, typiquement des variables qui ne se *somment* pas. L'équation de continuité (4.2) exprime bien en revanche le transport d'une variable extensive (la masse, ou le nombre d'entité), au travers de la variable intensive ρ qui est la densité de la mesure "masse" ou "nombre" d'entités" relativement à la mesure de volume sous-jacente.

On considère un champ de vecteur $u(x, t)$ défini sur $\mathbb{R}^d \times \mathbb{R}$, régulier. On note $X_s(x, t)$ le flot associé (supposé ici défini pour tout temps), défini par

$$\begin{cases} \frac{\partial X_s}{\partial t}(x, t) = u(X_s(x, t), t) \\ X_s(x, s) = x. \end{cases} \quad (4.4)$$

On peut montrer que toute solution de l'équation de transport non conservative (ou conservative avec un champ à divergence nulle) est constante le long des caractéristiques

Proposition 4.5. Soit $(x, t) \mapsto u(x, t)$ un champ de vitesse régulier (continu par rapport au couple, C^1 par rapport à la variable d'espace), et $\rho(x, t)$ une solution régulière de l'équation

$$\partial_t \rho + u_t \cdot \nabla \rho = 0.$$

Alors ρ est constant le long des caractéristiques $t \mapsto X_s(x, t)$ définies par (4.4).

Démonstration. On a

$$\frac{d}{dt} \rho(X_s(x, t), t) = \partial_t \rho(X_s(x, t), t) + \frac{\partial}{\partial t} X_s(x, t) \cdot \nabla \rho(X_s(x, t), t) = \partial_t \rho(X_s(x, t), t) + u_t \cdot \nabla \rho(X_s(x, t), t) = 0.$$

□

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

Proposition 4.6. Soit $\rho(x, t)$ une solution de l'équation

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

avec u régulier (continu, et continûment différentiable par rapport à la variable d'espace). Alors ρ vérifie

$$\rho(X_0(x, t), t) = \rho(x, 0) \exp \left(- \int_0^t \nabla \cdot u(X_0(x, s), s) ds \right).$$

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

$$\rho(y, t) = \rho_0(X_t(y, 0)) \exp \left(- \int_0^t \nabla \cdot u(X_t(y, s), s) ds \right).$$

Flot d'un champ de vecteur et équation de transport conservative Ce qui suit peut être considéré comme une approche alternative pour construire l'équation de transport sous forme conservative, sans passer par la notion de flux¹.

Soit $u(x, t)$ un champ de vitesse régulier en espace-temps, que l'on notera dans ce qui suit $u_t(x)$, et ρ_0 une densité positive régulière. On note $X_s(x, t)$ le flot associé, défini par (4.4). On note μ_t la mesure associée à la densité $\rho_t : d\mu_t = \rho_t(x) dx$. Pour tous $t, s \in \mathbb{R}$, on note (en s'autorisant à assimiler ici mesure et densité associée) μ_t la mesure image de μ_0 par l'application $X_0(\cdot, t)$, de telle sorte que μ_t est aussi la mesure image de μ_s par $X_s(\cdot, t)$, ce que l'on pourra noter

$$X_s(\cdot, t) \sharp \mu_s = \mu.$$

La notation ci-dessus (couramment utilisée dans le cadre du transport optimal), signifie que, pour tout borélien $A \subset \mathbb{R}^d$,

$$\mu_t(A) = \mu_s(X_s(\cdot, t)^{-1}(A)).$$

Le fait que le flot pousse μ_s vers μ_t peut aussi s'exprimer à l'aide de la formule de changement de variable. On a en particulier, pour toute fonction régulière $\varphi \in \mathcal{D}(\mathbb{R}^d) = C_c^\infty(\mathbb{R}^d)$,

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

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

$$\begin{aligned} \int_{\mathbb{R}^d} \varphi(y) \partial_t \rho_t(y) dy &= \int_{\mathbb{R}^d} \nabla \varphi(X_s(x, t)) \cdot u_t(X_s(x, t)) \rho_s(x) dx \\ &= \int_{\mathbb{R}^d} \nabla \varphi(x) \cdot u_t(x) \rho_t(x) dx = - \int_{\mathbb{R}^d} \varphi(x) \nabla \cdot (u_t(x) \rho_t(x)) dx, \end{aligned}$$

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)

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

1. Cette approche, qui permet de passer d'une description lagrangienne à une description eulérienne est d'une certaine manière inverse de la précédente : l'équation de transport avait été introduite pour exprimer une conservation, et l'on en avait déduit des propriétés impliquant les caractéristiques associées au champ de transport. Dans cette seconde approche, on part des caractéristiques, et l'on établit des propriétés de conservation dans deux contextes : mesure images associées au flot, représentées par des densités régulières, puis plus loin (proposition 4.12) transport de mesures singulières. Cette seconde approche est d'une certaine manière plus légitime, et en tout cas formalisable rigoureusement. Nous avons néanmoins conservé l'approche basée sur le flux, malgré ses défauts, car dans d'autres contextes (et en particulier quand les particules transportées interagissent de façon complexe entre elles, comme en mécanique des fluides), elle permet d'obtenir formellement des modèles pertinents, quand l'autre approche est essentiellement inapplicable.

Remarque 4.7. L'équation (4.2) exprime par construction la conservation d'une certaine quantité représentée par une variable *extensive* (voir section 1.2), le sens physique en est clair. L'équation non conservative (4.3) lui est formellement équivalente dans le cas d'un champ à divergence nulle, mais son sens est moins clair dans le cas général. Le fait que la quantité transportée reste constante le long des caractéristiques (qui sont ici des trajectoires) permet de les interpréter comme des variables intensives, qui ne font pas l'objet de sommes entre elles. Il peut s'agir par exemple de la fraction (dans $[0, 1]$) d'une certaine substance dans un mélange, ou même de variables plus essentiellement intensives, à savoir qualitatives, comme un triplet RGB représentant une couleur, ou une caractéristique extraire d'un dictionnaire sémantique sans structure algébrique (il n'est pas nécessaire pour que le phénomène de transport ait un sens que l'espace sur lequel vit la variable soit muni d'une loi de composition interne de type '+'). On peut par exemple penser au suivis des supporters d'une certaine équipe dans un mouvement de foule au sortir d'un stade.

Remarque 4.8. 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, 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. Dans le monde réel, 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 n'importe quel champ de vitesse sous-jacent. L'approche proposée dans les propositions 4.5 et 4.6 pour construire des solutions à l'équation de transport (formes non conservative et conservative) est basée sur les caractéristiques (ou trajectoires) associées aux champs de vitesse u . Il s'agit donc d'un passage *lagrangien vers eulérien*, qui est donc bien posé, ou plus informellement "facile" conformément aux considérations de la section 1.1 (on dégrade la qualité de l'information). Dès que l'on perd la possibilité de définir de façon unique des trajectoires, i.e. (voir théorème de Cauchy-Lipchitz 11.10, page 237) dès que le champ n'est plus localement lipschitzien, cette équation de transport pose des problèmes théoriques extrêmement délicats. On pourra se reporter à l'article historique de Di Perna & Lions², 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. On pourra aussi se reporter à Ambrosio³ pour plus de détails.

Une deuxième manière de considérer cette équation de transport (nous considérons ici l'équation conservative dite de continuité) est la suivante : 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 foule, voitures, voire planètes). Il s'agit maintenant d'un passage *eulérien vers lagrangien*, qui est (nous renvoyons une nouvelle fois à la section 1.1) intrinsèquement mal posé en général. Les développements récents en théorie du transport optimal⁴ permettent de montrer la caractéristique bien posé de ce problème en termes d'existence de solutions (dans le sens généralisé de la définition 4.9 ci-après), sous des hypothèses assez générales.

Solutions faibles

L'équation de continuité introduite ci-dessus fait intervenir des dérivées en temps et en espace. On parlera de solution *classique* (ou forte) de cette équation une densité continûment différentiable en temps et en espace, pour un champ advectant lui-même C^1 en x et au moins continu en t . Le phénomène de transport d'une substance ne nécessitant aucune régularité pour avoir un sens, il est

2. 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>

3. Ambrosio & Trevisan, Lectures notes on the Di Perna-Lions theory in abstract measure spaces, <http://arxiv.org/pdf/1505.05292v1.pdf>

4. On pourra se reporter au théorème 8.3.1 dans L. Ambrosio, N. Gigli, G. Savaré, Gradient flows in metric spaces and in the space of probability measures, Lectures in Mathematics (ETH Zürich, 2005).

important de donner un sens à l'équation pour des densités (et des vitesses) moins régulières, pour lesquelles les dérivées intervenant dans l'équation ne sont pas définie au sens classique. On parlera alors de solution *faible*. Cette appellation ne couvre pas une notion précise et universelle, mais plutôt une approche permettant de définir un type de solutions avec un degré de généralisation qui dépend du contexte. Nous proposons ici une approche assez extrême, puisqu'elle permet de définir la notion de solution non seulement pour des densités peu régulières, mais même pour des mesures non absolument continues par rapport à la mesure de Lebesgue.

Commençons par quelques mots sur la démarche générale (qui en elle-même n'est pas mathématisée) permettant d'accéder à une notion généralisée de solution. On considère l'équation de continuité

$$\partial_t \rho_t + \nabla \cdot (\rho_t u_t).$$

On considère un couple (ρ_t, u_t) de fonctions régulières en espace-temps, solution de cette équation sur $\mathbb{R}^d \times [0, T]$. On considère une fonction φ elle-même régulière en espace-temps, nulle en dehors d'un compact de $\mathbb{R}^d \times]0, T[$, on multiplie l'équation par φ , on intègre sur $\mathbb{R}^d \times]0, T[$, et l'on effectue des intégrations par parties pour faire passer les dérivées sur la fonction test. On obtient

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

On peut effectuer la démarche dans l'autre sens, ce qui assure qu'un couple de fonctions régulières (ρ_t, u_t) est solution de l'équation sur $]0, T[$ si est seulement si, pour toute fonction φ régulière, l'identité ci-dessus est vérifiée. Mais l'intérêt principal de l'approche réside dans le fait que cette identité a un sens même si ρ_t et u_t ne sont pas régulières. L'idée est de se donner des conditions suffisantes sur ρ_t et u_t pour que l'identité ci-dessus ait un sens clair pour des fonctions φ régulières. On notera que ρ_t n'apparaît que sous la forme $\rho_t dx$, que l'on écrira par abus de langage $d\rho_t(x)$, qui est la mesure associée à la densité ρ_t relativement à la mesure de Lebesgue. On peut donc directement chercher l'inconnue sous la forme d'une mesure, qui n'est pas nécessairement absolument continue par rapport à la mesure de Lebesgue, tant que l'on assure que l'intégrande de l'expression ci-dessus est intégrable pour cette mesure. Ces considérations conduisent à la définition suivante.

Definition 4.9. Soient (ρ_t) une famille de mesures de Borel⁵ sur \mathbb{R}^d , (u_t) une famille de champs de vecteurs ρ_t -mesurables avec $u_t \in L^1_{\rho_t}$ pour presque tout t , telles que

$$\int_0^T \|u_t\|_{L^1_{\rho_t}} dt = \int_0^T dt \int_{\mathbb{R}^d} |u_t| d\rho_t < +\infty.$$

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

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

pour tout $\varphi \in C_c^1(\mathbb{R}^d \times]0, T[)$.

Le cadre précédent s'applique directement à des mesures définies sur un ouvert Ω . On prendra garde que, dans ce cas, l'équation ne donne aucune information sur ce qui se passe au bord de l'ouvert.

Remarque 4.10. On peut intégrer la prise en compte de conditions initiales sur ρ en considérant des fonctions tests qui ne s'annulent pas nécessairement pour $t = 0$. On pourra considérer plus précisément des fonctions φ à support compact sur $\mathbb{R}^d \times [0, T[$, et demander que

$$\int_0^T dt \int_{\mathbb{R}^d} (\partial_t \varphi + u_t \cdot \nabla \varphi) d\rho_t + \int_{\mathbb{R}^d} \varphi(x, 0) d\rho_0 = 0.$$

Remarque 4.11. On prendra garde au fait que les hypothèses sur la vitesse dépendent de la mesure ρ_t , ce qui rend difficile l'utilisation de cette définition pour montrer le caractère bien posé de problèmes du type : on se donne la vitesse u , une condition initiale ρ_0 , et l'on cherche à construire la mesure

5. Mesures boréliennes positives qui prennent une valeur finie sur tout compact de \mathbb{R}^d

ρ_t transportée, Considérer par exemple le champ u_t sur \mathbb{R} identiquement égal à 1, 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 chargerait le point 0 en question. On pourra ainsi vérifier que, pour la condition initiale $\rho_0 = \mathbf{1}_{]-1,0[}$, l'équation admet une infinité de solutions, parmi lesquelles on retrouve bien le transport à vitesse constante de la densité initiale

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

mais aussi

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

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'accumuler pour former une mesure singulière). Pour aller encore plus loin, si l'on ne fixe pas de conditions de positivité sur les mesures, on peut avoir dans certains cas une infinité de solution à l'équation de transport associée à une donnée initiale nulle (voir exercice 4.1).

Cette définition permet une description eulérienne du mouvement de particules ponctuelles, comme l'exprime la proposition suivante.

Proposition 4.12. (Écriture eulérienne du transport de masse ponctuelles)

On considère une famille de masses ponctuelles mobiles $(x_j(t))_{1 \leq j \leq N}$ (masse $m_j > 0$ en x_j), animées de vitesses $u_j(t)$ continues, de telle sorte que

$$\dot{x}_j = v_j(t) \quad j = 1, \dots, N.$$

On suppose que les particules restent toujours distinctes deux à deux. On introduit la mesure atomique

$$\rho_t = \sum_{i=1}^N m_i \delta_{x_i(t)}.$$

On note u_t la mesure vectorielle supportée par le nuage des $x_j(t)$, qui prend la valeur $v_j(t)$ en $x_j(t)$. Alors le couple (ρ_t, u_t) est solution faible de l'équation de transport (définition 4.9).

Démonstration. On vérifie tout d'abord la propriété pour un seul atome ($N = 1$) de masse 1. On considère donc une trajectoire $t \mapsto x(t)$, avec $\dot{x}(t) = v(t)$, et la densité associée ρ_t . On a

$$\begin{aligned} \int_0^T dt \int_{\mathbb{R}^d} (\partial_t \varphi + u_t \cdot \nabla \varphi) d\rho_t &= \int_0^T (\partial_t \varphi(x(t), t) + v(x(t), t) \cdot \nabla \varphi(x(t), t)) \\ &= \int_0^T \frac{d}{dt} \varphi(x(t), t) = \varphi(x(T), T) - \varphi(x(0), 0) = 0. \end{aligned}$$

On en déduit la propriété générale pour N masses, la linéarité de l'équation vis à vis de la densité permettant de sommer les N contributions. \square

4.3 Diffusion

4.3.1 Loi de Fick, équation de la chaleur

Modèle 4.13. (Loi de Fick)

On dit qu'un phénomène de propagation d'une substance, dont la densité est représentée par $\rho(x, t)$, suit la loi de Fick s'il existe un paramètre $D > 0$ tel que

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

Remarque 4.14. 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.

Équation de la chaleur On considère une substance qui diffuse dans un milieu selon la loi de Fick (modèle 4.13). L'équation de conservation (4.1) s'écrit ici

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

ou, dans le cas où D est uniforme,

$$\frac{\partial \rho}{\partial t} - D \Delta \rho = 0. \quad (4.5)$$

Diffusion non isotrope. Dans le cas où le milieu n'est pas isotrope (i.e. la diffusion est plus ou moins facile selon la direction), on peut introduire une matrice de diffusion définie positive D qui conduit à 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 D ne sera pas scalaire. Cette situation est courante dans le cas de milieux *fibreux* (une direction longitudinale très diffusive, les deux autres moins), comme le sont par exemple les muscles dans le corps humain, ou de milieux stratifiés (deux directions typiquement plus diffusives que la direction transverse).

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 ρ_{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

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

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

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

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 Dirichlet⁶ en faisant tendre β vers $+\infty$.

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.

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

$$K_y(x, t) = \frac{1}{(4\pi Dt)^{d/2}} e^{-\frac{|x-y|^2}{4Dt}}, \quad (4.6)$$

qui représente la densité de probabilité de présence d'un mouvement brownien issu de y , est solution de l'équation de la chaleur (4.5). On pourra vérifier en particulier que $K_y(\cdot, t)$ se concentre en y quand t tend vers 0^+ , plus précisément que l'on a convergence au sens des mesures (ou des distributions) vers la masse de Dirac en y :

$$\int_{\mathbb{R}^d} K_y(x, t) \varphi(x) dx \longrightarrow \varphi(y) \text{ quand } t \rightarrow 0^+.$$

On peut ainsi écrire, pour toute fonction u_0 suffisamment régulière,

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

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

Déplacement moyen d'une particule. L'écart type σ de la loi ci-dessus vérifie $2\sigma^2 = 4Dt$, soit $\sigma = \sqrt{2Dt}$. Le déplacement moyen est donc proportionnel à la racine carrée du temps, et à la racine carrée du coefficient de diffusion. Dans le cas d'une (ou plusieurs) petites particules dans un fluide qui, du fait de collision avec les molécules du fluide environnant, suit un mouvement erratique, la formule de Stokes-Einstein 5.2 (page 102) permet d'estimer le coefficient de diffusion en fonction du diamètre de la particule et de la viscosité du fluide environnant, et donc d'après ce qui précède le déplacement moyen de la particule.

Remarque 4.15. L'expression (4.6) ci-dessus correspond également à la densité de présence d'une particule brownienne issue de y à $t = 0$, et dont la position X vérifie $dX = \sigma dW_t$, où W_t est un processus de Wiener. Le coefficient de diffusion est lié à σ par $D = \sigma^2/2$, ou réciproquement $\sigma = \sqrt{2D}$.

Structure de flot de gradient

Nous présentons ici de façon très informelle comment l'équation de la chaleur peut être interprétée comme un flot de gradient pour une certaine fonctionnelle. On considère l'équation de la chaleur sur un domaine Ω borné régulier, avec condition de Neuman homogènes. On multiplie l'équation par une fonction-test v , et on intègre sur le domaine

$$\int \partial_t h = \int_{\Omega} \Delta \rho h + \int_{\Omega} f h = - \int_{\Omega} \nabla \rho \cdot \nabla h + \int_{\partial \Omega} \frac{\partial \rho}{\partial n} h + \int_{\Omega} f h = - \langle \nabla J | h \rangle,$$

où J est la fonctionnelle

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

On peut donc écrire formellement l'équation de la chaleur

$$\frac{d\rho_t}{dt} = -\nabla J(\rho_t).$$

(N.B. on utilise une dérivée en temps droite ci-dessus car $t \mapsto \rho_t$ est comme une trajectoire dans un espace fonctionnel.)

Le sens physique de la fonctionnelle ci-dessus n'est pas limpide. On peut simplement dire que, dans le cas homogène ($f = 0$), cette fonctionnelle mesure la variation locale en moyenne quadratique. L'équation exprime donc une évolution suivant la ligne de plus grande pente d'une fonctionnelle qui pénalise les variations. On retrouve l'effet régularisant de cette équation.

Décroissance de l'entropie

Il existe une autre interprétation de l'équation de la chaleur comme flot de gradient d'une fonctionnelle qui a une sens beaucoup plus clair, puisqu'il s'agit de l'entropie. Cette interprétation trouve sa source dans la théorie du transport optimal, présentée dans un cadre discret au chapitre 14. On pourra se reporter à Santambrogio⁷ pour un exposé général de ce cadre.

Nous nous contentons ici de montrer la décroissante de l'entropie pour toute solution de l'équation de la chaleur avec conditions de Neuman homogène (qui assurent la conservation de la masse totale).

On considère Ω un domaine bornée régulier de \mathbb{R}^d , et ρ une densité de probabilité définie sur Ω . On définit 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 conditions aux limites de Neuman homogènes (de façon à garder une masse unitaire constante). On a

$$\frac{d}{dt} S(\rho) = \int_{\Omega} (1 + \log \rho) \frac{\partial \rho}{\partial t} = \int_{\Omega} (1 + \log \rho) \Delta \rho = - \int_{\Omega} \frac{1}{\rho} \nabla \rho \cdot \nabla \rho + \int_{\Gamma} \frac{\partial \rho}{\partial n} (1 + \log \rho) = - \int_{\Omega} \frac{1}{\rho} |\nabla \rho|^2 \leq 0.$$

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.

Remarque 4.16. Le taux de décroissance de l'énergie est

$$\int_{\Omega} \frac{1}{\rho} |\nabla \rho|^2 = \int_{\Omega} |\nabla \log \rho|^2 \rho.$$

Cette dernière quantité est appelée *information de Fisher*⁸.

4.3.2 Cadre mathématique pour le problème de Poisson et l'équation de la chaleur

L'équation de la chaleur stationnaire, appelée problème de Poisson :

$$\begin{cases} -\Delta u &= f & \text{dans } \Omega \\ u &= 0 & \text{sur } \Gamma, \end{cases} \quad (4.7)$$

7. F. Santambrogio, Flots de gradient dans les espaces métriques et applications, Séminaire Bourbaki, 65ème année, 2012-2013, no 1065. <https://cvgmt.sns.it/media/doc/paper/2056/Exp1065.FSantambrogio2.pdf>

8. De façon très frappante, le flot de gradient associé à cette fonctionnelle d'information de Fisher, dans le cadre du transport optimal, peut être rattaché à l'équation de Schrödinger, voir <https://arxiv.org/abs/0804.4621>.

a fait l'objet d'un nombre de travaux considérable, et suscite encore une certaine activité de recherche, notamment sur les questions de régularité de la solution pour des domaines peu réguliers et des conditions aux limites panachées. L'approche la plus directe, présentée succinctement ici (voir section 12.5, page 255, pour une présentation plus détaillée) consiste à écrire le problème sous forme variationnelle, et à identifier dans cette forme un problème qui rentre dans le cadre du théorème de Riesz-Fréchet (18.17). Le problème posé sur un domaine Ω borné et régulier, avec conditions de Dirichlet homogènes, consiste à chercher u dans l'espace de Sobolev $V = H_0^1(\Omega)$ (voir chapitre 12) tel que

$$\int_{\Omega} \nabla u \cdot \nabla v = \int_{\Omega} fv \quad \forall v \in V, \quad (4.8)$$

ou, de façon abstraite, $a(u, v) = \langle \varphi, v \rangle$ pour tout $v \in V$. Les propriétés de symétrie, de continuité, et de coercivité de la forme bilinéaire en font un produit scalaire qui induit une norme équivalente à la norme de départ, et le théorème de Riesz-Fréchet (18.17) assure l'existence et l'unicité d'une solution. La formulation variationnelle ci-dessus rentre dans le cadre plus général du théorème de Lax Milgram (théorème 18.25, page 363). Dans le cas d'une forme bilinéaire symétrique, ce qui est le cas ici, la solution est également caractérisée comme minimiseur de la fonctionnelle

$$v \mapsto J(v) = \frac{1}{2} \int_{\Omega} |\nabla v|^2 - \int_{\Omega} fv.$$

Dans le cas de conditions de Dirichlet non homogènes, la solution est également définie comme minimiseur de cette même fonctionnelle sur l'espace affine des fonctions de $H^1(\Omega)$ qui vérifient la condition au bord (voir corollaire 18.26, page 364, ainsi que la section 12.2, page 249, pour des précisions sur la notion délicate de "valeur au bord" pour des fonctions de l'espace de Sobolev $H^1(\Omega)$).

On notera que l'élaboration de la formulation variationnelle s'est faite de façon informelle, et que le résultat d'existence et d'unicité porte sur cette formulation variationnelle, et non pas sur l'équation de départ. Il s'agit donc de préciser en quoi la solution construite dans l'espace de Sobolev H^1 est bien solution de l'équation de départ avec ses conditions aux limites. Pour montrer que l'équation est vérifiée dans un sens classique, il faut établir la régularité H^2 de cette solution, i.e. montrer que les dérivées seconde sont bien dans L^2 . Cette étape peut être très délicate en tout généralité. On se reportera à la section 12.5, page 256 pour des développements autour de ces questions de régularité.

Équation de la chaleur instationnaire

Le problème instationnaire peut être mis lui-même sous forme variationnelle et, du fait de l'injection compacte de H_0^1 dans L^2 (théorème 12.32, page 253) rentre dans le cadre du théorème 18.45, basé sur la décomposition spectrale de l'opérateur auto-adjoint compact $(-\Delta)^{-1}$ dans L^2 . Plus précisément, le théorème 18.41, page 368, assure l'existence d'une famille infinie (λ_k, w_k) dans $]0, +\infty[\times H_0^1(\Omega)$ de couples propres, avec λ_k qui tend vers $+\infty$, vérifiant

$$-\Delta w_k = \lambda_k w_k.$$

La famille (w_k) est une base hilbertienne de $L^2(\Omega)$, et $(w_k/\sqrt{\lambda_k})$ est une base hilbertienne (voir définition 18.35, page 366) de $H_0^1(\Omega)$.

Le théorème 18.45 donne une forme explicite de la solution comme série infinie, qui s'écrit pour le problème homogène ($f \equiv 0$)

$$u(t) = \sum_{k=1}^{+\infty} u_0^k e^{-D\lambda_k t} w_k$$

où (w_k) est la base Hilbertienne (voir théorème 18.41) des fonctions propres du Laplacien avec conditions de Dirichlet, et $0 < \lambda_1 \leq \lambda_2 \leq \dots$ les valeurs propres associées. Les u_0^k sont les coefficients de la décomposition de la condition initiale u_0 dans la base hilbertienne des w_k .

En général (sauf cas très particulier où u_0 est orthogonal au premier vecteur propre), le terme le plus lent à s'amortir est donc le premier, et le temps caractéristique correspondant est $\tau_1 = 1/D\lambda_1$, ce

qui fait donc jouer un rôle essentiel à cette première valeur propre du Laplacien, qui s'exprime aussi, selon le théorème de Courant-Fisher (théorème 18.43, page 369)

$$\lambda_1 = \inf_{u \neq 0} \frac{\int_{\Omega} |\nabla u|^2}{\int_{\Omega} u^2}. \quad (4.9)$$

Remarque 4.17. L'opérateur du Laplacien apparaît donc, dans ce contexte comme l'opérateur de divergence composé avec le gradient, deux opérateurs mutuellement adjoint. Ce fait conduit, lorsque l'on écrit la formulation variationnelle du problème de Poisson $-\Delta u = f$ sur un domaine Ω , à une forme bilinéaire symétrique :

$$\int_{\Omega} \nabla u \cdot \nabla v = \int_{\Omega} fv,$$

avec des rôles parfaitement symétriques joués par le gradient appliqué à la fonction inconnue, qui provient de la loi de Fick (ou tout autre loi constitutive du même type, comme la loi de Darcy (9.10) pour les milieux poreux par exemple), et le second gradient appliqué à la fonction test, qui apparaît comme adjoint de l'opérateur de divergence. On prendra garde au fait que les sens de ces opérateurs respectifs sont, en terme de modélisation, très différents. Le premier exprime une loi phénoménologique, qui pourrait fort bien être invalidée dans certains contextes (comme dans les cas de diffusion non-linéaire par exemple), et devoir être remplacée par un autre opérateur. La divergence à l'origine du second gradient est en revanche plus universelle, puisqu'elle exprime simplement le principe conservation de la matière, ou plus généralement de bilan de matière dans le cas où l'on a un terme source. Sa linéarité est essentielle, du fait qu'il s'agit d'un opérateur de bilan (qui somme des quantités). On se reportera au chapitre 2 (en particulier la remarque 2.15, page 46) pour des considérations analogues dans un cadre discret.

4.3.3 Interprétations stochastique du Laplacien

Nous énonçons ici quelques propriétés qui dépassent largement le cadre de ce cours, mais qui donnent une interprétation très féconde de l'opérateur laplacien, et sont les versions continues de propriétés démontrées dans le cas discret (voir section 6.2, page 128).

Proposition 4.18. On considère un domaine borné et régulier Ω dans \mathbb{R}^d , et P un champ régulier sur la frontière $\Gamma = \partial\Omega$. Pour tout $x \in \Omega$, on considère w_x^t le mouvement brownien issu de x . On note τ_x le temps de première rencontre avec la frontière $\Gamma = \partial\Omega$ par w_x^t , i.e.

$$\tau_x = \inf \{t, w_x^t \in \Gamma\}.$$

On définit p comme l'espérance de P au point de rencontre avec la frontière :

$$p(x) = \mathbb{E}(P(w_x^{\tau_x})).$$

Alors p est solution du problème de Laplace avec condition de Dirichlet P sur la frontière.

Démonstration. On se reportera à Doob⁹ pour une démonstration de cette propriété □

Une autre propriété similaire est basée sur la notion de *measure harmonique* relativement à un point.

Proposition 4.19. On considère un domaine Ω borné régulier dans \mathbb{R}^d , et $x \in \Omega$. Soit $w_{x_0}(t)$ le mouvement brownien issu de x_0 et τ_{x_0} le temps de rencontre avec $\Gamma = \partial\Omega$, comme dans la proposition précédente. Alors $Y_{x_0} = w_{x_0}^{\tau_{x_0}} \in \Gamma$ est une variable aléatoire à valeurs dans Γ , donc on appelle μ_{x_0} la

9. J. L. Doob, Classical Potential Theory and Its Probabilistic Counterpart, Grundlehren der mathematischen Wissenschaften, Springer New York, NY, 1984.

loi. Cette mesure admet une densité relativement à la mesure de Lebesgue sur Γ , qui est $-\partial q/\partial n$, où q est solution du problème de Dirichlet avec second membre singulier :

$$\begin{cases} -\Delta q &= \delta_{x_0} & \text{dans } \Omega \\ q &= 0 & \text{sur } \Gamma. \end{cases} \quad (4.10)$$

En d'autres termes, pour tout $A \subset \Gamma$ mesurable

$$\mathbb{P}(Y_{x_0} \in A) = - \int_A \frac{\partial q}{\partial n}.$$

La mesure μ_{x_0} de densité $-\partial q/\partial n$ est appelé *mesure harmonique* relativement à x_0 .

Démonstration. On trouvera une démonstration de cette propriété dans Gustafsson¹⁰. \square

4.4 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 l'équation

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

Definition 4.20. (Nombre de Péclet)

Le nombre de Péclet est défini par

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

où L représente la taille caractéristique 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 $\text{Pe} >> 1$, 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.

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 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) $t^* = t/T$ et $x^* = x/L$. On note par ailleurs $u^* = u/U$. Dans ces nouvelles variables, l'équation s'écrit

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

10. B. Gustafsson, "Lectures on Balayage", Report series, no. 7, p. 17-63, Report series 7, Sirkka-Liisa Eriksson, 2002.

Exemple 4.4.1. (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 $\eta > 0$.

Remarques additionnelles

Sur la notion de flux, sur l'équation de conservation

La définition formelle 4.1, 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 ponctuelles*¹¹, 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 étudié¹².

La notion n'a en particulier pas 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 4.21. On peut se demander quelle est la nature de l'objet mathématique qui résulterait de l'application à *la lettre* de la définition 4.1, dans le cas où l'on a un nombre fini de particules, de masses m_i et vitesses $u_i(t)$, $i = 1, \dots, N$. En dimension 1, considérons le cas d'une particule de masse m parcourant la trajectoire $t \mapsto X(t)$, animée d'une vitesse $V(t) = \dot{X}(t)$, supposée positive pour fixer les idées. 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 faiblement¹³, (ou au sens des distributions) vers $mV\delta_{X(t)}$.

11. Cette notion-même de particule ponctuelle est une idéalisation de la situation où la taille des grains considérés est petite devant les autres grandeurs caractéristiques du phénomène étudié, en particulier la distance interparticulaire.

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

13. Dans ce contexte, la convergence faible correspond à une convergence faible- \star dans le dual de $C_0(\mathbb{R})$, espace des fonctions continues qui tendent vers 0 en $\pm\infty$.

4.5 Exercices

Exercice 4.1. On considère un champ de vitesse sur \mathbb{R} égal à $-V$ sur $]-\infty, 0[$, V sur $]0, +\infty[$, et 0 en 0, avec $V > 0$. Montrer que, si l'on n'impose pas aux mesures d'être positives, il existe une infinité de solutions associées à la donnée initiale nulle.

Exercice 4.2. On considère une distribution régulière de particules sur l'axe réel se déplaçant à vitesse constante U , et l'on suppose que chaque particule porte une masse proportionnelle à la distance commune h qui les séparent. La mesure à l'instant t est donc du type *peigne de Dirac* en espace :

$$\rho_h = \sum_{k \in \mathbb{Z}} h \delta_{kh+tU}.$$

a) Écrire le flux J_h correspondant comme mesure (ou distribution) sur \mathbb{R} , et préciser le comportement de J_h quand h tends vers 0.

b) Pour h fixé maintenant, on définit par $Q_h(\varepsilon)$ la masse qui traverse le point $x = 0$ pendant l'intervalle $]0, \varepsilon]$, et le flux moyen par $J_h(\varepsilon) = Q_h(\varepsilon)/\varepsilon$. Étudier la limite de $J_h(\varepsilon)$ quand ε et h tendent vers 0, selon la manière dont le couple (n, ε) tend vers 0. (On pourra en particulier proposer une condition suffisante pour que la limite soit le flux $1 \times U$ attendu, et donner des exemples pour lesquels on converge vers une autre valeur.)

Exercice 4.3. On se place sur l'intervalle $I =]0, 1[$, et l'on définit un champ de vitesse u par $u(x) = x(1-x)$. Décrire le comportement quand t tend vers $+\infty$ de $\Phi(\cdot, t)$ et $\Psi(\cdot, t)$, solutions respectives sur $I \times [0, +\infty[$, de

$$\frac{\partial \Phi}{\partial t} + \frac{\partial(u\Phi)}{\partial x} = 0,$$

et

$$\frac{\partial \Psi}{\partial t} + u \frac{\partial \Psi}{\partial x} = 0,$$

pour une même condition initiale $\Phi(x, 0) = \Psi(x, 0) = \varphi(x)$, où φ est une fonction continue de I dans \mathbb{R}_+ , non nulle en 0.

Exercice 4.4. On cherche à modéliser la dynamique d'une population de parasites dans un jardin. On suppose que ces parasites diffusent selon la loi de Fick, et se reproduisent à un taux c , de telle sorte que la dynamique est décrite par l'équation

$$\partial_t \rho - a \Delta \rho = c \rho.$$

On représente le jardin par un domaine Ω , et l'on suppose les alentours du jardin hostiles, de telle sorte que ρ est pris nul sur $\partial\Omega$.

1) Montrer que, selon les cas, on peut avoir extinction de la population ou croissance exponentielle de celle-ci.

2) On dispose d'un produit contre ces parasites, et l'on modélise son action de la façon suivante : lorsque l'on épand ce produit sur un domaine ω , on a disparition complète et définitive des parasites sur ω . Supposant que l'on dispose d'une quantité de produit permettant de traiter une surface $S < |\Omega|$, comment choisir le domaine ω à traiter pour assurer l'éradication des parasites ?

Exercice 4.5. (Temps de cuisson d'un rôti)

Montrer que, à forme donnée, le temps de cuisson d'un rôti (supposé borné et connexe) est proportionnel à la puissance $2/3$ de sa masse. Donner des conditions suffisantes assurant que l'on puisse cuire un rôti non borné en un temps fini.

Généraliser à la cuisson d'un hyper-rôti dans \mathbb{R}^d , $d \in \mathbb{N}$ quelconque.

Exercice 4.6. (Décroissance de l'entropie relative à la mesure stationnaire pour l'équation de diffusion / transport par un champ de gradient)

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 :

$$\frac{\partial \rho}{\partial t} - D\Delta\rho + \nabla \cdot (\rho u) = 0, \quad u = -\nabla\Psi, \quad (4.11)$$

dans un domaine Ω borné. On suppose que $u \cdot n = -\partial\Psi/\partial n = 0$, et l'on considère des conditions de neumann homogènes sur ρ :

$$\partial\rho/\partial n = 0.$$

On se donne une donnée initiale ρ_0 qui est la densité (supposée lisse) d'un mesure de probabilité sur Ω , et l'on considère une solution $t \mapsto \rho(\cdot, t)$ associée à cette solution initiale.

- 1) Montrer que l'intégrale de $\rho(\cdot, t)$ reste égale à 1 pour tout t .
- 2) On note $\pi = e^{-\Psi/D}/Z$, où Z est une constante de normalisation assurant que π est la densité d'une mesure de probabilité. Montrer que l'équation (4.11) s'écrit

$$\frac{\partial \rho}{\partial t} - D\nabla \cdot \rho \left(\nabla \log \left(\frac{\rho}{\pi} \right) \right) = 0,$$

et en déduire que π est solution stationnaire de l'équation.

- 3) On définit l'entropie relative d'une densité de probabilité ρ par rapport à π comme

$$S(\rho|\pi) = \int \rho \log \left(\frac{\rho}{\pi} \right). \quad (4.12)$$

Montrer que $S(\rho(\cdot, t)|\pi)$ décroît au cours du temps, et que cette décroissance est stricte tant que ρ n'est pas égale à π .

Exercice 4.7. On considère le modèle monodimensionnel d'un mur d'épaisseur L conducteur de la chaleur, que l'on suppose constitué de deux matériaux de conductivités thermiques différentes. On se place ainsi sur l'intervalle $]0, L[$, et l'on suppose que les matériaux 1 et 2, de conductivités respectives a_1 et a_2 , occupent respectivement les intervalles $]0, \alpha L[$ et $]\alpha L, L[$. Aux extrémités de l'intervalle on impose des températures T_1 et T_2 .

- 1) Écrire la relation de saut exprimant le bilan de chaleur en αL , et préciser le champ de température solution de l'équation de la chaleur stationnaire.
- 2) On note J le flux de chaleur passant à travers le mur (compté positivement dans le sens gauche vers droite). Exprimer la conductivité équivalente du mur, c'est à dire le nombre a tel que

$$J = -a \frac{T_2 - T_1}{L}.$$

- 3) On considère maintenant un matériau d'épaisseur L obtenu par juxtaposition alternée de N couches de chacun des matériaux précédents, avec la même proportion : matériau 1 sur $]0, \alpha L/N[$, matériau 2 sur $]\alpha L/N, L/N[$, etc... Préciser la conductivité équivalente du matériau composite.

- 4) Généraliser à un nombre quelconque de matériaux.

Chapitre 5

Modèles en vrac

Sommaire

5.1	Bilan radiatif	99
5.2	Montée / descente d'une rame de RER un jour de grève	101
5.3	Aérosols & concentration de CO ₂	102
5.4	Qualité de l'air dans un bâtiment	103
5.4.1	Modèle mono-compartiment	103
5.4.2	Modèles multi-compartiment	105
5.5	Inertie thermique & capacité thermique apparente	108
5.6	Modélisation de la survie d'un dialecte au sein d'une population	111
5.7	Positionnement de postes de secours	111
5.8	Affluence au supermarché	113
5.9	Densités de population et contacts induits	114
5.10	Modèles de mobilité sur réseaux	115
5.11	Cadre général, problématiques	116
5.12	Exercices	119
5.13	Mobilité individuelle	121

5.1 Bilan radiatif

On se propose d'étudier un modèle simplifié de bilan radiatif de la terre, basé sur l'écriture d'un équilibre entre l'énergie solaire reçue par la terre et l'énergie ré-émise par rayonnement, selon la loi de Stefan-Boltzman. On note Q le flux de rayonnement solaire reçu en moyenne par unité de surface sur terre, $Q \approx 341 \text{ W m}^{-2}$. On considère qu'une fraction $A \in [0, 1]$ de cette énergie est immédiatement réfléchie, où A , appelé *albedo*, est autour de 0.3. L'énergie émise en moyenne par unité de surface par la terre s'écrit σT^4 , où T est la température moyenne (exprimée en Kelvin), et $\sigma = 5.67 \times 10^{-8} \text{ W m}^{-2} \text{ K}^{-4}$. On considère qu'une fraction de cette énergie n'est pas rayonnée vers l'espace, du fait de l'effet de serre. On note $S \in [0, 1]$ la fraction d'énergie qui n'est pas évacuée vers l'espace. Ce paramètre est estimé à $S = 0.4$. En supposant que l'on est à l'équilibre, on écrit le bilan entre les énergies reçue et émises :

$$\sigma T^4(1 - S) = (1 - A)Q.$$

Avec les valeurs de référence données, on trouve

$$T_0 \approx 289 \text{ K} = 16^\circ\text{C}.$$

La valeur correspondant à un effet de serre nul ($S = 0$) est de -18°C .

Le modèle d'évolution le plus sommaire est basé sur l'hypothèse qu'à chaque instant l'énergie totale de la terre est produit de la température moyenne et d'une constante C , qui représente la capacité thermique globale de la terre. Pour se conformer au cadre ci-dessus, nous désignerons par C la capacité thermique globale par mètre carré, i.e. C multiplié par la surface de la terre est l'énergie qu'il faut pour faire monter d'un degré la température moyenne, exprimé en $\text{JK}^{-1}\text{m}^{-2}$.

On obtient le modèle suivant :

$$C \frac{dT}{dt} = f(T) = (1 - A)Q - (1 - S)\sigma T^4.$$

En supposant que tous les paramètres sont constants, fixés aux valeurs indiquées ci-dessus, on a un point fixe d'équilibre unique $T_{eq} \approx 16^\circ\text{C}$. On a

$$f'(T_{eq}) = -4(1 - S)\sigma T_{eq}^3$$

qui est strictement négatif, ce point d'équilibre est donc stable.

L'estimation du paramètre C est très délicate, car la température est fortement non uniforme au sein de la planète, et la variations de température au sein des différentes zones induites par l'excédent ou déficit de chaleur en surface se font à des échelles de temps très différentes. La difficulté vient du fait que la notion même de capacité thermique globale, pour un système au sein duquel la propagation de la chaleur est régie par des mécanismes complexes, n'a de sens que pour un temps caractéristique de variation fixé (voir section 5.5). Nous nous baserons¹ sur les deux hypothèses suivantes :

1. la chaleur (tout du moins sa part variable) est essentiellement stockée dans les océans ;
2. la partie thermiquement réactive des océans est limitée à une certaine profondeur.

Considérons dans un premier temps que toute la masse d'eau est impliquée dans les variations de la température moyenne. Ces océans totalisent une masse $M = 1.4 \times 10^{21} \text{ kg}$, la capacité calorifique de l'eau étant de $\approx 4200 \text{ JK}^{-1}\text{kg}^{-1}$, on trouve une capacité calorifique globale ramenée au mètre carré de

$$C = 1.4 \times 10^{21} \text{ kg} \times 4200 \text{ JK}^{-1}\text{kg}^{-1} / 510 \times 10^{12} \text{ m}^2 \approx 11.5 \times 10^9 \text{ J K}^{-1} \text{ m}^{-2}.$$

Si l'on fait l'hypothèse (suivant le document cité en note de bas de page) que la partie réactive correspond à une profondeur de 100 m, on trouve directement

$$C' = 4200 \text{ JK}^{-1}\text{kg}^{-1} \times 10^3 \times \text{kg m}^{-3} \times 100 \text{ m} = 4.2 \times 10^8 \text{ J K}^{-1} \text{ m}^{-2},$$

soit 27 fois moins que l'estimation précédente.

Le temps caractéristique de retour à l'équilibre pour l'estimation haute est

$$\tau = \frac{C}{4(1 - S)\sigma T_{eq}^3} \approx \frac{11.5 \times 10^9}{4 \times (1 - 0.4) \times 5.67 \times 10^{-8} \times 289^3} \approx 3.5 \times 10^9 \text{ s} \approx 110 \text{ années},$$

soit $\tau' = 4$ ans pour l'estimation basse C' .

L'albedo, quantifié par la variable A dans ce qui précède, dépend de multiples paramètres, dont la part de surface terrestre recouverte de glace (très réfléchissante, alors que la mer l'est assez peu). Il est donc naturel de supposer que A dépend lui-même de la température. En supposant que T_{eq} correspond bien à un point d'équilibre, on écrit

$$A = A_0 - \beta(T - T_{eq}),$$

avec $\beta > 0$. On obtient une équation de la forme $\dot{T} = f(T)$, avec

$$Cf(T) = (1 - A)Q - (1 - S)\sigma T^4 + \beta Q(T - T_{eq}),$$

dont T_{eq} est toujours un point d'équilibre. La stabilité de ce point d'équilibre dépend du signe de

$$Cf'(T_{eq}) = -4(1 - S)\sigma T_{eq}^3 + \beta Q,$$

qui reste devient positif (instabilité) dès que

$$\beta > \beta_c = \frac{1}{Q} 4(1 - S)\sigma T_{eq}^3 \approx 0.015 \text{ K}^{-1}.$$

¹ Voir par exemple : http://www.atmos.albany.edu/facstaff/brose/classes/ATM623_Spring2015/Notes/index.html

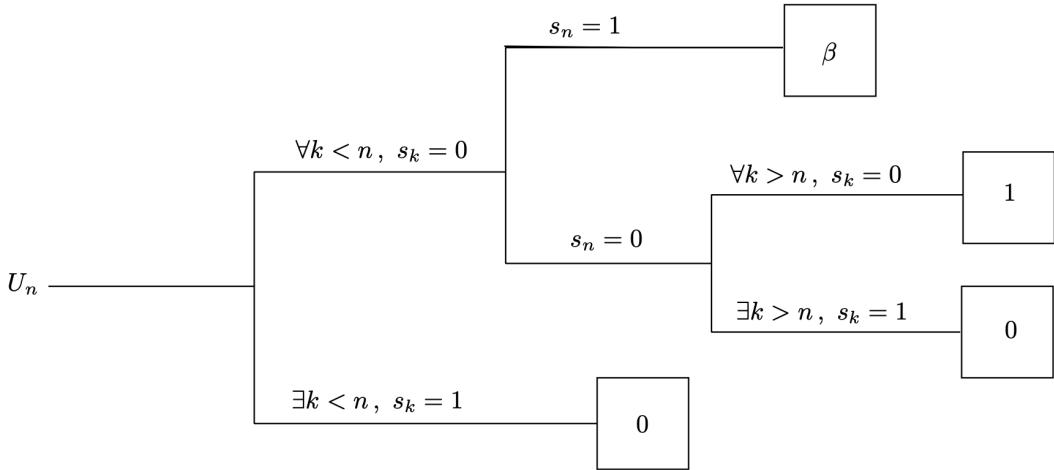


FIGURE 5.1 – Calcul du gain U_n

5.2 Montée / descente d'une rame de RER un jour de grève

On considère la situation suivante : à l'instant initial, des voyageurs sont dans une rame de RER, et d'autres attendent sur le quai. Lorsque les premiers commencent à descendre, il n'est pas rare que deux voyageurs sur le quai se positionnent de part et d'autre du passage, pour garantir leur accès au wagon, mais au prix d'une réduction drastique du flux sortant, qui peut aboutir à ce que certaines personnes ne puissent pas montrer dans le wagon.

On cherche ici à modéliser cette situation, c'est à dire à montrer que ce comportement globalement contre productif peut être retrouvé sous des hypothèses de comportement assez simple. On considérera pour simplifier qu'un seul voyageur est susceptible de se positionner en position bloquante. On modélise simplement les voyageurs sur le quai, numéroté selon leur proximité à la porte

$$n = 1 \text{ (le plus près de l'accès)}, 2, 3, \dots, N.$$

On considère que les agents accèdent dans leur ordre de numérotation à la porte. Si la place n'est pas déjà prise, le joueur n a le choix entre deux stratégies, que l'on encode par $s_n \in \{0, 1\}$.

$s_n = 0$: il reste en retrait (comportement civique)

$s_n = 1$: il se positionne en position bloquante (comportement égoïste)

On considère que tout le monde se comporte de façon civique, l'ensemble des agents bénéficiera d'une utilité égale à 1 (tout le monde peut monter dans le train). Si l'un des agents choisit 1, alors tous les autres ont une utilité nulle (ce qui correspond à un risque important de ne pas pouvoir monter dans le train). Celui qui choisit 1 a une utilité β strictement positive (son accès au train est assuré), mais potentiellement inférieure à 1 (sentiment de culpabilité, risque de conflit avec les autres voyageurs, ...). L'arbre des possibilités pour le joueur n est représenté sur la figure 5.1.

Focalisons nous maintenant sur le joueur 1, dont on suppose qu'il a un comportement rationnel : il tend à maximiser l'espérance de son utilité, tout en considérant que chacun des autres voyageurs est soit civique (il joue 0), soit incivique (il joue 1 si c'est possible). Le joueur 1 pense qu'un individu pris au hasard a la probabilité $p \in [0, 1]$ d'être civique.

On peut calculer l'espérance du gain du joueur en fonction de la stratégie choisie : si $s_1 = 1$, alors $U_1 = \beta$, si $s_1 = 0$, $\mathbb{E}(U_1) = p^{N-1}$. Le joueur 1 n'a donc "intérêt" à laisser la place que si $p^{N-1} > \beta$.

Diamètre	1 µm	5 µm	50 µm	200 µm
Vitesse	30 µms ⁻¹	0.8 mms ⁻¹	8 cms ⁻¹	1.3 ms ⁻¹
Temps chute 1 m	8h 20min	20 min	12 s	0.76 s

TABLE 5.1 – Vitesses de sédimentation et temps de chute (1 m)

5.3 Aérosols & concentration de CO₂

Nous nous intéressons ici au mouvement de gouttelettes²de liquide (typiquement de l'eau) dans un gaz (typiquement de l'air).

Vitesse de sédimentation

On considère une particule sphérique (rayon a , densité ρ) en suspension dans un fluide visqueux de viscosité μ . Le poids de la particule est

$$\frac{4}{3}\pi a^3 \rho g.$$

La force exercée par le fluide environnant sur la particule est proportionnelle à la vitesse, avec un coefficient de proportionnalité donné par la loi de Fáxen (9.25, page 221), égal à $6\pi\mu a$. La vitesse de sédimentation s'obtient en écrivant l'équilibre des forces

$$\frac{4}{3}\pi a^3 \rho g = 6\pi\mu a U_s,$$

d'où

$$U_s = \frac{1}{18} \frac{\rho g}{\mu} d^2, \quad (5.1)$$

où d est le diamètre de la particule.

Cette formule simple permet de comprendre le danger potentiel de petites gouttelettes en suspension dans l'air. A titre d'exemple, pour une particule de liquide de type eau (densité $\rho \approx 1000 \text{ kg m}^{-3}$) de diamètre 1 µm, la table 5.3 représente les vitesses de sédimentation et le temps mis pour chuter d'une hauteur de 1 mètre. Il apparaît clairement que les petites particules de l'ordre du micron vont flotter dans l'air un temps important avant de potentiellement sédimerter sur le sol.

Diffusion

Les petites particules au sein d'un gaz sont soumises au choc avec les particules de gaz environnant, ce qui entraîne un phénomène de diffusion, quantifié par un coefficient de diffusion donné par la loi de Stokes-Einstein

$$D = \frac{k_B T}{6\pi\mu a}, \quad (5.2)$$

où a est le rayon de la particule. A titre d'exemple, le coefficient de diffusion pour une particule de diamètre 1 micron, dans l'air à température ambiante, est

$$D = \frac{k_B T}{6\pi\mu a} \approx \frac{1.4 \times 10^{-23} \times 300}{6 \times 3.14 \times (2 \times 10^{-5}) \times (0.5 \times 10^{-6})} \approx 2.2 \times 10^{-11} \text{ m}^2 \text{s}^{-1}.$$

2. Une toute petite quantité d'un liquide comme de l'eau en suspension garde une forme sphérique du fait de la tension surfacique.

2. Les particules comme des particules de diesel présentent un danger lorsqu'elles sont inhalées, et a fortiori lorsqu'elles sont susceptibles de se déposer en profondeur dans l'appareil respiratoire.

Particule dans un fluide en mouvement

On considère une particule isolée au sein d'un fluide en mouvement. Par isolée nous entendons que la particule voit le fluide environnant comme un milieu fluide infini, animé d'une vitesse constante qui est la vitesse locale $u(x)$. Cette vision repose sur plusieurs échelles : l'échelle microscopique correspond au diamètre d de la particule, l'échelle macroscopique L est la taille du domaine global d'intérêt (par exemple une pièce d'habitation). L'échelle mésoscopique η correspond au voisinage de la particule, significativement plus grand que d , mais suffisamment petit pour la vitesse du fluide puisse y être considérée comme uniforme, de telle sorte que la particule "voit" ce η -voisinage comme un milieu infini entraîné à vitesse constante, de telle sorte que l'on puisse utiliser la loi de Fàxen pour exprimer la force exercée sur la particule en fonction de sa vitesse relative par rapport au fluide. Si m est la masse de la particule, et $x(t)$ sa position au cours du temps, l'équation du mouvement s'écrit.

$$m\ddot{x} = \beta(u(x) - \dot{x}), \quad \beta = 6\pi\mu a, \quad (5.3)$$

où a est le rayon, et $u(x)$ la vitesse du fluide au voisinage de la particule.

Nombre de Stokes

Pour une particule en suspension dans un fluide visqueux, l'importance relative de l'inertie et des forces de frottement visqueux est caractérisé par un nombre sans dimension, le *nombre de Stokes*. Considérons une particule lancée à vitesse v_0 dans un fluide visqueux immobile $u \equiv 0$. On a

$$m\dot{v} = -6\pi\mu av \implies v(t) = v_0 \exp\left(-\frac{6\pi\mu a}{m}t\right),$$

d'où une relaxation exponentielle de la vitesse vers 0 en un temps caractéristique $\tau_r = m/(6\pi\mu a)$, qui s'écrit aussi

$$\tau_r = \frac{2\rho a^2}{9\mu},$$

où ρ est la densité de la particule. Considérons maintenant une particule dans un fluide en mouvement, et suivons sa trajectoire pendant le temps τ_r . Si la vitesse du fluide environnant n'a pas significativement changé durant τ_r , celle de la particule aura relaxé vers cette vitesse. A contrario, si la vitesse a significativement changé, les vitesses du fluide et de la particule seront différentes. La distance parcourue s'écrit $\tau_r U$, où U est l'ordre de grandeur de la vitesse du fluide. Si l'on note L l'échelle en espace correspondant à des changements significatifs de vitesses du fluide, les deux régimes ci-dessus peuvent être distingués par un nombre sans dimension, appelé *nombre de Stokes*, qui est le rapport entre le temps caractéristique de relaxation et le temps mis par une particule pour passer d'une zone à une zone où la vitesse environnante est significativement différente.

Definition 5.1. (Nombre de Stokes)

Pour une particule sphérique de densité ρ et de rayon a , en suspension dans un fluide visqueux, le nombre de Stokes est défini comme

$$\text{St} = \frac{2\rho a^2 U}{9L\mu}.$$

5.4 Qualité de l'air dans un bâtiment

5.4.1 Modèle mono-compartiment

On cherche ici à construire un modèle permettant de simuler les variations du taux de CO₂ au cours du temps, dans une pièce unique que l'on suppose échanger de l'air seulement avec le monde extérieur. On note $c = c(t) \in [0, 1]$ la fraction de molécules de CO₂ dans l'air ambiant (supposée homogène dans la pièce), au temps t . Si l'on estime les quantités de gaz sur la base du volume équivalent aux

conditions normales de température et de pression, la quantité contenue dans la pièce à un instant donné s'écrit $c(t)V$, où V est le volume de la pièce³.

On note n le nombre de personnes dans la pièce, et l'on suppose que chacune de ces personnes produit du CO₂ à un taux F . On suppose que l'air dans la pièce se renouvelle régulièrement par de l'air extérieur, et l'on note sous la forme RV ce taux (en m³ h⁻¹). Cela signifie que de l'air extérieur entre régulièrement dans la pièce au taux ci-dessus, et que de l'air intérieur sort de la pièce au même taux. L'équation décrivant l'évolution de la quantité totale de CO₂ dans la pièce s'écrit, sous ces hypothèses

$$\begin{aligned}\frac{d(Vc)}{dt} &= RV(c_{ext} - c) + nF \\ \frac{dc}{dt} &= R(c_{ext} - c) = \frac{nF}{V}\end{aligned}$$

Si aucun des paramètres ne varie, on a un point d'équilibre correspondant à un plateau :

$$c^\infty = c_{ext} + \frac{nF}{RV}, \quad (5.4)$$

qui est asymptotiquement stable, et l'on a même une solution explicite de cette équation comme combinaison barycentrique de la condition initiale et de l'état d'équilibre

$$c(t) = e^{-Rt}c^{init} + (1 - e^{-Rt})c^\infty.$$

Dans le cas où le nombre de personnes varie au cours du temps, ou si la production de CO₂ individuelle varie (exercice physique, ou au contraire repos) au cours du temps⁴, la méthode de variation de la constante permet d'obtenir l'expression intégrale suivante :

$$\begin{aligned}c(t) &= e^{-Rt}c^{init} + (1 - e^{-Rt})c_{ext} + R \int_0^t e^{-R(t-s)} \frac{n(s)F(s)}{RV} ds \\ &= e^{-Rt}c^{init} + (1 - e^{-Rt})c_{ext} + \frac{1}{V} \int_0^t e^{-R(t-s)} n(s)F(s) ds.\end{aligned} \quad (5.5)$$

Le paramètre R , qui encode le mode d'utilisation de la salle (fenêtre / porte ouverte ou fermée, VMC active ou pas, ...), est susceptible de changer au cours du temps. Si ce paramètre est constant par morceaux, on obtiendra la solution globale en raboustant par morceaux des solutions données par l'expression (5.5).

Ordres de grandeur.

Dans un contexte épidémiologique chargé, on utilise le CO₂ comme marqueur indirect de la concentration d'aérosols potentiellement contaminants. Dans cette optique, estime en général à 800 ppm le taux en-dessous duquel le risque de contamination est faible. Les mesures dans les locaux d'habitation conduisent à des valeurs très dispersées, entre 600 et 3000 ppm, selon la qualité de la ventilation.

Un taux de brassage R courant pour un local d'habitation sans VMC, porte et fenêtres fermées, est de l'ordre de 0.5 h⁻¹. Pour une salle d'opération dans un hôpital, les normes prescrivent un taux au delà de 15 h⁻¹.

La production de CO₂ pour un adulte au repos, exprimé en volume équivalent aux conditions normales de température et de pression, est de l'ordre de 20 Lh⁻¹.

Estimation du coût énergétique de la ventilation. On cherche ici à estimer le coût énergétique de la ventilation, plus précisément la puissance qu'il faut fournir pour chauffer l'air extérieur que l'on fait rentrer pour compenser la production de CO₂ de personnes dans une pièce. Pour fixer les idées,

3. Si l'on a par exemple une pièce de 100m³ vide d'occupants, à une fraction de CO₂ correspondant à celle du monde extérieur, soit 400 ppm = 400 10⁻⁶ = 0.0004 = 0.04 %

4. On prendra garde au fait que le modèle s'appuie sur un hypothèse d'homogénéité du CO₂ dans chaque pièce. En conséquence, il n'est pas destiné à prendre en compte de façon fine les variations brusques de certains paramètres.

on considère que l'on vise une valeur-cible de $c_p = 800$ ppm. D'après la relation (5.4), le débit d'air entrant nécessaire au maintien d'un taux de 800, par personne, est

$$RV = \frac{F}{c_p - c_{ext}},$$

La capacité calorifique de l'air étant de $C_a = 1.25 \text{ kJm}^{-3} \text{ K}^{-1}$ on obtient, pour une différence de température entre l'extérieur et l'intérieur de $\Delta T = 10^\circ\text{C}$, de

$$\mathcal{P} = C_a \times \frac{F}{c_p - c_{ext}} \Delta T = 1.25 \text{ kJm}^{-3} \text{ K}^{-1} \frac{0.02 \times 3600^{-1} \text{ Ls}^{-1}}{600 10^{-6}} \times 10\text{K} \approx 115 \text{ W}.$$

Noter que, pour un taux-cible de 1000 ppm, on trouve les 2/3 de cette valeur, c'est à dire 115 W, qui correspond approximativement à la production de chaleur d'une personne au repos, et donc l'énergie que sa présence permet a priori d'économiser en chauffage. Le maintien d'un taux raisonnable de CO₂ (disons autour de 1000) se fait donc d'une certaine manière, si le processus est contrôlé avec précision, à *coût nul*.

On notera aussi la dépendance de cette puissance vis à vis du taux cible \bar{c} , en $1/(\bar{c} - c_{ext})$, de telle sorte que le coût croît fortement lorsque l'on cherche à maintenir un taux proche du taux extérieur. Ainsi maintenir 500 ppm consiste à diviser par 4 l'écart à c_{ext} par rapport aux 800 de l'exemple ci-dessus, et donc à multiplier par 4 le coût énergétique de 174 W.

Estimation effective du taux de renouvellement

Estimation R à partir de la valeur plateau, n et V étant supposés connus.

Si, dans des conditions données (supposées fixes), on connaît la valeur limite c_p , la valeur extérieure c_{ext} , le nombre de personnes présentes n , et le volume V de la pièce, et le flux de CO₂ produit par personne, on peut estimer R à l'aide de l'expression

$$c_p - c_{ext} = \frac{nF}{RV} \implies R = \frac{nF}{V(c_p - c_{ext})}.$$

On prendra garde au fait que les concentrations sont exprimées comme des fractions (400 ppm correspond par exemple à 0.0004), et que la production est exprimée en m³ par heure, i.e. $F = 0.02 \text{ m}^3 \text{ h}^{-1}$.

A titre d'exemple, considérons le cas (réel) d'une salle de travaux dirigés d'un volume $V = 165 \text{ m}^3$, avec 27 personnes présentes, émettant chacune une quantité de CO₂ égale à $F = 0.02 \text{ m}^3 \text{ h}^{-1}$, avec un écart au taux extérieur de 1200 ppm = 0.0012, on obtient

$$R = \frac{nF}{V(c_p - c_{ext})} = \frac{27 \times 0.02}{165 \times 0.0012} \approx 2.7 \text{ h}^{-1}.$$

Application à la vie réelle

Si l'on connaît le volume V d'une salle, et que l'on dispose du taux de renouvellement R , on peut estimer le nombre de personnes à ne pas dépasser pour rester en dessous d'un seuil fixé c_{seuil} . Par exemple pour $V = 165 \text{ m}^3$, $R = 3 \text{ h}^{-1}$, et un seuil de 800 ppm, on trouve

$$n = \frac{RV(c_{seuil} - c_{ext})}{F} = \frac{3 \times 165 \times 400 \times 10^{-6}}{0.02} \approx 10.$$

5.4.2 Modèles multi-compartiment

Pour modéliser la situation d'un bâtiment comprenant plusieurs espaces (salles de réunion, de loisir, couloirs, halls, ...), on associe à chaque espace un sommet d'un graphe, et l'on définit simplement les

arêtes en prescrivant que $(x, y) \in E$ si x et y sont en contact, ou plus précisément sont susceptibles d'échanger directement de l'air. On suppose connu, pour tout $(x, y) \in E$, le débit d'air $\tilde{u}_{xy} \geq 0$ de x vers y . On prendra garde au fait qu'il ne s'agit pas ici d'un débit algébrique, qui représenterait le bilan des flux entre x et y , mais bien de la quantité d'air qui va de x vers y . Le débit dans l'autre sens \tilde{u}_{yx} est lui aussi un nombre positif. En termes de suivi de quantité de matière, il pourrait être tentant d'introduire la différence des deux débits, mais il est ici important de conserver ces deux quantités indépendamment, car les concentrations en x et en y sont susceptibles d'être différentes. On peut avoir $\tilde{u}_{xy} = \tilde{u}_{yx}$ comme dans le cas d'une pièce unique en interaction avec le monde extérieur, ou des valeurs différentes si l'on a des pièces en série.

On représente le monde extérieur par l'un des sommets (auquel toutes les pièces bordant l'immeuble vont donc être connectées), que nous noterons $o \in V$, et nous noterons $\mathring{V} = V \setminus \{o\}$ l'ensemble des points "intérieurs". On note enfin v_x le volume de la pièce x (le volume du monde extérieur n'a pas à être défini, comme nous le verrons). Le système s'écrit

$$\frac{d(v_x c_x)}{dt} = - \sum_{y \sim x} (\tilde{u}_{xy} c_x - \tilde{u}_{yx} c_y) + n_x F \quad , \quad \forall x \in \mathring{V}. \quad (5.6)$$

$$\sum_{y \sim x} \tilde{u}_{xy} - \sum_{y \sim x} \tilde{u}_{yx} = 0.$$

Il faut admettre que ce système est difficilement utilisable en pratique, car il nécessite la connaissance de l'ensemble des flux entre les pièces (et pas seulement des bilans de flux entre elles), qui peuvent être très délicats à mesurer, et variables en temps.

Il peut être pertinent de séparer dans les flux la partie échange symétrique, qui va conduire à un phénomène de diffusion sur le modèle discret, de la partie signée, qui correspondra à un transport. Plus précisément, on introduit

$$Q_{xy} = \min(\tilde{U}_{xy}, \tilde{U}_{yx}), \quad U_{xy} = \tilde{U}_{xy} - Q_{xy}, \quad u_{yx} = \tilde{U}_{xy} - Q_{xy}.$$

On obtient l'équation

$$\frac{d(v_x c_x)}{dt} = - \sum_{y \sim x} Q_{xy} (c_x - c_y) - \sum_{y \sim x} (U_{xy} c_x - U_{yx} c_y) + n_x F \quad , \quad \forall x \in \mathring{V}. \quad (5.7)$$

Ce modèle fait apparaître un opérateur de type laplacien discret

$$c \in \mathbb{R}^V \longmapsto Lc, \quad Lc(x) = \sum_{y \sim x} Q_{xy} (c_x - c_y),$$

un opérateur de type transport

$$- \sum_{y \sim x} (U_{xy} c_x - U_{yx} c_y),$$

comme on peut s'en convaincre en considérant une rangée de pièces $1, 2, \dots, N$ en série, transversées par un courant d'air ($u_{j,j+1} = U > 0$ et $U_{j-1,j} = 0$, pour tout j), et un terme source (production de CO₂).

La partie "courant d'air" des termes d'échange peut être modélisée en considérant que le flux d'air d'une pièce à l'autre est proportionnelle au saut de pression entre les deux pièces (on suppose la pression uniforme dans chaque pièce) :

$$U_{xy} = C_{xy}(p_x - p_y)_+, \quad U_{yx} = C_{xy}(p_y - p_x)_+.$$

L'une des quantités ci-dessus est nulle, par définition, et $C_{xy} > 0$, champ symétrique défini sur les arêtes du graphe, encode la perméabilité de l'interface entre x et y . Il sera d'autant plus grand qu'il existe des ouvertures (bordure des portes, trous dans la serrure) importantes sur la paroi délimitant les pièces.

On obtient donc, couplé à l'équation d'évolution discrète de type transport diffusion, un problème de Darcy discret posé sur un réseau de même topologie, avec ces nouvelles caractéristiques afférentes aux arêtes de type perméabilité. Il est naturel ici de revenir à la notation du chapitre 2, en introduisant le champ $u = (u_{xy}) \in \mathbb{R}^E$ antisymétrique défini sur les arêtes, relié à (U_{xy}) par

$$U_{xy} = (u_{xy})_+, \quad U_{yx} = (u_{yx})_+.$$

Le problème de Darcy s'écrit alors

$$\begin{cases} u_{xy} + c_{xy}(p_y - p_x) &= 0 \quad \forall (x, y) \in E, \\ \sum_{y \sim x} u_{yx} &= 0 \quad \forall x \in \mathring{V} \end{cases} \quad (5.8)$$

ou, avec les opérateurs discrets ∂ et ∂^* ,

$$\begin{cases} u + c\partial^* p &= 0 \quad \text{sur } E \\ \partial u &= 0 \quad \text{sur } \mathring{V}. \end{cases} \quad (5.9)$$

La question des conditions aux limites est délicate. Les courants d'air au sein d'un bâtiment ou d'une maison (c'est bien de ça qu'il s'agit ici) sont induit par des différences de pressions à l'extérieur, au niveau des différentes fenêtres du bâtiment, elles même induites par des écoulements de l'air autour du bâtiment. Il est donc nécessaire de considérer des conditions aux limites d'une nature différente de celles envisagées pour le CO₂ (avec une même valeur imposée à l'extérieur). On peut considérer par exemple qu'une valeur de pression est imposée à l'extérieur de chaque pièce en contact avec l'extérieur, avec des valeurs différentes selon les pièces. On peut rajouter des points extérieurs en lesquels on impose un condition de Dirichlet, ou imposer une condition de type Robin au niveau des pièces elles mêmes (il n'y a alors aucun sommet qui corresponde au monde extérieur).

Compléments

Estimation directe de R

On se place dans une phase où les paramètres n , V et R sont fixes, de telle sorte que l'évolution est donnée (on note t_0 l'instant initial)

$$c(t) - c(t_0) = (c^\infty - c(t_0))(1 - e^{-Rt}).$$

On suppose que l'on dispose de 3 valeurs c_0 , c_1 et c_2 à des temps $t_0 < t_1 < t_2$, respectivement. On suppose pour fixer les idées que l'on est dans une situation où c est croissante, on a donc une relaxation vers $c^\infty > c_2 > c_1 > c_0$. La fonction $t \mapsto c(t)$ est, dans ce cas, concave.

En supposant que l'évolution suit exactement le modèle ci-dessus, on écrit, pour $i = 1, 2$,

$$c_i - c_0 = (c^\infty - c_0)(1 - e^{-R(t_i - t_0)}).$$

On obtient ainsi

$$G_{t_1, t_2}(R) = \frac{1 - e^{-R(t_2 - t_0)}}{1 - e^{-R(t_1 - t_0)}} = \frac{c_2 - c_0}{c_1 - c_0}. \quad (5.10)$$

On peut vérifier que la fonction $R \mapsto G_{t_1, t_2}(R)$ est strictement décroissante sur \mathbb{R}_+ de $(t_2 - t_0)/(t_1 - t_0)$ vers 1. Comme la fonction $c(\cdot)$ est concave, on a

$$(c_2 - c_0)/(t_2 - t_0) < (c_1 - c_0)/(t_1 - t_0).$$

L'équation (5.10) admet donc une solution unique, qui est le taux de renouvellement recherché. Cette équation ne peut pas se résoudre analytiquement, mais on peut utiliser une méthode numérique (dichotomie ou méthode de Newton) pour approcher sa solution.

On peut obtenir une expression plus directe en utilisant le fait que la fonction

$$F : R \mapsto \log \left(\frac{1 - e^{-R(t_2-t_0)}}{1 - e^{-R(t_1-t_0)}} - 1 \right)$$

est très proche d'être affine en R , en particulier⁵ si $t_2 - t_0 = 2(t_1 - t_0)$.

On la remplace par la fonction affine dont le graphe passe par $(0, (t_2-t_0)/(t_1-t_0))$ et $(R_{\text{ref}}, F(R_{\text{ref}}))$, et l'on cherche maintenant le R qui donne à cette fonction affine la valeur

$$\log \left(\frac{t_2 - t_0}{t_1 - t_0} - 1 \right),$$

ce qui conduit à l'estimateur approché

$$\hat{R} = R_{\text{ref}} \frac{\log \left(\frac{c_2 - c_0}{c_1 - c_0} - 1 \right) - \log \left(\frac{t_2 - t_0}{t_1 - t_0} - 1 \right)}{\log \left(\frac{1 - e^{-R_{\text{ref}}(t_2-t_0)}}{1 - e^{-R_{\text{ref}}(t_1-t_0)}} - 1 \right) - \log \left(\frac{t_2 - t_0}{t_1 - t_0} - 1 \right)}$$

5.5 Inertie thermique & capacité thermique apparente

On cherche ici à donner un sens à la notion de capacité thermique apparente pour un système complexe, et plus précisément d'explorer la manière dont cette capacité apparente est susceptible de varier selon la nature de l'excitation. Pour un corps simple, au sein duquel on suppose à chaque instant la température uniforme u , l'énergie thermique est le produit de la capacité thermique C (variable extensive⁶) avec la température (variable intensive). Si l'on fournit à ce corps une puissance Q , on a

$$C \frac{du}{dt} = Q \implies u = \frac{1}{C} \int Q.$$

La température va donc évoluer d'autant plus vite que C est petit, et l'on parle d'*inertie thermique* par analogie avec l'équation de mouvement d'un point matériel, où u est une vitesse, C une masse, et Q une force. Toutes choses égales par ailleurs, u augmente d'autant plus lentement que la masse est grande. On s'intéresse maintenant à un système complexe, au sein duquel la température est susceptible d'être non uniforme, qui est chauffé *de l'exterieur*, ce que l'on va simplement modéliser par le fait que seule une partie du système reçoit directement la chaleur. Nous choisissons de représenter le système complexe par un réseau \mathcal{N} , dont les sommets $x \in V$ sont les constituants élémentaires du système, caractérisés par une capacité thermique propre C_x . Les sommets sont reliés entre eux par des conductances c_{xy} , qui conditionnent le flux de chaleur Q_{yx} de y vers x , selon la loi de Fourier discrète

$$Q_{xy} = c_{xy}(u_y - u_x).$$

On s'intéresse au problème consistant à injecter en un point particulier du réseau, noté $o \in V$, un flux de chaleur, Q (en W) les autres points n'échangeant de la chaleur qu'avec leurs voisins.

Intuitivement, si l'on injecte très lentement de la chaleur en o , de façon à laisser le temps à la température de s'uniformiser au sein du réseau, l'énergie thermique globale est la sommes des énergies, soit $(C_o + \dots + C_N)u_0$, de telle sorte que la capacité thermique apparente du système sur lequel on interagit au travers du seul point o est

$$C_{app} = \bar{C} = \sum_{x \in V} C_x,$$

5. Si l'on prend $t_0 = 0$ pour alléger l'écriture, on a, en développant par rapport à e^{-Rt_1} considéré comme une quantité petite

$$\frac{1 - e^{-Rt_2}}{1 - e^{-Rt_1}} - 1 = e^{-Rt_1} - e^{-Rt_2} + e^{-2Rt_1} + \mathcal{O}(e^{-3Rt_1}),$$

qui est donc égale à e^{-Rt_1} à l'ordre 2 dès que $t_2 = 2t_1$. En prenant le log on obtient donc une fonction affine en R à l'ordre 2 en e^{-Rt_1} .

qui exprime bien le fait que la capacité thermique, définie comme grandeur afférente à une zone donnée de l'espace, est une grandeur extensive.

Toujours intuitivement, si l'on effectue un forçage thermique très chahuté, induisant des variations de u_o trop rapides pour que les voisins de o aient le temps d'en être affectés, on peut s'attendre à ce que le système réagisse comme s'il était limité au seul point o , ce telle sorte que $C_{app} = C_o$, qui peut être très inférieur à \bar{C} . On se propose dans ce qui suit de préciser ces considérations informelles, et de quantifier ce que l'on entend par hautes et basses fréquences d'excitation, pour essayer de formaliser la manière dont la capacité thermique apparente dépend de la fréquence d'excitation.

Le système décrit ci-dessus s'écrit

$$C_x \frac{du_x}{dt} = - \sum_{y \sim x} c_{xy} (u_x - u_y) + Q\delta_{xo},$$

le dernier terme exprimant que le forçage Q n'est exercée qu'au point o . Ce système d'équation s'écrit matriciellement, avec des notations évidentes,

$$C \frac{du}{dt} + Lu = Qe_o \text{ ou } \frac{du}{dt} + C^{-1}Lu = QC^{-1}e_o, \quad (5.11)$$

qui est une équation de la chaleur discrète.

La matrice L du laplacien étant symétrique, il existe $N + 1$ couples propres (λ_j, w_j) solutions du problème aux valeurs propres généralisé

$$Lu = \lambda Cu,$$

où les λ_j sont des réels, et les w_i forment une base C -orthonormée :

$$\langle Cw_j | w_k \rangle = \delta_{jk}.$$

Comme la matrice L est symétrique positive, les valeurs propres sont toutes positives ou nulle. La plus petite valeur propre est 0, qui est associée au vecteur constant w_0 dont toutes les composantes sont égales à $1/\sqrt{\bar{C}}$. On ordonne les valeurs propres

$$0 = \lambda_0 < \lambda_1 \leq \dots \leq \lambda_N.$$

N.B. : Pour tout $j > 0$, la C -orthogonalité de w_j avec w_0 impose que w_j est à C -moyenne nulle, c'est à dire que la variations de températures correspondante ne changent pas l'énergie thermique globale. Seul le premier mode, correspondant aux températures uniformes, est porteur d'énergie.

On cherche une solution à (5.11) (forme de droite) sous la forme

$$u(t) = \sum_{j=1}^N \alpha_j(t) w_j. \quad (5.12)$$

On décompose en premier le second membre dans la base des vecteurs propres :

$$C^{-1}e_o = \sum b_j w_j, \quad b_j = \langle Cw_j | C^{-1}e_o \rangle = \langle w_j | e_o \rangle = w_j^o$$

qui est la valeur en o du vecteur propre w_j (vu comme une fonction propre sur V). En injectant 5.12 dans la forme de droite de (5.11), et en écrivant le système vectoriel composante par composante (sur la base des w_j), on obtient

$$\dot{\alpha}_j + \lambda_j \alpha_j = Qw_j^o.$$

Dans le cas général où Q dépend du temps, on obtient la solution

$$\alpha_j(t) = \alpha_j(0) + \int_0^t e^{-\lambda_j(t-s)} Q(s) w_j^o ds.$$

6. La capacité thermique massique d'un matériau, exprimée en $\text{WK}^{-1}\text{kg}^{-1}$, est la variable intensive associée à cette variable extensive, que l'on peut considérer comme une mesure.

si l'on part d'une température uniforme sur le réseau, fixée à 0, on a donc

$$u(t) = \sum_{j=1}^N \left(\int_0^t e^{-\lambda_j(t-s)} Q(s) w_j^o ds \right) w_j.$$

Le cas d'un forçage périodique permet de préciser le comportement du système selon la fréquence d'excitation $\omega/2\pi$. De façon à travailler avec des forçages qui induisent des variations de température analogues, avec une quantité d'énergie fournie fixée sur chaque demie période, on se place dans le cas d'un forçage donné par $\omega \tilde{Q} e^{i\omega t}$, où \tilde{Q} est un réel fixe (en J). Les équations correspondant aux différentes composantes s'écrivent

$$\dot{\alpha}_j + \lambda_j \alpha_j = \omega \tilde{Q} w_j^o \quad j = 0, \dots, N.$$

Les solutions périodiques des ces équations s'écrivent

$$\alpha_j(t) = \beta_j e^{i\omega t}, \text{ avec } \beta_j = w_j^o \tilde{Q} \frac{\omega}{\lambda_j + i\omega}$$

d'où l'on déduit que la solution globale s'écrit

$$u(t) = e^{i\omega t} \tilde{Q} \sum_{j=1}^N \left(w_j^o \frac{\omega}{\lambda_j + i\omega} \right) w_j,$$

avec la convention que, pour $j = 0$, comme $\lambda_j = 0$, $\omega/(\lambda_j + i\omega) = 1/i$.

Une excitation de basse fréquence correspond à un ω petit. Pour ω tendant vers 0, tous les termes de la suite ci-dessus tendent vers 0 sauf le premier, et l'on a

$$u(t) \approx \frac{1}{i} w_0^o w_0 \tilde{Q} e^{i\omega t},$$

qui correspond à un mode où toutes les températures sont les mêmes au sein du réseau. D'après les propriétés de normalisation de w_0 , on a $w_0^o w_0 = 1/\bar{C}$, d'où, si l'on note $Q = \omega \tilde{Q} e^{i\omega t}$

$$u(t) \approx \frac{1}{\bar{C}} \frac{F}{i\omega}.$$

Si l'on se place dans le position d'un observateur extérieur qui fournit de la chaleur à la racine o , à basse fréquence, la capacité thermique apparente est *celle du réseau entier*, à savoir \bar{C} . Noter que l'on observe ce comportement lorsque ω devient plus petit que λ_1 , qui correspond au taux d'amortissement du système (le plus lent).

Si maintenant on excite le système à haute fréquence, ce que l'on modélise par $\omega \rightarrow +\infty$, tous les rapports dans la somme ci-dessus tendent vers $1/i$, et l'on a

$$u(t) \approx e^{i\omega t} \frac{1}{i} \tilde{Q} \sum_{j=1}^N w_j^o w_j \approx e^{i\omega t} \frac{1}{i} \tilde{Q} C^{-1} e_0,$$

où e_o est le champ $(\delta_{xo})_x$ (Dirac en o). La variation de température n'affecte donc que la racine o , et l'on a, en réécrivant les choses à l'aide du terme de forçage $Q = \omega \tilde{Q} e^{i\omega t}$,

$$u_o(t) = \frac{1}{\bar{C}_o} \frac{F}{i\omega}.$$

L'observateur extérieur voit donc le système comme s'il était réduit au point o , avec une *capacité thermique apparente égale à celle de o* , à savoir C_o . Ce type de comportement est observé lorsque ω devient plus grand que la plus grande des valeurs propres, notée ici λ_N .

5.6 Modélisation de la survie d'un dialecte au sein d'une population

On considère un réseau social représenté par un graphe non orienté (V, E, K) , avec $E = \text{supp}K$, où $K = (K_{xy})$ représente l'intensité des contacts entre x et y . On considère chaque personne caractérisée par une variable $u \in [0, 1]$ qui quantifie sa connaissance et sa pratique d'un dialecte régional, toutes les personnes parlant par ailleurs la langue officielle du pays (disons le français pour fixer les idées) : $u_x = 0$ si x ne pratique ni ne maîtrise le dialecte, et $u_x = 1$ si x le parle couramment et régulièrement. Pour encoder le fait que la préservation de ce dialecte passe par une pratique régulière avec les personnes avec qui l'on est en contact, on considère le système d'équations différentielles suivant (de type Allen-Cahn discret)

$$\frac{du_x}{dt} = - \sum_{y \sim x} K_{xy}(u_x - u_y) + \beta u_x(1 - u_x)(u_x - a),$$

où $a \in [0, 1]$ représente le “point de bascule”, c'est-à-dire le niveau de pratique et / ou de connaissance en dessous duquel la maîtrise du dialecte tend à se détériorer.

5.7 Positionnement de postes de secours

(Les considérations qui suivent font suite à une discussion avec le responsable des services médicaux Paris 2024.)

On se place dans la position de l'organisateur d'une grande manifestation (sportive ou autre), prenant place sur une zone étendue. Du fait de l'afflux attendu de personnes, il souhaite mettre en place un certain nombre de postes de secours pour pré-traiter sur place les urgences médicales (malaises, blessures, déshydratation, ...). Nous présentons une formalisation très simplifiée de la problématique associée, avant d'envisager plus loin des extensions possibles.

On note $X = (x_1, \dots, x_N) \in \mathbb{R}^{2N}$ la collection des zones d'intérêt, qui vont être occupées par des personnes à secourir potentiellement. Le point $x_i \in \mathbb{R}^2$ représente la position du lieu (stade, fan zone, portion de quartier dédiée aux visiteurs, ...), et μ_i représente le nombre de personnes que contient i .

On note $Y = (y_1, \dots, y_M) \in \mathbb{R}^{2M}$ l'ensemble des positions des postes de secours, et l'on note ν_j la capacité d'accueil du poste j , c'est à dire que j peut traiter ν_j personnes par unité de temps. Dans cette version simplifiée de la réalité, on considère que chaque lieu est caractérisé par un coefficient β_i , qui est la fraction de population en i qui va nécessiter une prise en charge par unité de temps⁷.

On note c_{ij} le temps⁸ qu'il faut pour aller de x_i à y_j .

On se place pour l'instant dans la situation où les positions y_j des postes de secours sont arrêtées, ainsi que leurs capacités d'accueil. On se place dans la situation idéalisée d'une capacité de soin parfaitement calibrée, c'est à dire que l'ensemble des postes sont exactement dimensionné pour accueillir le flux attendu, soit

$$\sum_i \beta_i \mu_i = \sum_j \nu_j.$$

7. Ce coefficient peut dépendre de divers facteurs, comme le degré d'entassement (susceptible de générer des malaises), la typologie du public attendu, la météo (risques de déshydratation en cas de forte chaleur), ...

8. On pourra considérer par exemple le temps qu'il faut à une personne en x_i nécessitant des soins légers pour rejoindre à pieds, au travers de la foule, le poste j . Ou si l'on s'intéresse à des urgences absolues, le temps qu'il faut à une équipe de soignants équipés du poste j pour rejoindre la zone i , éventuellement avec un véhicule de transport.

Optimisation des affectations

Un premier problème de transport optimal (voir chapitre 14 pour une présentation générale du cadre théorique sous-jacent) associé consiste à minimiser

$$C(\gamma) = \sum_{ij} c_{ij} \gamma_{ij}$$

sur

$$\Pi = \left\{ \gamma \in \mathbb{R}_+^{N \times M}, \sum_{i=1}^N \gamma_{ij} = \nu_j, \sum_{j=1}^M \gamma_{ij} = \beta_i \mu_i \right\}.$$

On cherche ici à minimiser le temps moyen d'accès au soin, en cherchant à identifier les affectations optimales. Plus précisément, γ_{ij} représente le nombre de malades provenant de la zone i qu'il faut diriger vers le centre j , par unité de temps. Il est éclairant d'introduire la variable intensive associée à cette variable extensive : $\tilde{\gamma}_{ij} = \gamma_{ij}/(\beta_i \mu_i)$ est la fraction de malades en i qu'il faut envoyer au poste j . Cette variable se traduit directement en termes de consignes : il s'agit au personnel affecté à la zone i (s'il existe...), d'envoyer une fraction $\tilde{\gamma}_{ij}$ de malades vers la zone j , pour tous les j tels que $\tilde{\gamma}_{ij} \neq 0$.

Le problème rentre dans le cadre du chapitre 14, plus précisément le problème (14.1), page 287. La proposition 14.3 assure l'existence d'une solution, qui peut être approchée par un algorithme du type régularisation entropique (algorithme 14.21, page 299).

Nous nous sommes intéressés au temps moyen, mais il est plus pertinent de considérer le temps maximal mis pour rejoindre un poste de secours où que l'on soit. On est alors amené à minimiser

$$\max_{ij, \gamma_{ij} > 0} c_{ij},$$

ou une quantité approchante inspirée de la norme ℓ^p ,

$$\sum_{ij} c_{ij}^p \gamma_{ij},$$

pour p grand (qui tend vers la quantité précédente quand p tend vers $+\infty$).

Positionnement optimal des postes

On se place maintenant très en amont : il s'agit de déterminer un positionnement optimal (dans un sens à préciser) des postes de secours. On suppose dans un premier temps que le nombre de ces postes est fixé, égal à M , ainsi que leurs capacités ν_1, \dots, ν_M . Pour toute collection de positions y_1, \dots, y_M , on peut estimer le coût associé définit précédemment, dont nous explicitons la dépendance vis-à-vis des données en position et capacités

$$C = C(y, \nu),$$

qui est le coût optimal, qui admet au moins un minimiseur $\gamma = \gamma(y, \nu)$.

On peut formuler différents problèmes d'optimisation, selon les contraintes que l'on se fixe.

Optimisation à nombre de postes fixé

Le problème le plus général (à M fixé) consiste à trouver y (positions) et ν (capacité d'accueil) qui minimise la quantité ci-dessus :

$$\min_{y, \nu} C(y, \nu)$$

la minimisation se faisant sur les couples (y, ν) , avec $y \in \mathbb{R}^M$, et ν une mesure sur l'ensemble des positions, de masse totale $|\nu|$ égale aux besoins $|\beta \mu|$. Noter que le nombre de postes M ne sera pas

forcément saturé par le minimiseur (certains ν_j peuvent être nul), même si l'on peut s'attendre à ce qu'il le soit.

Noter que la proposition 15.5, page 313, permet de calculer le gradient du coût par rapport aux positions, et la proposition 15.6, page 314 donne l'expression du gradient de ce même coût par rapport à la mesure ν .

On peut aussi fixer la capacité des centres et minimiser simplement en position.

Optimisation libre

Un problème encore plus général consiste à laisser ouvert le nombre de postes de secours. Si l'on considère le problème consistant à minimiser $C(y, \nu)$ sur les couples (y, ν) , avec $y \in \mathbb{R}^M$, et ν une mesure sur l'ensemble des positions, de masse totale $|\nu|$ égale aux besoins $|\mu|$, avec M quelconque, on peut s'attendre à ce qu'il n'y ait pas de minimiseur. En effet, si l'on considère le problème pour M fixé considéré précédemment, et que l'on augmente d'une unité de nombre de poste, on va en général pouvoir diminuer strictement la distance moyenne ou la distance maximale. Une construction selon le principe d'une suite minimisante conduira à un nombre de postes augmentant à l'infini, avec possible convergence faible de l'ensemble des postes vues comme une mesure atomique vers une mesure diffuse, qui correspondrait à un étalement complet des postes de secours sur l'ensemble de la zone d'intérêt, ce qui n'est manifestement pas réaliste. On peut rendre l'approche plus réaliste en considérant par exemple que multiplier le nombre de postes conduit à des coûts et / ou des difficultés de mise en place rédhibitoires. On peut intégrer cela en rajoutant à la fonctionnelle à minimiser un terme qui pénalise le nombre de postes, ce qui conduit à minimiser une fonctionnelle du type

$$C(y, \nu) + \Psi(M)$$

où Ψ est une fonction croissante de M , qui tend vers l'infini quand M tend lui-même vers l'infini.

5.8 Affluence au supermarché

On considère une population de clients d'un supermarché répartie en N groupes, de tailles $\mu_1, \mu_2, \dots, \mu_N$ (considérés comme des nombres réels). On considère par ailleurs des créneaux horaires $1, \dots, M$ qui recouvrent la plage d'ouverture du supermarché. Les groupes de clients correspondent à des préférences communes en termes d'horaire. On notera u_{ij} l'utilité des clients de type i pour le créneau j .

Approche administrée

On se place en premier lieu dans la situation caricaturale d'un système complètement administré. On suppose que chaque créneau j est affecté d'une stricte capacité d'accueil ν_j , de telle façon que la capacité totale est égale à la population totale :

$$\sum \mu_i = \sum \nu_j.$$

Le problème d'optimisation naturellement associé consiste à écrire

$$\max_{\Pi_{\mu, \nu}} \sum_{ij} \gamma_{ij} u_{ij}, \quad \Pi_{\mu, \nu} = \left\{ \gamma \in \mathbb{R}^{M \times N}, \sum_i \gamma_{ij} = \nu_j, \sum_j \gamma_{ij} = \mu_i \right\}$$

Il correspondrait au cas d'un administration toute puissante disposant de capacité de traitement des clients / administrés exactement égale à la demande, et cherche à maximiser la satisfaction moyenne des personnes. Ce problème n'a en effet de sens que si le choix des créneaux horaires est décidé par une entité extérieure, qui a la visio de l'ensemble des besoins, des ressources, et des souhaits des personnes.

Un problème voisin consiste à se placer dans une situation moins tendue en termes de ressources, en supposant simplement que la capacité de l'offre totale dépasse les besoins. On suppose alors

$$\sum \mu_i \leq \sum \nu_j, \quad \Pi_{\mu, \nu} = \left\{ \gamma \in \mathbb{R}^{M \times N}, \sum_i \gamma_{ij} \leq \nu_j, \sum_j \gamma_{ij} = \mu_i \right\}.$$

On peut aussi considérer la situation de ressources insuffisantes, avec une inégalité dans l'autre sens. Le problème consiste toujours à maximiser l'utilité globale, en acceptant de "sacrifier" une partie de la population (qui n'aura pas le droit d'aller faire ses courses).

Système "libre"

On se place maintenant dans une optique différente, plus proche du monde qui nous entoure, considérant que les choix de créneaux vont être effectués par les individus eux-mêmes. On ne suppose plus que les créneaux horaires disposent d'une capacité limitée, mais on considère en revanche que l'affluence est pénalisante (attente aux caisses, rayons encombrés, ...). On modélise cette congestion en définissant une utilité effective correspondant au choix de i pour j , qui dépend des choix effectués par les autres clients, et réduit l'utilité d'autant plus que le nombre de clients est important. On écrit cette utilité effective sous la forme

$$u_{ij} - \varphi \left(\sum_i \gamma_{ij} \right),$$

où φ est une fonction strictement croissante, que nous supposeront continue. Il s'agit donc d'une utilité *a posteriori*, qui n'a pas de sens avant que la journée ne se soit déroulée, puisqu'elle dépend des choix effectués par les autres personnes.

On peut considérer une version "administrée" de ce problème, en cherchant à maximiser l'utilité effective globale sous contrainte de simple marginale

$$\max_{\Pi_\mu} \sum_{ij} \gamma_{ij} \left(u_{ij} - \varphi \left(\sum_i \gamma_{ij} \right) \right), \quad \Pi_\mu = \left\{ \gamma \in \mathbb{R}^{M \times N}, \sum_j \gamma_{ij} = \mu_i \right\}.$$

Il s'agit d'un problème d'optimisation classique (maximisation d'une fonction continue sur un polyèdre), potentiellement difficile à résoudre effectivement du fait de la non concavité de la fonctionnelle à minimiser.

Pour sortir de ce monde administré, on peut s'intéresser à la notion d'équilibre de Nash, qui correspond à un plan d'affectation γ_{ij} tel que personne n'a strictement intérêt à changer son choix. Plus précisément, on appellera équilibre de Nash un plan γ_{ij} tel que, pour tout i ,

$$\forall j, \gamma_{ij} > 0 \implies u_{ij} - \varphi \left(\sum_i \gamma_{ij} \right) = \max_{j'} \left(u_{ij'} - \varphi \left(\sum_i \gamma_{ij'} \right) \right).$$

5.9 Densités de population et contacts induits

(En collaboration avec P. Poncet, géographe, et S. Faure)

On considère une zone urbaine représentée par un domaine Ω , dans lequel la densité d'habitants $\rho = \rho(x)$ est donnée. On cherche à explorer des moyens de localiser les zones dans lesquelles les contacts entre habitants seront susceptibles d'être les plus nombreux. Nous proposons en premier lieu un modèle macroscopique basé sur les considérations suivantes : chaque individu de la population émet un flux continu d'*avatars*, à un taux β , qui correspondent à des cheminements possibles de la personne en question. Le mouvement des ces avatars est de type diffusif, on pourra imaginer que chaque avatar suit un mouvement de type brownien. La portée de la zone couverte par ces avatars est contrôlée par

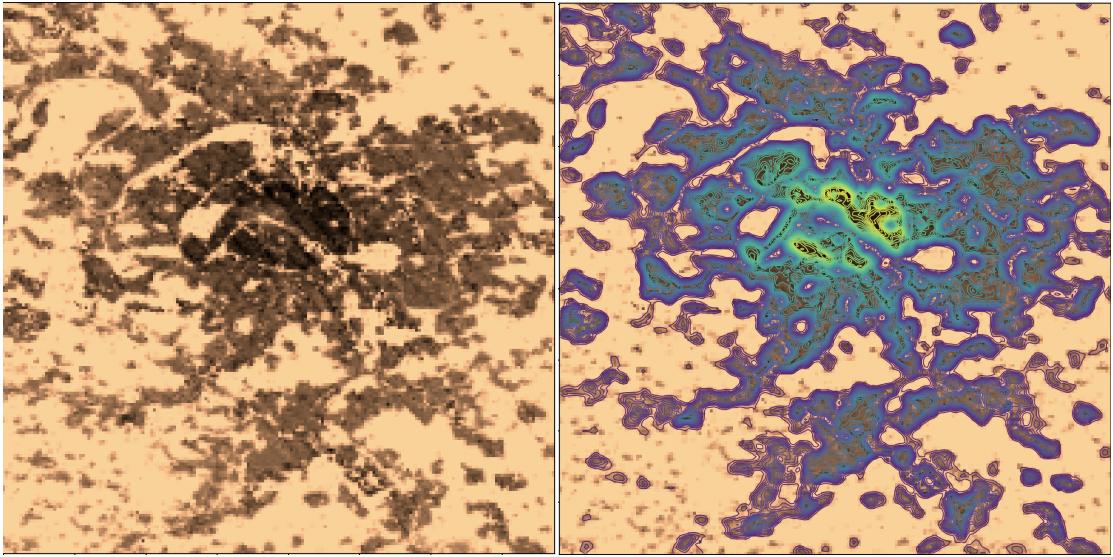


FIGURE 5.2 – Représentation de la densité de résidents, et des solutions stationnaires obtenues pour $\sigma = 0.01$.

un paramètre γ , tel que $1/\gamma$ est le temps caractéristique de vie. La distance moyenne parcourue en un temps t étant de l'ordre de $\sqrt{2Dt}$ (voir équation 4.6, page 91), l'étendue de la zone visitée est $\sigma = \sqrt{2D/\gamma}$.

On note u la densité d'avatars (émis par l'ensemble des individus de la population considérée). Cette densité suit l'équation d'évolution suivante

$$\partial_t u - D\Delta u = \beta\rho(x) - \gamma u,$$

avec des conditions aux limites de type Dirichlet homogène sur la frontière de Ω (les avatars sont supposés disparaître lorsqu'il atteignent le bord de Ω). Ce champ converge vers la solution unique de l'équation

$$\gamma u - D\Delta u = \beta\rho.$$

5.10 Modèles de mobilité sur réseaux

Nous proposons ici une représentation d'un réseau de mobilité sous la forme la forme d'un graphe (réseau routier et/ ou ferroviaire). Les utilisateurs du réseau (appelés *agents*, ou *voyageurs* dans la suite) ne sont pas représentés individuellement, mais par un nombre réel positif affecté à certains sommets du graphe, nombre quantifiant le nombre de personnes résidant dans la zone associée au sommet (qui peut correspondre à une station de transport en commun, ou à un quartier de ville). La difficulté d'analyse des modèles proposés ci-dessous tient en particulier au fait que le temps de transport associé à un itinéraire choisi par un agent peut dépendre du choix des autres agents (phénomène de congestion). Comme dans le cadre des *jeux à champ moyen* (dont les modèles présentés ici sont en quelque sorte une version discrète particulière), on considérera qu'un agent seul ne pèse rien : son temps de parcours dépend du choix de l'ensemble des autres, mais sa propre décision individuelle n'affecte pas à elle seule le temps de parcours d'un chemin selon qu'il le choisisse ou pas.

5.11 Cadre général, problématiques

On définit un réseau de transport $\mathcal{N} = (V, E)$ de la façon suivante. Un sommet $x \in V$ correspond à une zone géographique de taille réduite⁹ dans laquelle habitent ou travaillent des gens (proximité d'une station de RER / métro, quartier d'une ville), ou un point de bifurcation (échangeur routier, gare de correspondance, croisement de rue...). Une arête $e \in E$ correspond à un tronçon routier (ou autre) orienté entre deux points¹⁰ de bifurcation, à un tronçon ferré reliant deux gares successives,

On attribuera un type C (transports en commun) ou R (route) à chaque arête. On note $X \subset V$ un ensemble de sommets correspondant aux zones d'habitation, qui constituent les origines des chemins effectués le matin, et $Y \subset V$ l'ensemble des sommets correspondant aux lieux d'activité, qui sont les destinations des parcours. On se donne une mesure $\mu = (\mu_x) = \mathbb{R}_+^X$ sur X quantifiant les nombres de personnes rattachées aux différentes zones de X .

On suppose connu un plan Origine-Destination $(\gamma_{xy}) \in \mathbb{R}_+^{X \times Y}$. La quantité γ_{xy} représente le nombre de personnes qui souhaitent aller le matin de x à y . Ce plan admet μ comme marginale en X , i.e.

$$\sum_{y \in Y} \gamma_{xy} = \mu_x \quad \forall x \in X,$$

et sa marginale ν en Y correspond aux poids des différentes destinations¹¹ :

$$\nu_y = \sum_{x \in X} \gamma_{xy} \quad \forall y \in Y.$$

On notera $\Lambda_{\mu, \nu}$ l'ensemble des plans OD qui satisfont ces conditions de marginales. On introduit le support de (γ_{xy}) :

$$\Lambda = \text{supp}(\gamma_{xy}) = \{(x, y) \in X \times Y, \gamma_{xy} > 0\}, \quad (5.13)$$

qui est l'ensemble des couples origine-destination qui concernent une quantité non nulle de voyageurs.

Pour $x \in X$ et $y \in Y$, on définit un chemin comme une suite d'arêtes consécutives, la première étant issue de x , la dernière se terminant en y . On dit qu'un chemin c est *admissible* s'il ne contient que des arrêtes de type C (transports en commun), ou que des arêtes de type R (route), ou un premier tronçon de type R suivi d'un second de type C , dans cet ordre¹².

On suppose Y fortement connecté à X relativement à Λ , ce qui signifie que, pour tout couple $(x, y) \in \Lambda$, il existe au moins un chemin admissible de x vers y .

On note C_{xy} l'ensemble des chemins admissibles de x vers y (non vide par hypothèse). Un scénario correspond à un choix de distribution, pour tout $(x, y) \in \Lambda$, entre les différents chemins possibles pour aller de x vers y , ce que l'on peut représenter, pour tout $(x, y) \in \Lambda$, par

$$(\gamma_{xy}^c)_{c \in C_{xy}} \in \mathbb{R}_+^{C_{xy}},$$

où γ_{xy}^c représente le nombre de personnes qui vont de x à y en passant par le chemin $c \in C_{xy}$. On a donc par construction

$$\sum_{c \in C_{xy}} \gamma_{xy}^c = \gamma_{xy}.$$

9. Ces zones sont définies de façon à ce que, pour une personne et une destination données, son choix de parcours ne dépend essentiellement pas de sa position dans la zone définie.

10. On représentera donc une voie à double sens entre z et z' par deux arêtes (z, z') et (z', z) , auxquelles on associera des flux positifs. Cette convention l'importance de ne pas effectuer a priori (sauf dans certaines situations) le bilan des flux. Elle permet par ailleurs nativement de prendre en compte des voies à sens unique, ou des voies à double sens dont l'un des axes est momentanément supprimé ou modifié (accident sur une voie rapide, travaux,...).

11. Cette marginale ν résulte de γ_{xy} , qui lui-même exprime les contraintes professionnelles des agents, données a priori, même si elles ne sont pas toujours *connues* en pratique. Il ne s'agit pas d'une contrainte qui doit être vérifiée par γ , comme c'est le cas en transport optimal.

12. On s'autorise ces parcours mixtes pour prendre en compte les personnes qui rejoindraient une gare le matin à l'aide de leur véhicule, pour le retrouver le soir. La contrainte que le véhicule termine sa journée à la maison dissymétrise les rôles de X et de Y .

De Lagrange à Euler

Le plan γ_{xy}^c est une donnée de nature lagrangienne (qui va où, et par quel chemin), alors que les données observées sur un tel réseau peuvent être de nature eulérienne (on compte par exemple le nombre de personnes qui passent sur une période donnée au travers d'une arête particulière du réseau). Pour simplifier, considérons en premier lieu que toutes les personnes se rendant de x à y empruntent le même chemin, noté c_{xy} . Si l'on note f_e le nombre de personnes qui passent pendant une phase de transport au travers de l'arête (orientée) e , on a

$$f_e^\gamma = \sum_{(x,y), e \in c_{xy}} \gamma_{xy}.$$

Cette formule explicite correspond au passage d'une vision lagrangienne à une vision eulérienne $L \rightarrow E$, qui ne pose pas de problème particulier. La correspondance inverse $E \rightarrow L$ est beaucoup moins simple, du fait notamment qu'il existe a priori plusieurs scénarios possibles qui peuvent conduire à la même collection de flux observés (f_e). Ce problème de reconstruction peut prendre plusieurs formes, conditionnées par la nature des informations dont on dispose. On peut par exemple considérer que les marginales μ (habitants) et ν (lieux de travail) sont connues, et que l'on observe les flux journaliers sur une partie du réseau, i.e. on se donne une collection de flux $\tilde{f}_{\tilde{E}} = (\tilde{f}_{\tilde{e}})_{\tilde{e} \in \tilde{E}}$, où $\tilde{E} \subset E$ est la collection des arêtes (orientées) sur lesquelles on mesure les flux. Le problème consiste alors à trouver $\gamma \in \Lambda_{\mu,\nu}$ tel que

$$\tilde{f}_{\tilde{e}} = \sum_{(x,y), \tilde{e} \in c_{xy}} \gamma_{xy} = \tilde{f}_{\tilde{e}} \quad \forall \tilde{e} \in \tilde{E}.$$

Ce problème est de façon évidente mal posé en général, au moins en termes d'unicité. On peut poser le problème en termes d'univers des possibles, en cherchant par exemple à décrire l'ensemble des plans compatibles avec les observations, c'est à dire l'ensemble

$$\Lambda_{\tilde{f}_{\tilde{E}}} = \{\gamma \in \Lambda_{\mu,\nu}, f_{\tilde{e}}^\gamma = \tilde{f}_{\tilde{e}} \quad \forall \tilde{e} \in \tilde{E}\}.$$

Sans information supplémentaire, il peut être pertinent dans cette situation de minimiser l'entropie sur l'ensemble des champs compatibles :

$$\min_{\gamma \in \Lambda_{\tilde{f}_{\tilde{E}}}} \sum_{(x,y)} \gamma_{xy} \log \gamma_{xy},$$

qui est un problème de minimisation d'une fonctionnelle strictement convexe sur un compact convexe, et admet donc une solution unique.

Temps de parcours.

On cherche ici à élaborer un critère selon lequel les agents seraient susceptibles de se déterminer au moment de choisir entre plusieurs options. Nous considérons dans un premier temps que le choix se fait uniquement sur la durée du parcours. Nous notons T_{xy}^c le temps de parcours associé au chemin $c \in C_{xy}$ permettant d'aller de x vers y . Ce temps dépend en premier lieu du chemin choisi, mais aussi a priori du choix des autres agents, du fait par exemple que si un grand nombre de personnes empruntent un même axe routier, le temps de parcours va être rallongé. On a donc en toute généralité

$$T_{xy}^c = T_{xy}^c(\gamma),$$

où $\gamma = (\gamma_{xy}^c)$ est le plan de transport décrivant l'ensemble des choix des agents. Il s'agit donc d'un temps de parcours *a posteriori*, puisqu'il dépend, pour un jour donné, de l'ensemble des choix effectués par les agents, qui n'est pas connu *a priori*.

Remarque 5.2. Le temps $T_{xy}^c(\gamma)$ a un sens aussi pour un chemin qui n'est emprunté par personne, i.e. tel que $\gamma_{xy}^c = 0$. Le chemin $c \in C_{xy}$, même s'il est à un moment dédaigné par les gens qui vont de x vers y , peut être en effet une option possible, qui peut à l'occasion être plus intéressante que les chemins effectivement empruntés.

Observables, problématiques associées

Si l'on suppose connue la fonction $T^c(\cdot)$ d'estimation des temps de parcours, on peut associer à un plan $\gamma = (\gamma_{xy}^c)$ le *temps moyen de parcours* défini par

$$\bar{T} = \bar{T}(\gamma) = \frac{1}{\mu_X} \sum_{(x,y) \in X \times Y} \sum_{c \in C_{xy}} \gamma_{xy}^c T_{xy}^c(\gamma), \quad \text{avec } \mu_X = \sum_{x \in X} \mu_x. \quad (5.14)$$

Remarque 5.3. On prendra garde au fait que ce temps moyen n'est pas une fonction linéaire de γ , du fait de la dépendance des temps de parcours vis à vis de γ . On notera aussi le caractère non local des termes intervenant dans la sommation : le temps de parcours associé à un chemin c de x vers y est susceptible d'être influencé par le nombre de voyageurs se rendant de $x' \neq x$ à $y' \neq y$, si le chemin c' en question partage un tronçon avec c .

Remarque 5.4. Si l'on note T_{xy}^0 le temps de parcours “à vide” de x vers y , c'est à dire le temps minimal que l'on peut mettre pour aller de x vers y (sans congestion), la différence $T_{xy}^c - T_{xy}^0 \geq 0$ mesure d'une certaine manière la *frustration* des gens qui vont de x à y par c en un temps T_{xy}^c , conscients qu'ils auraient pu mettre T_{xy}^0 s'ils étaient tous seuls.

On peut associer à un plan γ d'autres observables, comme la quantité totale de gaz à effet de serre émise sur l'ensemble des trajets, ou la quantité de micro-particules résultant des fortement des pneumatiques et de la chaussée, qui dépend des conditions de circulation.

Équilibre de Wardrop

Definition 5.5. (Équilibre de Wardrop pour un plan de mobilité)

Dans le cadre des notations introduites précédemment, on dit que le plan $\gamma = (\gamma_{xy}^c)$ réalise un équilibre de Wardrop si

$$T_{xy}^c \geq T_{xy}^{max} = \max_{c \in C_{xy}, \gamma_{xy}^c > 0} T_{xy}^c \quad \forall c \in C_{xy}, \quad \text{avec égalité si } \gamma_{xy}^c > 0,$$

En d'autres termes les temps de parcours associés aux chemins c de C_{xy} qui sont effectivement empruntés sont tous les mêmes, et cette valeur commune est inférieure ou égale à celles des chemins non choisis. Cette propriété exprime qu'un agent devant aller de x à y , et ayant opté pour le chemin c , n'a pas d'option alternative qui lui permettrait de diminuer strictement son temps de parcours, au vu de ce que font les autres agents.

Dépendance du temps de parcours vis-à-vis de γ

Le temps de parcours effectif d'un agent dépend de nombreux facteurs, et se trouve conditionné par des phénomènes ponctuels et locaux qu'il est difficile de prévoir. Pour les parcours sur route, le temps de parcours d'un conducteur dépend en particulier de l'apparition ou non de bouchons, qui peuvent apparaître ponctuellement et se résorber, de telle sorte que pour un même parcours, un décalage du moment du départ pourra affecter significativement ce temps. On peut néanmoins considérer que le temps effectif moyen est une fonction décroissante de la quantité de personnes présentes. Dans cette optique, pourra considérer par exemple que le temps mis pour parcourir une arête e s'écrit comme un temps de référence T_0^e (longueur de e divisée par la vitesse limite sur ce tronçon) multiplié par un facteur correctif inférieur à 1 qui dépend du nombre de personnes qui empruntent le parcours

$$T^e = T_0^e \Psi \left(\sum_{x,y} \sum_{c \in C_{xy}, c \ni e} \gamma_{xy}^c \right).$$

Si l'on s'inspire des modèles classiques de trafic routier, on pourra prendre par exemple¹³

$$\Psi(m) = \left(1 - \frac{m}{m_{max}^e} \right)^{-1},$$

où m_{max}^e est la quantité de véhicules qui conduit, sur la période de temps considérée, à une saturation complète du trafic.

5.12 Exercices

Exercice 5.1. (Passage Lagrange → Euler et Euler → Lagrange pour un réseau linéaire)

On s'intéresse ici à une ligne de bus, dont on note $\{0, \dots, N\}$ les arrêts (on se focalise sur les déplacements vers les indices croissants). On note μ_j le nombre de personnes qui montent dans le bus à l'arrêt j , ν_j le nombre de personnes qui en descendent, et ρ_j le nombre de voyageurs dans le bus entre l'arrêt j et l'arrêt $j+1$. On notera $\mu, \nu \in \mathbb{R}^{N+1}$ et $\rho \in \mathbb{R}^N$ les vecteurs associés (N.B. : on pourra raisonnablement considérer que $\nu_0 = \mu_N = 0$). On note $\gamma = (\gamma_{ij})$ le plan origine-destination : γ_{ij} est le nombre de personnes montées en i qui descendent en j . On définit le plan origine destination normalisé $\tilde{\gamma}$ par

$$\tilde{\gamma}_{ij} = \frac{\gamma_{ij}}{\mu_i} \quad i < N, \quad i < j \leq N.$$

Ainsi défini, $\gamma_i = (\gamma_{ij})_j$ est une loi de probabilité supportée par $\{i+1, \dots, N\}$, qui décrit la distribution des destinations des voyageurs montant en i . On note enfin $b_j = \mu_j - \nu_j$ le bilan montée-desccente à l'arrêt j .

1) Écrire $\rho = (\rho_j)$ en fonction de $b = (b_j)$, ainsi que la réciproque. (On écrira les relations entre les ρ_j et les b_j , puis on proposera une formulation matricielle).

2)a) Écrire la relation entre γ , μ , et b sous la forme

$$G\mu = b,$$

où G est une matrice qui dépend du plan renormalisé $\tilde{\gamma}$, dont on précisera l'écriture.

b) Montrer que si l'on connaît $\rho = (\rho_j)$ (historique de l'occupation du bus sur les différents tronçons) et $\tilde{\gamma}$, on peut reconstruire μ .

c) La reconstruction précédente est-elle stable ?

3)a) On ne fait plus d'hypothèse sur $\tilde{\gamma}$. On suppose que l'on est capable de mesurer les entrées-sorties à chaque station (i.e. μ et ν sont supposés connus). Décrire le plus précisément possible l'ensemble des plans OD renormalisés $\tilde{\gamma}$ compatibles avec les observations. (N.B. : on pourra introduire les quantités ρ_j^i qui représentent le nombre de voyageurs montés en i présents dans le bus sur le tronçon $[j, j+1]$).

b) Décrire un algorithme de reconstruction explicite de $\tilde{\gamma}$ sous une hypothèse First In First Out (FIFO), c'est à dire que tout individu sort avant tous les voyageurs montés après lui.

c) Même question sous une hypothèse Last In First out, i.e. le personnes qui sortent en premier sont celles qui sont montées le plus récemment.

3) Proposer une formulation continue du cadre précédent, i.e. la situation d'un bus imaginaire dont l'ensemble des arrêts est un continuum.

Exercice 5.2. (Équilibre de Nash pour un problème-jouet)

On considère la situation représentée sur la figure 5.3, qui représente de façon très schématique un réseau routier. Une quantité μ_1 de personnes doit se rendre, à une période donnée, du point x_1 au point y_1 , et une quantité μ_2 du point x_1 au point y_2 . La population 1 a le choix entre le chemin c_1 et le chemin c_{12} , la 2 entre le même c_{12} et un autre chemin c_2 . On suppose au départ que le temps de

13. Cette expression du temps dérive d'une expression de la vitesse du type

$$v(\eta) = V(1 - \rho/\rho_{max}),$$

où ρ est la densité linéique de véhicules, et ρ_{max} la densité de saturation. Nous considérons ici que cette densité linéique moyenne sur la période considérée est proportionnelle à la quantité de personnes qui vont emprunter cet axe routier, ce qui est discutable.

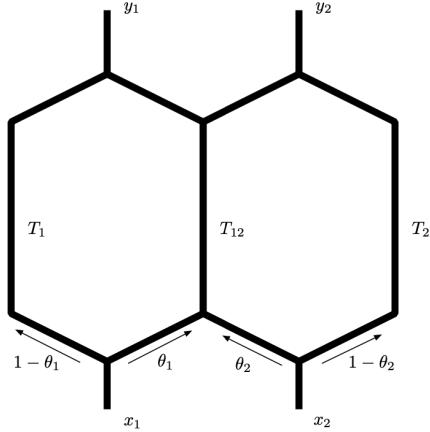


FIGURE 5.3 – Réseau routier

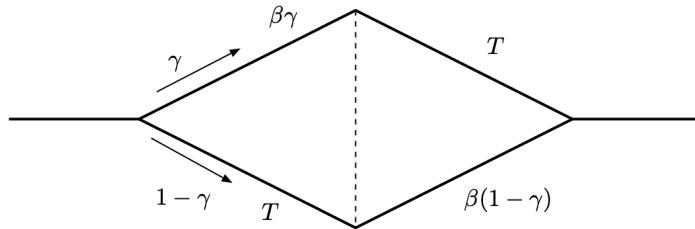


FIGURE 5.4 – Paradoxe de Braess

parcours associé au chemin i est $T_i > 0$, et que le temps de parcours associé à c_{12} croît avec la densité de véhicules qui l'utilisent. On supposera plus précisément

$$T_1 = T_2 = T, \quad T_{12} = T_{12}^0 + \beta\mu, \quad T_{12}^0 < T,$$

où μ est la quantité de gens qui empruntent la route c_{12} . On suppose ici que les flux en 1 et 2 sont unitaires, et l'on note θ_i la fraction de i qui utilise la route partagée. Les temps de parcours sont donc

$$T, \quad T_{12}^0 + \beta(\theta_1 + \theta_2), \quad T.$$

- 1) Préciser les équilibres de Nash du système
- 2) On suppose que le comportement des gens évolue d'un jour sur l'autre, en partant de la situation où chaque agent choisit la stratégie qui correspond au plus petit temps de parcours à vide.
 - a) Décrire l'évolution du système si l'on suppose que chaque agent choisit au jour $k+1$ la stratégie qui se serait révélée la meilleure au jour k .
 - b) Proposer des modèles d'évolution susceptibles d'évoluer vers un équilibre de Nash, ou au moins de s'en rapprocher.

Exercice 5.3. (Paradoxe de Braess pour un problème-jouet.)

On considère la situation représentée sur la figure 5.4 : un flux unitaire de personnes venant de la gauche cherche à rejoindre le point droit du réseau. En prenant vers le sud au départ (en bas sur le dessin), un conducteur utilise un tronçon dont le temps de parcours est fixe égal à T , puis un tronçon au temps de parcours “à vide” négligeable, mais qui augmente linéairement avec le trafic (temps $\beta\mu$, où μ mesure le trafic). S'il part à gauche, il traverse deux tronçons du même type, dans l'ordre inverse. On note γ le flux de gens qui prennent la voie du haut, et $1 - \gamma$ le flux vers le bas.

- 1) Montrer que le système admet un unique équilibre de Nash.
- 2) On suppose maintenant que, dans l'optique de fluidifier l'écoulement, on rajoute une voie (en pointillé vertical sur la figure) qui relie les milieux des deux routes. On suppose que le temps de parcours de cette voie rajoutée est négligeable, on le prend égal à 0. On suppose en outre que $\beta = \alpha T$ avec $\alpha < 1$. Montrer que le scénario précédent n'est plus un équilibre de Nash. Montrer que, ce nouveau réseau admet un équilibre de Nash différent du premier, et que le temps de parcours correspondant, pour certaines valeurs de $\alpha < 1$, est *supérieur* à celui du premier réseau.

5.13 Mobilité individuelle

On cherche à représenter les pratiques de mobilité d'une population X de N individus. On note Y l'ensemble des lieux visités par la population, de cardinal M . Pour tout individu $x \in X$, on note K_{xy} la fraction du temps qu'il passe au lieu y . La matrice $K = (K_{xy})$

$$K = \begin{pmatrix} K_{x_1 y_1} & K_{x_1 y_2} & \dots & K_{x_1 y_M} \\ K_{x_2 y_1} & & & \\ \vdots & \vdots & \ddots & \\ \vdots & \vdots & \ddots & \\ \vdots & \vdots & \ddots & \\ K_{x_N y_1} & K_{x_1 y_2} & \dots & K_{x_N y_M} \end{pmatrix}$$

est donc stochastique par construction. Ses lignes donnent une information de nature *lagrangienne* sur les pratiques de mobilités.

Si l'on note $e = (1, 1, \dots, 1)^T \in \mathbb{R}^N$, on a

$$K^T e = (N_y),$$

où N_y est le nombre moyens de visiteurs en y . La matrice K^T , ou plutôt le bilan de ses lignes, encode donc une information de nature eulérienne.

Chapitre 6

Propagation d'opinion sur réseaux sociaux

Sommaire

6.1	Modèle d'évolution	122
6.2	Cadre stochastique	128
6.3	Liens avec les schémas de discréétisation des EDP	131
6.4	Monitoring of a network though influencers / influence coefficients	132
6.5	Modèle continu et structure de flot de gradient	135
6.6	Réseaux charismatiques	140
6.7	Niveau de certitude	143
6.8	Clivage des opinions au sein d'une population	145
6.9	Propagation sur des réseaux du type "application"	146
6.10	Exercices	148

6.1 Modèle d'évolution

Nous présentons ici un modèle simple de propagation d'opinion sur un réseau. Les noeuds de ce réseaux sont des personnes, ou "agents", auxquels on affecte un nombre représentant une opinion sur un certain sujet à un instant donné. Ce nombre peut par exemple représenter la tendance qu'a un individu à voter pour tel ou tel candidat au second tour d'une élection présidentielle, ou l'idée que l'on peut se faire de la probabilité de gain d'une équipe nationale à une finale de coupe du monde. Dans un autre contexte, on pourra penser à la valeur d'une quantité qui fait l'objet d'un débat public, comme l'augmentation de la température moyenne sur la planète dans 20 ans.

On considère un ensemble V de N individus, on note u_x^t l'opinion de l'individu x à l'instant t , et par $u^t = (u_x^t)_{x \in V}$ la collection des opinions. L'influence de $y \in V$ sur x est quantifiée par un coefficient $K_{xy} \in [0, 1]$, et l'on suppose :

$$\sum_{y \in V} K_{xy} = 1 \quad \forall x \in V.$$

La collection de l'ensemble des coefficient est donc encodée par une matrice (sans choix de numérotation des sommets) $(K_{xy})_{(x,y) \in V^2}$ stochastique.

On notera que toute l'information sur le graphe est dans la collection des influences : $E \subset V \times V$ est défini par

$$E = \text{supp}(K_{xy}) = \{(x, y) \in V \times V, K_{xy} > 0\}.$$

On conservera néanmoins la notation (redondante) (V, E, K) pour désigner le réseau.

On écrira $x \rightarrow y$ si $K_{xy} > 0$, qui signifie que x écoute y , ou x suit y , ou plus généralement x est influencé par y . Avec cette convention, l'opinion / influence remonte le sens des flèches.

Le coefficient diagonal K_{xx} peut être non nul (présence de boucles dans le réseau)), ce qui correspond à une certaine inertie de x , ou résistance de x à modifier son opinion sous l'effet d'influences extérieures, jusqu'à éventuellement ne plus se préoccuper de l'opinion des autres (cas extrême $K_{xx} = 1$) On note $\Gamma \subset V$ l'ensemble des sommets qui ne pointent que vers eux-mêmes

$$\Gamma = \{x \in V, K_{xx} = 1\}, \quad (6.1)$$

et par $\mathring{V} = V \setminus \Gamma$ l'ensemble des sommets intérieurs. La frontière Γ correspond aux individus qui n'écoutent qu'eux mêmes, et étant éventuellement suivis par d'autres. Si l'on considère que ces agents affichent une opinion dans le dessein de modifier l'opinion d'autres agents du réseau, on peut voir ces individus comme des *influenceurs*¹, ou plus simplement, s'ils ne nourrissent aucun dessein particulier, de personnes non influençables, ou *têtues*.

Modèle discret d'évolution

On se donne une collection d'opinions initiales $(u_x^0)_{x \in V}$, et l'on considère que l'opinion évolue d'un jour à l'autre selon la relation

$$u_x^{k+1} = \sum_{x \rightarrow y} K_{xy} u_y^k \quad \forall x \in V. \quad (6.2)$$

qui peut s'écrire aussi matriciellement. On notera que l'indication ' $x \rightarrow y$ ' n'est pas à strictement parler obligatoire, du fait que $K_{xy} = 0$ dès que $(x, y) \notin E$. Elle sera parfois omise, et l'on écrira alors simplement $\sum K_{xy} u_y^k$.

Definition 6.1. (Point d'équilibre)

On dit que $u = (u_x)_V \in \mathbb{R}^V$ est point d'équilibre pour (6.2) si

$$u_x = \sum_{x \rightarrow y} K_{xy} u_y \quad \forall x \in V. \quad (6.3)$$

L'ensemble des points d'équilibre est un sous-espace vectoriel, dont la dimension peut prendre n'importe quelle valeur entre 1 et le nombre de sommets N , selon la structure de (K_{xy}) . Tout point d'équilibre est dans le noyau de $L = \text{Id} - K$, défini comme le laplacien associé au réseau (voir définition 10.33, page 230), on parlera donc de *champ harmonique*.

Un point d'équilibre correspond à une situation où les opinions ne varient plus. Du fait des hypothèses sur K le vecteur $e = (1, \dots, 1)$ est point d'équilibre, ainsi que tout λe pour tout $\lambda \in \mathbb{R}$. Un tel point d'équilibre est appelé un *consensus*. Il peut exister d'autres types de points d'équilibre, par exemple dans le cas extrême où toutes les arêtes sont des boucles, tout vecteur est point d'équilibre. En termes d'opinion, cela signifie simplement qu'il n'y a aucune communication entre les membres d'un groupe, et qu'ainsi chacun campe sur son opinion initiale, qui est arbitraire. Dans la situation intermédiaire avec quelques influenceurs qui ont des opinions différentes, on verra qu'il existe d'autres types de points d'équilibres, définis comme solution d'un problème de type Dirichlet présenté plus loin.

Remarque 6.2. L'équation (6.2) peut être écrite

$$u_x^{k+1} = u_x^k + \sum_{x \rightarrow y} K_{xy} (u_y^k - u_x^k) = u_x^k + \sum_{x \rightarrow y} F_{yx}^k, \quad (6.4)$$

1. Ces influenceurs peuvent aussi être vus comme des agents influencés par une entité extérieure (entreprise, groupe de pression, ...) qui les contrôle.

où $F_{yx}^k = K_{xy} (u_y^k - u_x^k)$ peut être interprété comme un *flux* d'opinion de y vers x . La définition de ce flux peut être vue comme une sorte de loi d'Ohm's (ou loi de Poiseuille dans un contexte fluide), qui énonce une relation de proportionnalité entre un flux et une différence de potentiels (ou différence de pression en mécanique des fluides). De ce point de vue, K_{xy} apparaît comme un coefficient de *conductance* qui conditionne le passage de y vers x . On notera que, contrairement aux lois phénoménologique du monde physique, cette loi n'est pas symétrique : on peut en particulier avoir $K_{xy} > 0$ alors $K_{yx} = 0$. Cette non-symétrie est caractéristique d'interactions entre entités dotées de capacités cognitives qui permettent le transfert d'information.

Proposition 6.3. (Principe du maximum, version dynamique)

Soit $u^0 \in \mathbb{R}^V$ un état initial, et u^1, \dots, u^k, \dots les champs définis par (6.2). Pour tout $k \geq 0$, tout $x \in V$, on a

$$u_x^{k+1} \in [\min u^k, \max u^k] \subset [\min u^0, \max u^0].$$

Démonstration. Comme, pour tout x , les K_{xy} sont positifs et sont de somme égale à 1, l'équation (6.2) exprime que u_x^{k+1} est une combinaison convexe des u_y^k . On a donc

$$u_x^{k+1} \in [\min u^k, \max u^k].$$

La propriété s'ensuit par récurrence sur k . □

Le principe du maximum existe aussi sous forme statique, voir la proposition 6.8 ci après.

Remarque 6.4. La variable d'opinion u est non conservative, au sens où

$$\sum_{x \in V} u_x^k,$$

qui est l'opinion moyenne (au facteur $1/N$ près), est susceptible d'évoluer au cours des itérations. Considérons par exemple un réseau de N agents, avec un unique influenceur : $\Gamma = \{x\}$, et supposons que $V - \cdot \rightarrow \Gamma$, c'est à dire que tout agent est influencé (directement ou indirectement) par x . On suppose que l'opinion de x est 1, et l'opinion initiale de tous les autres est 0. L'opinion moyenne (si l'effectif est grand) est donc proche de 0, alors qu'elle converge 1 au cours de l'évolution (comme prouvé en toute généralité par la proposition 6.9 ci-après).

Reconnaissons que cet exemple n'est pas complètement convaincant car un système avec influenceur ne peut être considéré comme un système fermé évoluant librement. On peut voir l'influenceur comme un agent exerçant une influence extérieure, même si elle lui est propre, comme il n'y a pas de rétroaction avec le système, elle peut être vue comme une action extérieure (ce qu'elle est d'ailleurs effectivement si l'influenceur est considéré comme un point d'entrée vers le système par une "puissance" extérieure). Si l'on poursuit l'analogie avec la loi d'Ohm (voir remarque 6.2 ci-dessus), que l'on voit plutôt comme une loi de Fick pour la diffusion de la chaleur, l'influenceur peut être vu comme un point de la frontière en lequel on impose des conditions de Dirichlet. Et de fait, dans le cas d'un objet sur le bord duquel on impose la température, l'énergie thermique n'est pas conservée car le maintien de la température voulue nécessite un flux de chaleur rentrant ou sortant.

On considérera plutôt un réseau constitué d'un cycle et d'une branche connectée au cycle (graphe Lollipop de la figure 6.4, page 147). Si initialement l'opinion vaut 1 sur le cycle, et 0 ailleurs, elle va rester égale à 1 sur le cycle, et cette valeur 1 va se propager sur la branche. Si cette dernière est longue, la valeur moyenne va passer d'un valeur petite à la valeur 1.

Remarque 6.5. (Thermodynamics of opinion propagation)

Si l'on considère la somme des opinions comme une énergie, la remarque précédente indique une non-vérification du Premier Principe de la thermodynamique. If one sees the sum of opinion as an energy, the previous remark shows that the model does not obey the *First Law of Thermodynamics* : the total energy is not conserved. On the other hand, the Ohm-like law (see Remark 6.2) expresses that the opinion flows from higher values to smaller values, thereby tending to make the distribution more uniform by decreasing the maximum and increasing the minimum (Maximum Principle in Proposition 6.3). From this standpoint, the model can be said to obey the *Second Law of Thermodynamics*,

although a proper entropy may not always be defined, and convergence toward an equilibrium state may be ruled out. As we shall see in Section 6.2, the transpose problem will present a reverse structure (see in particular Diagram (6.8)).

Because of its hyper-academic character, we shall not consider in this section the non-diffusive case, i.e. the case of graphs which correspond to a mapping, with K_{xy} taking values in $\{0, 1\}$. Yet, for the sake of mathematical completeness, we develop this situation in Section 6.9.

Proposition 6.6. (Convergence vers un consensus)

On suppose² qu'il existe un entier p tel que, pour tous $x, y \in V$, il existe un chemin de longueur exactement p entre x et y (y compris pour $y = x$). Alors, pour toute condition initiale, la suite associée au problème (6.2) converge vers un consensus (champ uniforme sur V).

Démonstration. D'après le principe du maximum, le max décroît et le min croît. Comme la suite des max est minorée (par le min initial), elle converge vers une valeur M . De même la suite des min converge vers m . Si les deux valeurs sont égales, on a convergence vers le consensus correspondant. Si les valeurs sont distinctes, on extrait une sous-suite $u^{\varphi(k)}$ qui converge vers $w^0 \in \mathbb{R}^V$, avec $\max(w^0) = M$, w^0 non identiquement égal à M . Il existe donc au moins un y tel que $w_y^0 < M$.

On se place maintenant dans le cas $p = 1$, c'est-à-dire que le graphe est supposé *complet*. On considère le problème d'évolution associé à la condition initiale w^0 . On note (w^k) la suite correspondante. Comme le graphe est complet, chaque w_x^1 est combinaison convexe de valeurs $\leq M$, dont une au moins est $< M$, il est donc lui-même $< M$. On a donc $\max(w^1) < M$, d'où

$$\max(Ku^{\varphi(k)}) \rightarrow \max(Kw^0) = \max(w^1) < M,$$

ce qui implique que le max de $K^1 u^{\varphi(k)} = u^{\varphi(k)+1}$ est strictement inférieur à M , ce qui est absurde car ce max décroît vers M . Dans le cas général, $p > 1$, on remplace dans ce qui précède K par K^p .

Il est donc impossible que M soit différent de m , d'où la convergence du champ vers un consensus. \square

Remarque 6.7. Dès qu'il existe au moins une boucle dans le graphe, l'hypothèse de forte connexité suffit pour assurer l'hypothèse de la proposition précédente (voir exercice 6.1, page 148)

Problème d'évolution avec influenceurs

On suppose maintenant $\Gamma \neq \emptyset$, i.e. certains agents gardent leur opinion initiale constante. On se donne $U \in \mathbb{R}^\Gamma$, qui représente les opinions des influenceurs. On introduit le problème de Dirichlet :

$$\begin{cases} \sum_{y \leftarrow x} K_{xy}(u_x - u_y) = 0 & \forall x \in \mathring{V} \\ u_x = U_x & \forall x \in \Gamma. \end{cases} \quad (6.5)$$

On peut aussi écrire ce problème sous forme matricielle :

$$(\mathring{I} - \mathring{K})u = \mathring{A}u = b,$$

2. L'hypothèse qui suit s'exprime dans le contexte des chaînes de Markov à l'aide de la notion de matrice primitive. Une matrice carrée est dite primitive s'il existe une puissance de cette matrice dont tous les coefficients sont strictement positifs. On suppose donc ici que la matrice (K_{xy}) est primitive. La propriété énoncée ici peut être vue comme une conséquence du théorème de Perron-Frobenius appliqué à une telle matrice donne l'existence d'une valeur propre réelle strictement positive simple qui domine strictement les modules des autres valeurs propres. Nous en donnons néanmoins une démonstration qui n'utilise pas cette approche spectrale, démonstration qui peut s'étendre à des situations plus générales, en particulier non linéaires.

avec

$$(\mathring{A}u)_x = \sum_{y \leftarrow x, y \in \mathring{V}} K_{xy}(u_x - u_y) + \sum_{y \leftarrow x, y \in \Gamma} K_{xy}u_x, \quad b_x = \sum_{y \leftarrow x, y \in \Gamma} K_{xy}U_x.$$

Proposition 6.8. (Principe du maximum, version statique)

On suppose que $\Gamma \neq \emptyset$, et que tout $x \in V$ est connecté à Γ , i.e. $V - \cdot \rightarrow \Gamma$ (voir définition 10.13, page 226). Alors pour tout u solution de (6.5), le maximum et le minimum de u sont atteints sur Γ .

Démonstration. On considère un point x qui réalise le maximum de u . Pour tout y connecté à x , u_y réalise nécessairement aussi ce maximum. En suivant un chemin de x to Γ , on obtient que le maximum est aussi réalisé en au moins un point de Γ . On montre de la même manière que le minimum est atteint sur Γ . \square

Proposition 6.9. On suppose $\Gamma \neq \emptyset$ et $V - \cdot \rightarrow \Gamma$. Alors, pour toute donnée U sur Γ , le problème de Dirichlet (6.5) admet une solution unique $u = (u_x)$, avec

$$u_x \in [\min(U_x), \max(U_x)] \quad \forall x \in V.$$

De façon équivalente, l'équation de point fixe (6.3) admet une unique solution $u \in \mathbb{R}^V$ telle que $u_x = U_x$ pour tout $x \in \Gamma$.

Démonstration. Commençons par établir l'unicité. On considère deux solutions associées à la même condition U . Par linéarité, la différence u est solution du même problème avec des conditions homogènes sur le bord. Comme le maximum est atteint sur Γ , ce maximum est nul 0. Le minimum est nul de la même manière, et u est donc identiquement nul. Comme le problème consiste à résoudre un problème linéaire à N° (nombre de sommets intérieurs) équations et N° inconnues, l'unicité implique l'existence. La matrice est donc inversible, d'où l'on déduit l'existence d'une unique solution. Le fait que u_x soit dans l'enveloppe convexe des valeurs au bord pour tout x est conséquence directe du principe du maximum.

Pour l'existence on peut alternativement utiliser un argument topologique, qui s'applique à des situations plus générales (en particulier à des problèmes non linéaires). On introduit $\Lambda_a^b \subset \mathbb{R}^N$ l'ensemble des champs $u \in \mathbb{R}^V$ (identifié à \mathbb{R}^N), dont tous les éléments sont dans l'intervalle $[a, b] = [\min U, \max U]$, et tels que u s'identifie à U sur Γ . L'application

$$K : u \in \mathbb{R}^N \mapsto v \in \mathbb{R}^N, \quad v_x = \sum_{x \rightarrow y} K_{xy} u_y,$$

est continue et laisse Λ_a^b invariant. Comme Λ_a^b est convexe et compact, le théorème de Brouwer (voir par exemple Th. 6.6 dans [Border]³) assure l'existence d'un point fixe. \square

Remarque 6.10. Nous avons vu que la solution $u = (u_x)$ du problème de Dirichlet, qui correspond au point d'équilibre en termes de distribution d'opinion, est telle que les opinions u_x sont dans l'enveloppe convexe des valeurs au bord. Pour tout $x \in V$ il existe donc une collection de coefficients $(\mu_{xy})_{y \in V}$ dans $[0, 1]$ telle que

$$u_x = \sum_{y \in \Gamma} \mu_{xy} U_y, \quad \sum_{y \in \Gamma} \mu_{xy} = 1.$$

Pour chaque $y \in V$, μ_{xy} est le poids de U_y dans la combinaison barycentrique, il quantifie ainsi l'importance de y dans la formation de l'opinion de x . On peut ainsi voir (μ_{xy}) comme la matrice d'une application de \mathbb{R}^Γ dans \mathbb{R}^V . On peut aussi voir μ comme une application de V dans \mathbb{R}^Γ , plus précisément comme une collection de lois de probabilités sur Γ , indexée par les sommets du graphe : à tout x on associe $(\mu_{xy})_y \in \mathcal{P}(\Gamma)$. Pour x fixé, cette loi s'appelle la *mesure harmonique* associée à x (voir définition 6.17, basée sur une interprétation stochastique).

³ K. C. Border, *Fixed point theorems with applications to economics and game theory*, Cambridge university press, 1985.

Proposition 6.11. Sous les hypothèses de la proposition précédente, pour toute condition initiale u^0 , la suite des itérés converge vers l'unique point d'équilibre⁴, i.e. l'unique solution à (6.5) où U est la restriction de u^0 à Γ .

Démonstration. Soit u^{eq} l'unique solution de (6.5) avec la condition au bord $U = u_{|\Gamma}^0$. Soit $(u^k)_k$ la suite obtenue par itération de (6.2). La différence $e^k = u^k - u^{eq}$ vérifie la même équation d'évolution

$$e_x^{k+1} = \sum_{x \rightarrow y} K_{xy} e_y^k,$$

et s'annule sur Γ . D'après le principe du maximum, $\max(e^k)$ décroît, en restant positif (à cause de la condition au bord). Il converge donc vers une limite $\ell \geq 0$. Nous allons montrer que $\ell = 0$. Il sera alors possible de montrer de la même manière que $\min(e_k) \leq 0$ converge vers 0 par valeur inférieures, de telle sorte que (e^k) converge vers 0, et donc que (u^k) converge vers u^{eq} .

On raisonne par l'absurde : supposons $\ell > 0$. Comme (e^k) est bornée, elle admet une sous-suite $e^{\varphi(k)}$ qui converge vers $w^0 \in \mathbb{R}^V$, avec $\max(w^0) = \ell$.

Considérons maintenant le problème d'évolution associé à la condition initiale w^0 . On note (w^k) la suite correspondante. On cherche à montrer que $\max(w^k)$, qui vaut ℓ pour $k = 0$, diminue strictement au delà d'un certain temps. Le cas de figure est proche de celui de la démonstration de la proposition 6.6, avec une petite difficulté supplémentaire : comme les points ne sont plus nécessairement connectés à eux-mêmes, il est a priori possible qu'un point en lequel la valeur est $< \ell$ passe à la valeur ℓ à l'étape suivante (s'il est cerné de points qui réalisent le max). On définit X_k comme l'ensemble des sommets x qui sont connectés à la frontière Γ par un chemin constitué de sommets en lesquels u^k est $< \ell$.

Considérons un sommet x en lequel w^0 est ℓ . Ce sommet est connecté à la frontière par un chemin $x = x_0 \rightarrow x_1 \rightarrow \dots \rightarrow x_n \in \Gamma$. Considérons maintenant le plus grand indice p tel que $w_{x_p}^0 = \ell$. Comme x_p est connecté à x_{p+1} en lequel la valeur est strictement inférieure à ℓ , la valeur va décroître, i.e. $w_{x_p}^1 < \ell$, de telle sorte que $x_p \in \bar{X}_1$, alors qu'il n'était pas dans X_0 . Comme $X_0 \subset X_1$, X_1 contient au moins un sommet de plus que X_0 . De la même manière, tant que $X_k \neq V$, il est strictement inclus dans X_{k+1} . Comme le nombre de sommets est fini, il existe un indice m tel que X_k couvre V pour tout $k \geq m$. On a donc $\max(\bar{w}^m) < \ell$.

On note maintenant, comme dans la démonstration de la proposition 6.6, K^m l'application qui correspond à m itérations de la relation de récurrence, de telle sorte que $w^m = K^m w^0$. Cette application, linéaire, est continue, tout comme

$$w \mapsto \max(K^m w).$$

On a donc

$$\lim \max(K^m w^{\varphi(k)}) = \lim \max(w^{\varphi(k)+m}) = \max(K^m w^0) = \max(w^m) < \ell,$$

qui est une contradiction car $\max(w^n) \geq \ell$ pour tout $n \in \mathbb{N}$. \square

Dans le cas où certains points ne sont connectés à aucun influeur (i.e. aucun chemin de x vers Γ), la situation est plus compliquée. Il peut par exemple y avoir des cycles déconnectés de Γ , pouvant entraîner une solution asymptotiquement périodique.

Remarque 6.12. (Oubli)

Dans le cadre de la proposition précédente, on notera que le champ limite ne dépend que des valeurs initiales au bord : l'opinion des "influençables" (i.e. non strictement influenceurs) n'a aucune incidence sur la distribution d'opinions finale.

4. Noter que le point d'équilibre dépend de la condition initiale, plus précisément que la restriction à Γ de la condition initiale, du fait que les valeurs sur Γ ne changent jamais, joue le rôle d'une condition aux limites.

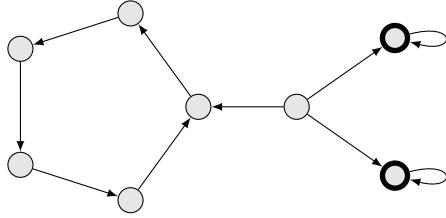


FIGURE 6.1 – Un sommet influencé par Γ (cercles en gras) et par un cycle.

Remarque 6.13. (Fonction de Lyapunov)

La preuve de la proposition (13.30) est une instantiation d'un principe général assurant la stabilité asymptotique, basé sur la notion de fonction de Lyapunov. Considérons un système dynamique dans \mathbb{R}^N défini par $u^{k+1} = T(u^k)$, où T est continue. On suppose que toutes les trajectoires sont bornées, et que le système admet un unique point fixe, i.e. un unique $u^{eq} \in \mathbb{R}^N$ vérifiant $u^{eq} = T(u^{eq})$. On suppose maintenant qu'il existe une fonctionnelle continue $F : \mathbb{R}^N \rightarrow \mathbb{R}$, qui admet u^{eq} comme unique minimiseur sur \mathbb{R}^N . On suppose de plus que F est strictement décroissante le long des trajectoires au sens suivant : pour tout $u^0 \in X \setminus \{u^{eq}\}$, $(F(u^k))$ est décroissante et il existe k tel que $F(u^k) < F(u^0)$. Alors la suite (u^k) converge vers u^{eq} quand k tend vers $+\infty$.

6.2 Cadre stochastique

Nous donnons ici une interprétation stochastique du problème de Dirichlet (6.3) étudié dans la section suivante. On se place dans le cas où $\Gamma \neq \emptyset$, et $V - \cdot \rightarrow \Gamma$ (voir definition. 10.13). On note $U \in \mathbb{R}^\Gamma$ une condition aux limites sur Γ .

Une marche⁵ peut être canoniquement définie sur le réseau (V, E, K) , avec une matrice de transition simplement égale à K : K_{xy} représente la probabilité d'aller de x à y . Pour tout x , on considère une marche aléatoire issue de x , selon les probabilités de transition (K_{zy}) , et l'on note j l'indice correspondant au premier passage sur Γ :

$$X_0 = x, X_1, \dots, X_j \in \Gamma, X_i \notin \Gamma \quad \forall i < j.$$

La valeur de U en X_j est une variable aléatoire associée au point de départ x . On note u_x son espérance.

Proposition 6.14. Le champ $u \in \mathbb{R}^V$ est la solution au problème de Dirichlet (6.5), qui s'écrit

$$\begin{cases} u_x - \sum_{x \rightarrow y} K_{xy} u_y &= 0 \quad \forall x \in \mathring{V}, \\ u_x &= U_x \quad \forall x \in \Gamma. \end{cases} \quad (6.6)$$

Démonstration. Notons en premier lieu que u s'identifie à U sur Γ (quand $x \in \Gamma$ alors $j = 0$). Par ailleurs, K_{xy} étant la probabilité d'aller de x à y , on a (espérance conditionnelle)

$$u_x = \sum_{x \rightarrow y} K_{xy} u_y,$$

de telle sorte que u vérifie (6.3). □

5. Cette section utilise des propriétés élémentaires des marches aléatoires, qui sont présentées ici de façon très informelle.

On note ρ_x^k la probabilité que la particule soit en x à l'étape k . La collection des lois (ρ^k) et les champs d'opinion (u^k) vérifient respectivement les relations de récurrence,

$$\rho_x^{k+1} = \sum_{y \rightarrow x} K_{yx} \rho_y^k \quad \text{et} \quad u_x^{k+1} = \sum_{x \rightarrow y} K_{xy} u_y^k. \quad (6.7)$$

Remarque 6.15. On notera que la propagation de u^k s'effectue dans la direction contraire. les deux problèmes d'évolution sont mutuellement adjoint : le premier (sur ρ , associé à K^T) implique une loi de probabilité, et vérifie une propriété de conservation. En effet ,

$$\sum_x \rho_x^{k+1} = \sum_x \sum_{y \rightarrow x} K_{yx} \rho_y^k = \sum_y \rho_y^k \underbrace{\sum_{y \rightarrow x}}_{=1} = \sum_y \rho_y^k.$$

La seconde (équation sur u associée à K) jouit d'une propriété de type principe du maximum (proposition 6.3). Cela fixe un cadre naturel en termes de normes pour ces deux matrices vues comme des opérateurs entre espaces de Banach : la norme naturelle pour u est la norme ℓ^∞ , alors que la norme naturelle pour ρ est ℓ^1 , comme illustré par le diagramme :

$$\begin{array}{ccc} (\mathbb{R}^N, \ell^\infty) & \xrightarrow{K} & (\mathbb{R}^N, \ell^\infty) \\ (\mathbb{R}^N, \ell^1) & \xleftarrow{K^T} & (\mathbb{R}^N, \ell^1), \end{array} \quad (6.8)$$

et les propriétés immédiates.

$$\|K\|_\infty = \sup_{\mathbb{R}^N} \frac{\|Ku\|_\infty}{\|u\|_\infty} = 1 = \sup_{\mathbb{R}^N} \frac{\|K^T \rho\|_1}{\|\rho\|_1} = \|K^T\|_1.$$

De ce point de vue, les propriétés de principe du maximum et de conservation peuvent être considérées comme mutuellement adjointes.

Remarque 6.16. Les variables associées u et ν sont des archétypes de variables respectivement *intensive* et *extensive* (voir section 1.2, page 17), ce que l'on peut exprimer grossièrement : ça n'a pas de sens de sommer deux valeurs d'opinion, mais cela a du sens de sommer des valeurs de ν sur un ensemble $X \subset V$ (cela donne la probabilité que la particule soit dans X).

On considérera dans la section 6.6 où les matrices sont autoadjointes dans un certain sens (i.e. pour un certain produit scalaire), ce qui confère au problème d'évolution un principe du maximum et un principe de conservation, qui est la préservation de la moyenne pondérée de l'opinion selon une certaine mesure, que nous appellerons *charisme*, et qui joue le rôle d'une mesure stationnaire pour la marche aléatoire considérée (voir la proposition 6.37 , ou directement l'équation (6.24)).

L'interprétation probabiliste proposée nous conduit naturellement à la notion de *mesure harmonique*.

Definition 6.17. (Mesure harmonique)

Soit $x_0 \in V$ un sommet. La mesure harmonique de pôle x_0 , notée $\mu_{x_0, \cdot}$, est une mesure de probabilité sur Γ telle que $\mu_{x_0, y}$ est la probabilité que la marche aléatoire issue de x_0 , avec matrice de transition K , rencontre Γ pour la première fois en y . Cette mesure peut être étendue à V en posant simplement $\mu_{x_0, y} = 0$ pour $y \notin \Gamma$.

Proposition 6.18. Soit $U \in \mathbb{R}^\Gamma$ et $u \in \mathbb{R}^V$ la solution unique du problème de Dirichlet (voir proposition 6.9). Alors, pour tout $x_0 \in V$,

$$u_{x_0} = \sum_{y \in \Gamma} \mu_{x_0, y} U_y,$$

où $\mu_{x_0, y}$ est la mesure harmonique de pôle x_0 (voir définition 6.17).

Démonstration. On utilise la proposition 6.14. On sait que u_{x_0} est l'espérance de U_Y , où Y est le point de première rencontre avec Γ pour une marche aléatoire issue de x_0 . On a donc

$$u_{x_0} = \mathbb{E}[U_Y] = \sum_{y \in \Gamma} U_y \underbrace{P(Y = y)}_{=\mu_{x_0, y}}. \quad \square$$

La valeur u_{x_0} est donc une combinaison convexe des valeurs de U sur le bord, dont les coefficients sont les $(\mu_{x_0,y})_{y \in \Gamma}$. En particulier, $\mu_{x_0,y}$ est l'opinion de x_0 associée à l'opinion sur le bord $U = \delta_{y,\cdot}$, i.e. y pense 1, et les autres influenceurs pensent 0.

Proposition 6.19. Soit x_0 fixé. On considère le problème linéaire :

$$\begin{cases} \nu_y - \sum_{x \rightarrow y} K_{xy} \nu_y = \delta_{x_0,y} & \text{pour tout } y \in \mathring{V} = V \setminus \Gamma, \\ \nu_x = 0 & \text{pour } x \in \Gamma. \end{cases} \quad (6.9)$$

où $\delta_{\cdot,\cdot}$ est le symbole de Kronecker, i.e. $\delta_{x_0,y} = 1$ si $y = x_0$, et 0 sinon.

Ce problème admet une solution unique ν , et l'on a

$$\mu_{x_0,y} = -(\nu - K^T \nu)_y = -(\nu_y - \sum_{x \rightarrow y} K_{xy} \nu_x) \quad \forall y \in \Gamma. \quad (6.10)$$

Démonstration. L'équation vérifiée par ν est un système linéaire, dont la matrice est la transposée de celle du problème de Dirichlet (6.5). D'après la proposition 6.9, on sait que ce problème est bien posé, la matrice est donc non singulière. On fixe maintenant $y_0 \in \Gamma$. Soit u l'unique solution au problème

$$\begin{cases} u_x - \sum_{y \rightarrow x} K_{xy} u_y = 0 & \text{pour tout } x \in \mathring{V}, \\ u_{y_0} = 1, \\ u_y = 0 & \text{pour } y \in \Gamma, y \neq y_0. \end{cases}$$

Ainsi défini, u n'est autre que la solution du problème de Dirichlet décrite par la proposition 6.9 avec $\delta_{y_0,\cdot}$ comme donnée au bord. On a donc en particulier $\mu_{x_0,y_0} = u_{x_0}$ d'après la proposition 6.18.

Comme ν est nulle sur Γ , et que u est harmonique sur \mathring{V}

$$\sum_{x \in V} \nu_x \underbrace{\left(u_x - \sum_{y \rightarrow x} K_{xy} u_y \right)}_{=0 \text{ sur } \mathring{V}} = 0.$$

Cette quantité peut s'écrire

$$\begin{aligned} 0 &= \langle u - Ku | \nu \rangle = \langle u | \nu - K^T \nu \rangle = \sum_{y \in V} \underbrace{u_y}_{=\delta_{y_0,y} \text{ sur } \Gamma} \underbrace{\left(\nu_y - \sum_{x \rightarrow y} K_{xy} \nu_x \right)}_{=\delta_{x_0,y} \text{ sur } \mathring{V}}. \\ &= \nu_{y_0} - \sum_{x \rightarrow y_0} K_{xy_0} \nu_x + u_{x_0}, \end{aligned}$$

d'où

$$u_{x_0} = \mu_{x_0,y_0} = - \left(\nu_{y_0} - \sum_{x \rightarrow y_0} K_{xy_0} \nu_x \right),$$

qui est la propriété annoncée. \square

Remarque 6.20. La proposition 6.19 est l'analogie discret, dans le cas non symétrique, d'une propriété générale portant sur les mesures harmoniques associées à un domaine euclidien. On considère un domaine régulier $\Omega \subset \mathbb{R}^d$ de frontière $\Gamma = \partial\Omega$, $x_0 \in \Omega$, et l'on note ν la solution du problème de Poisson avec second membre singulier

$$\begin{cases} -\Delta \nu = \delta_{x_0} & \text{in } \Omega \\ \nu = 0 & \text{on } \Gamma. \end{cases}$$

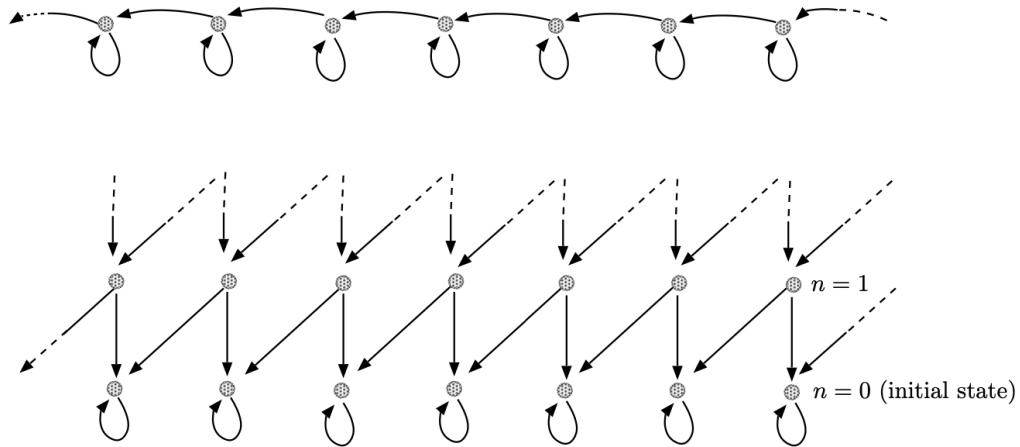


FIGURE 6.2 – Upwind scheme graphs : space graph (top) and full space-time graph (bottom).

Le champ $-\partial\nu/\partial n$ défini sur $\partial\Omega$, qui est la version continue de (6.10), est aussi la densité de probabilité du point de premier passage sur la frontière $\partial\Omega$ pour un mouvement brownien issu de x_0 . Le problème de Dirichlet ci-dessus est l'analogue continu du problème discret (6.9). Dans le même esprit, la proposition 6.14 est la contrepartie discrète d'une propriétés standard en théorie du potentiel. On considère un champ scalaire U défini sur la frontière Γ , et le problème de Laplace associé

$$\begin{cases} -\Delta u &= 0 \quad \text{dans } \Omega \\ u &= U \quad \text{sur } \Gamma, \end{cases}$$

qui est l'analogue continu de (6.6). For any $x \in \Omega$, consider the Brownian motion issued from x , denote by $X \in \Gamma$ the location of its first contact with Γ . The expected value of $U(X)$ is $u(x)$, where u is the solution to the Laplace problem above (see again [?], p. 651). Note that, in the discrete, non-symmetric setting, the two problems are natively set on different spaces in terms of norms : Problem (6.6) is set in $(\mathbb{R}^N, \ell^\infty)$, whereas Problem (6.9) is set in (\mathbb{R}^N, ℓ^1) (see Diagram (6.8)).

6.3 Liens avec les schémas de discréétisation des EDP

We describe the discrete dynamical systems which come from the space and time discretization of standard Partial Differential Equations. This section is independent of the rest, it aims at underlying the fact that evolution problems of the type (6.2) can be seen as discrete generalization of standard PDE's. In the one-dimensional setting, we shall consider the transport equation at constant velocity $U > 0$

$$\partial_t u + U \partial_x u = 0, \quad (6.11)$$

and the heat equation

$$\partial_t u - D \partial_{xx} u = 0, \quad (6.12)$$

where $D > 0$ is the diffusion coefficient. The space-time variable (x, t) belongs to $[0, L] \times [0, T]$. We shall assume periodic boundary conditions in space, i.e. L is identified to 0. The Finite Difference space-time discretization of such equations is based on a time step $\Delta t = T/N_T$ and space step $\Delta x = L/N$. We shall denote by u_j^k the approximation of a solution to one of those equations at $(j\Delta x, k\Delta t)$. Given an initial condition u^0 , we build an initial discrete approximation according to $u_j^0 = u^0(j\Delta x)$. The so-call upwind scheme for (6.11) writes

$$\frac{u_j^{k+1} - u_j^k}{\Delta t} + U \frac{u_j^k - u_{j-1}^k}{\Delta x} = 0 \quad j = 1, \dots, N \equiv 0, \quad k = 0, \dots, N_T - 1.$$

The previous scheme can be written

$$u_j^{k+1} = \underbrace{\left(1 - \frac{U\Delta t}{\Delta x}\right)}_{=K_{jj}} u_j^k + \underbrace{\frac{U\Delta t}{\Delta x}}_{=K_{j,j-1}} u_{j-1}^k, \quad (6.13)$$

which fits into the framework described at the beginning of Section 6.1 (Equation 6.2), as soon as the so-called CFL condition $\theta = U\Delta t/\Delta x \leq 1$ is satisfied. Vertices identify here with space discretization points, and the underlying graph is represented in Figure 6.2 (top). Note that for $\theta = 1$, we recover the one-to-one and onto mapping situation. For $\theta \in (0, 1)$, the problem fits in the assumptions of Corollary ?? (for $m = N$), thus u^k converges to a uniform field. This result may appear as conflicting with pure transport phenomena, which do not change the profile. It is a consequence of the so-called *numerical diffusion*, which is a consequence of space discretization. Note that the set of harmonic fields is the line spanned by the uniform field $(1, \dots, 1)$.

For the heat equation, the so-called explicit scheme writes

$$\frac{u_j^{k+1} - u_j^k}{\Delta t} + D \frac{-u_{j-1}^k + 2u_j^k - u_{j+1}^k}{(\Delta x)^2} = 0 \quad j = 1, \dots, N \equiv 0, \quad k = 0, \dots, N_T - 1,$$

which can be written

$$u_j^{k+1} = \underbrace{\left(1 - \frac{2D\Delta t}{(\Delta x)^2}\right)}_{=K_{jj}} u_j^k + \underbrace{\frac{D\Delta t}{(\Delta x)^2}}_{=K_{j,j-1}} u_{j-1}^k + \underbrace{\frac{D\Delta t}{(\Delta x)^2}}_{=K_{j,j+1}} u_{j+1}^k. \quad (6.14)$$

Under the stability condition $2D\Delta t/Dx^2 \leq 1$, this expresses again an evolution of the type (6.2). Note that the network associated to the heat equation (discretization scheme (6.14)) is charismatic (according to Definition 6.30), whereas the transport network (associated to the discretization scheme (6.13)) is not.

Space-time graph

Now consider the space-time discrete field (u_j^n) obtained by one of the previous schemes, for the transport equation or the heat equation. We replace the index k by n because the time index no longer plays the role of an incremental evolution index. It can be seen as a discrete field over the space-time grid

$$V = \{(k, j), \quad 0 = 1, \dots, N, \quad k = 0, \dots, N_T\}.$$

In this setting, the topology of the graph is represented in Figure 6.2 (bottom), and the vertices corresponding to the initial states, at which the values are prescribed, play the role of *influencers*. The full discrete field can thus be considered as *harmonic* in the sense of (6.1), over the space-time grid. For a given grid, there is obviously a unique harmonic field in this sense for any collection of initial values (i.e. values at influencers), so harmonic fields form a linear space of dimension N .

6.4 Monitoring of a network through influencers / influence coefficients

We describe in this section how the so-called adjoint method can be carried out to investigate the dependence of some global quantities upon some parameters, called in this context *control variables*. We shall consider here two options for the choice of control variables : opinions of influencers, and influence coefficients K_{xy} . We consider as previously a set V of agent, a collection of influence coefficients (K_{xy}) encoded in a stochastic matrix K . We recall that influencers are agents x such that $K_{xx} = 1$, that

Γ denotes the set of influencers, and that $\mathring{V} = V \setminus \Gamma$ is the set of interior vertices. We assume that $V - \cdot \rightarrow \Gamma$, i.e. each vertex is connected by a path to the boundary (voir définition 10.13, page 226).

The core of the approach is the so-called *state equation*, which is in our case the static Dirichlet problem (6.5). We shall reformulate it as a linear system on the unknowns associated to the interior vertices. We denote by $U \in \mathbb{R}^\Gamma$ the collection of influencers' opinions, and by $\dot{u} \in \mathbb{R}^{\mathring{V}}$ the collection of unknowns (opinions at interior nodes), so that $u = (\dot{u}, U) \in \mathbb{R}^V$ is the full collection of opinions. The matrix formulation of (6.5) is $Au = 0$, with $A = I - K$, and $u_x = U_x$ for any $x \in \Gamma$. It will be more convenient here to separate interior degrees of freedom and influencers' opinions. The problem can be written

$$\dot{u}_x - \sum_{x \rightarrow y \in \mathring{V}} K_{xy} \dot{u}_y = \sum_{x \rightarrow y \in \Gamma} K_{xy} U_y \quad \forall x \in \mathring{V},$$

which can be written in a matrix form

$$\mathring{A}\dot{u} = K_\Gamma U, \tag{6.15}$$

where \mathring{A} is a submatrix of A (restriction to interior degrees of freedom). The matrix K_Γ , which encodes a mapping from \mathbb{R}^Γ to $\mathbb{R}^{\mathring{V}}$, integrates the effect of influencers' opinions. For any $U \in \mathbb{R}^\Gamma$, Problem (6.15) admits a unique solution (Proposition 6.9), which we denote by \dot{u}_U . We introduce the so-called *cost function*⁶, that is a function of U given as an expression of \dot{u}_U , i.e.

$$\Phi(U) = J(\dot{u}_U),$$

where $J(\dot{u})$ is typically a function which one aims at minimizing. To fix the idea, we may consider that Φ is defined as

$$J(\dot{u}) = \frac{1}{2} |\dot{u} - \hat{u}|^2, \tag{6.16}$$

where $\hat{u} \in \mathbb{R}^{\mathring{V}}$ is a targeted collection of opinions. Minimizing $J(\dot{u}_U)$ with respect to U consists in finding the collection of influencers' opinions such that the associated opinions are the closest to \hat{u} . An essential ingredient to actually perform this minimization is the gradient of this functional with respect to U . We shall now describe a classical method to estimate this gradient. Note that, beyond the actual use of this gradient in a minimization procedure, it can be also useful in terms of modeling to actually estimate the sensitivity of a functional (possibly the opinion of a specific agent, see Remak 6.22), with respect to influencers' opinions, or any other parameter.

Description of the adjoint method

The approach is based on a so-called *Lagrangian* L defined as

$$L : (\dot{u}, U, p) \in \mathbb{R}^{\mathring{V}} \times \mathbb{R}^\Gamma \times \mathbb{R}^{\mathring{V}} \mapsto J(\dot{u}) + \langle \mathring{A}\dot{u} - K_\Gamma U, p \rangle.$$

It holds that, for any $p \in \mathbb{R}^{\mathring{V}}$,

$$\Phi(U) = L(\dot{u}_U, U, p).$$

As a consequence

$$D\Phi(U) = D_{\dot{u}} L(\dot{u}_U, U, p) \circ D_U \dot{u}_U + D_U L(\dot{u}_U, U, p).$$

(We underline here that $D_U L(\dot{u}_U, U, p)$ stands for the partial derivative of $L(\cdot, \cdot, \cdot)$ with respect to the second variable U , taken at (\dot{u}_U, U, p) , \dot{u}_U being *frozen* at its current value.)

The strategy consists in building a p such that $D_{\dot{u}} L(\dot{u}_U, U, p) = 0$ in the expression above. For such a p , $D\Phi(U)$ is simply $D_U L$, which depends on this very p . From the expression of the Lagrangian, we have that⁷

$$\langle D_{\dot{u}} L, \delta \dot{u} \rangle = \langle D_{\dot{u}} J, \delta \dot{u} \rangle + \langle \mathring{A} \delta \dot{u} | p \rangle = \langle \nabla \Phi + \mathring{A}^T p | \delta \dot{u} \rangle.$$

6. Also called *loss* or *objective* function in the context of Machine Learning.

The *adjoint problem* for the dual variable p is defined in order to vanish the quantity above. The problem writes

$$\mathring{A}^T p = -\nabla J \quad (6.17)$$

Let p be the solution to this problem, it holds that

$$\langle D\Phi(U), \delta U \rangle = \langle D_U L(\dot{u}_U, U, p), \delta U \rangle = -\langle K_\Gamma \delta U, p \rangle = -\langle K_\Gamma^T p, \delta U \rangle.$$

As a consequence, $\nabla\Phi(U) = -K_\Gamma^T p$, where p is the solution to the adjoint problem (6.17).

In the case of a cost function defined by (6.16), the right-hand side of (6.17) is $\hat{u} - \dot{u}_U$. In the case of a cost function that is the value of u at some point x_0 , the right-hand side is the discrete Dirac mass $-\delta_{x_0}$ (see Remark 6.22 below).

Remarque 6.21. The adjoint problem can be interpreted as a discrete Laplace problem (based on the matrix adjoint to the non-symmetric Laplacian $I - K$), with homogeneous Dirichlet boundary conditions. Indeed, Problem (6.17) expresses

$$p_y - \sum_{y \leftarrow x \in \mathring{V}} K_{yx} p_x = (\nabla\Phi)_x \quad \forall y \in \mathring{V},$$

and the sum over vertices can be extended to V if one sets p_x to 0 for any $x \in \Gamma$.

Remarque 6.22. This approach makes it possible to recover the notion of harmonic measure introduced in Section 6.2. The idea consists in choosing another functional, that is $\Phi(u) = u_{x_0}$ for a certain $x_0 \in \mathring{V}$. The adjoint problem reads $\mathring{A}^T p = -\delta_{x_0}$, which is another way to express (6.9). The gradient of $J(U) = (u_U)_{x_0}$ is then a vector defined on Γ , the entries of which are the weights of the opinions of the various influencers upon x_0 , i.e. the values of the harmonic measure $\mu_{x_0,y}$, for $y \in \Gamma$.

Effective minimization of the cost functional

Considering a quadratic functional of the type $\Phi(U) = J(\dot{u}_U)$, where J is given by (6.16), minimizing $\Phi(U)$ amounts to find a $U \in \mathbb{R}^\Gamma$ at which the gradient is 0. A straightforward computation leads to the following matrix formulation of this condition :

$$(\mathring{A}^{-1} K_\Gamma)^T \mathring{A}^{-1} K_\Gamma U = -(\mathring{A}^{-1} K_\Gamma)^T \hat{u}.$$

Note that the matrix $S = (\mathring{A}^{-1} K_\Gamma)^T \mathring{A}^{-1} K_\Gamma U$ is symmetric definite positive. Indeed, it is of the form $B^T B$, therefore symmetric and nonnegative. Besides, $SU = 0$ implies $BU = 0$, which means that the solution to Problem (6.15) (which is another way to write Dirichlet problem (6.5)) is zero, so that $U = 0$. A Conjugate Gradient Method can then be implemented to solve this problem.

Derivative with respect to influence coefficients

We consider now the problem of determining the derivative of some function of the interior opinions with respect to the influence coefficients K_{xy} . This field can be identified to a vector in \mathbb{R}^E , where $E = \text{supp}(K_{xy})$ is the set of edges. We shall assume that influencers remain influencers, i.e. K_{yy} remains equal to 1 for any $y \in \Gamma$, and variations are restricted to connections between interior vertices and their neighbors (possibly influencers). We shall consider variations of coefficients on existing edges, but the approach may be extended to variations on all possible links.

7. We use the notation $\langle \cdot, \cdot \rangle$ to represent the duality pairing between the differential of a mapping and a variation of the considered variable, whereas $\langle \cdot | \cdot \rangle$ still represents the scalar product between two elements of the same Euclidean space, $\mathbb{R}^{\mathring{V}}$ or \mathbb{R}^Γ .

Since the vector of influencers' opinions is no longer a variable, we shall drop the dependence of \dot{u} upon U , and replace it by a dependence upon K . We write the state equation as

$$\mathring{A}_K \dot{u}_K = b_K, \quad \text{with } b_K = K_\Gamma U \in \mathbb{R}^{\mathring{V}}, \quad \text{i.e. } (b_K)_x = \sum_{x \rightarrow y \in \Gamma} K_{xy} U_y \quad \forall x \in \mathring{V}, x \rightarrow \Gamma,$$

with $(b_K)_x = 0$ for any all vertices which are not directly connected to Γ . The Lagrangian now writes

$$L : (\dot{u}, K, p) \in \mathbb{R}^{\mathring{V}} \times \mathbb{R}^E \times \mathbb{R}^{\mathring{V}} \mapsto J(\dot{u}) + \langle \mathring{A}_K \dot{u} - b_K, p \rangle.$$

It can be used, as previously, to estimate the gradient of $K \mapsto \Phi(K) = J(\dot{u}_K)$. The adjoint problem is the same as previously, i.e. Equation (6.17). It remains to estimate the derivative of L with respect to K . The part of L which depends on K writes

$$\sum_{x \in \mathring{V}} p_x \sum_{x \rightarrow y} K_{xy} (u_x - u_y) = \sum_{e \in \mathring{E}} \left(\sum_{x \rightarrow y} p_x (u_x - u_y) \right) K_{xy}$$

where u_x is U_x whenever $x \in \Gamma$, and \mathring{E} is the set of edges except influencers' loops. We finally obtain

$$\nabla \Phi(K) = (p_x (u_x - u_y))_{(x,y) \in \mathring{E}},$$

where p is the solution to the adjoint problem (6.17).

N.B. : Implementing this approach to actually minimize the cost function necessitates to account for the constraints on the K_{xy} 's : linear constraints on the sums of coefficients on each row, and unilateral constraints $K_{xy} \in [0, 1]$. Notice also that the problem is quite different from the previous one, with influencers' opinions as control variables, because the mapping $K \mapsto u_K$ is not linear, and consequently the global functional $J(\cdot)$ is not quadratic.

6.5 Modèle continu et structure de flot de gradient

Le fait que cela prenne un certain temps pour x d'absorber l'influence de ses voisins peut être modélisé en introduisant un paramètre d'inertie $\theta \in [0, 1]$, et en écrivant le modèle relaxé

$$u_x^{k+1} = (1 - \theta) u_x^k + \theta \sum_{x \rightarrow y} K_{xy} u_y^k. \quad (6.18)$$

Pour $\theta = 1$, on retrouve le problème discret. Noter que ce nouveau problème rentre dans le cadre discret précédent, en introduisant les paramètres modifiés

$$K'_{xx} = (1 - \theta) + \theta K_{xx}, \quad K'_{xy} = \theta K_{xy} \quad \text{for } y \neq x.$$

Équation différentielle

On considère que θ s'écrit ε/η , où η est un temps de relaxation fixe (temps typique de propagation de l'influence), et ε un petit paramètre (également homogène à un temps). On a

$$\frac{u_x^{k+1} - u_x^k}{\varepsilon} = \frac{1}{\eta} \left(\sum_{x \rightarrow y} K_{xy} (u_y^k - u_x^k) \right).$$

L'évolution prend la forme de la discrétisation en temps d'une système d'équations différentielles ordinaires pour des quantités $t \mapsto u_x^t \in \mathbb{R}$ qui varient continûment en temps, et vérifient le système d'équations

$$\frac{d}{dt} u_x = \frac{1}{\eta} \left(\sum_{x \rightarrow y} K_{xy} (u_y - u_x) \right).$$

Le problème continu en temps s'écrit donc

$$\frac{du}{dt} = -\frac{1}{\eta} Au \quad \text{avec} \quad A = I - K. \quad (6.19)$$

Stabilité

D'après le théorème de Gershgorin, le spectre de K est dans l'union des disques fermés $D_x \in \mathbb{C}$, centrés en K_{xx} et de rayon égal à la somme des coefficients extradiagonaux :

$$\text{Sp}(K) \subset \bigcup_{x \in V} D_x, \quad D_x = \left\{ z \in \mathbb{C}, |z - K_{xx}| \leq \sum_{y \neq x} K_{xy} \right\} \quad (6.20)$$

Comme $K_{xy} \in [0, 1]$ et $\sum_y K_{xy} = 1$, tous les D_x sont dans le disque unité, de telle sorte que le spectre de $A = I - K$ est dans le disque fermé de centre 1 et de rayon 1. Comme K est stochastique, $Ke = e$, avec $e = (1, 1, \dots, 1)$. La matrice A admet donc 0 comme valeur propre (avec multiplicité potentielle). En conséquence, les valeurs propres de A autres que la valeur propre nulle ont une partie réelle strictement positive, tout point d'équilibre est donc *stable*.

Lien entre les modèles discret et continu

On considère le problème d'évolution continu (6.19). Pour toute condition initiale u^0 , la solution s'exprime à l'aide de l'exponentielle de la matrice A :

$$\begin{aligned} u^t &= \exp\left(-\frac{t}{\eta}A\right)u^0 = \exp\left(-\frac{t}{\eta}\text{Id} + \frac{t}{\eta}K\right)u^0 \\ &= e^{-t/\eta} \sum_{k=0}^{+\infty} \left(\frac{t}{\eta}\right)^k \frac{K^k}{k!} u^0 = \sum_{k=0}^{+\infty} a_k(t) u^k, \end{aligned} \quad (6.21)$$

où $u^k = K^k u^0$ est la solution du problème d'évolution discret (6.2), et

$$a_k(t) = \frac{1}{k!} e^{-t/\eta} \left(\frac{t}{\eta}\right)^k.$$

La solution exacte au temps t est ainsi une combinaison barycentrique infinie des itérés discrets (u^k) , dont les poids suivent une distribution de Poisson de paramètre t/η (on se reportera à l'exercice (6.3), page 148, pour plus de détails).

Flots de gradient

On se propose de caractériser ici les cas où le problème d'évolution (6.19) a une structure de flot de gradient pour un certain produit scalaire.

Le cas où l'on se restreint au produit scalaire euclidien canonique est immédiat :

Proposition 6.23. Le problème 6.19 a une structure de flot de gradient pour le produit scalaire canonique, i.e. il existe une fonctionnelle Ψ deux fois continûment différentiable telle que $Au = \nabla\Psi(u)$, si et seulement si A est symétrique.

Démonstration. Si A est symétrique on a $Au = \nabla\Psi(u)$ avec

$$\Psi(u) = \frac{1}{2}\langle u | u \rangle.$$

Si $Au = \nabla\Psi(u)$, alors $a_{ij} = \partial^2\Psi/\partial x_i \partial x_j = \partial^2\Psi/\partial x_j \partial x_i = a_{ji}$. □

Plus généralement, si l'on considère une matrice M symétrique définie positive, on note $\langle \cdot | \cdot \rangle_M$ le produit scalaire associé, i.e.

$$\langle u | v \rangle_M = \langle Mu | v \rangle.$$

Pour toute fonctionnelle $\Phi : \mathbb{R}^N \rightarrow \mathbb{R}$ continûment différentiable, on note $\nabla_M \Phi(u) \in \mathbb{R}^N$ son gradient en u selon le produit scalaire associé à M , c'est à dire le vecteur tel que

$$\Phi(u + h) = \Phi(u) + \langle \nabla_M \Phi | h \rangle_M + o(h) = \Phi(u) + \langle M \nabla_M \Phi | h \rangle + o(h).$$

On a donc $\nabla_M \Phi = M^{-1} \nabla \Phi$, ce qui permet d'énoncer une première caractérisation des flots de gradient.

Proposition 6.24. Le problème 6.19 a une structure de flot de gradient pour le produit scalaire associé à la matrice s.d.p. M , i.e. il existe une fonctionnelle Ψ deux fois continûment différentiable telle que $Au = \nabla_M \Psi(u)$, si et seulement A s'écrit $M^{-1}H$, où H est une matrice symétrique.

Démonstration. Si $A = M^{-1}H$, alors $Au = \nabla_M \Psi(u)$ avec $\Psi(u) = \frac{1}{2} \langle Hu | u \rangle$. Si $Au = \nabla_M \Psi(u)$ alors $MAu = \nabla \Psi(u)$, d'où, en différentiant cette identité, $MA = H$, où H est la matrice hessienne de Ψ , donc symétrique. \square

Proposition 6.25. Le problème (6.19) a une structure de flot de gradient pour un certain produit scalaire si et seulement si A (ou, de façon équivalente, K) est diagonalisable, et ses valeurs propres sont réelles.

Démonstration. Si Au est le gradient d'un fonctionnelle quadratique en u pour un produit scalaire $\langle \cdot | \cdot \rangle_M$, il existe une matrice symétrique H telle que $A = M^{-1}H$. Comme M est s.d.p., elle s'écrit $M = UDU^{-1}$, où U est une matrice orthogonale et D est diagonale. On définit alors $M^{1/2}$ comme $UD^{1/2}U^{-1}$. La matrice A est semblable à

$$M^{1/2}AM^{-1/2} = M^{1/2}M^{-1}HM^{-1/2} = M^{-1/2}HM^{-1/2},$$

qui est symétrique car $M^{-1/2}$ et H le sont. La matrice A est donc semblable à une matrice symétrique, elle est donc diagonalisable de valeurs propres réelles.

On suppose maintenant que A est diagonalisable, de valeurs propres réelles : $A = PDP^{-1}$ où D est diagonale réelle, et P est une matrice inversible. On écrit

$$A = PDP^{-1} = PP^T P^{-T} DP^{-1} = M^{-1}H,$$

où $M = (PP^T)^{-1} = P^{-T}P^{-1}$ est une matrice symétrique définie positive⁸, et $H = P^{-T}DP^{-1}$ est symétrique réelle. \square

Remarque 6.26. Dans la section 6.6, nous aborderons un cas particulier de systèmes présentant une structure de gradient, il s'agit que réseaux que nous appellerons *charismatiques* (voir la définition 6.30). Pour de tels réseaux, on aura $K = M^{-1}H$, où M est une matrice *diagonale*. Les coefficients diagonaux de M correspondent aux charismes $(m_x)_{x \in V}$ des agents. D'un point de vue probabiliste, cette situation correspond au cas d'une chaîne de Markov *réversible*. Dans le cas où l'on a une structure de flot de gradient pour une métrique M quelconque (non diagonale), le sens des coefficients de M est moins clair. On peut toutefois effectuer un changement de variable pour se ramener au cas diagonal.

On a

$$M \frac{du}{dt} = -\frac{1}{\eta} Hu \text{ avec } M = UDU^T, \quad U \text{ orthogonal },$$

d'où, en effectuant le changement de variable $u = Uv$,

$$UDU^T U \frac{dv}{dt} = -\frac{1}{\eta} H U v \implies D \frac{dv}{dt} = -\frac{1}{\eta} U^T H U v,$$

8. Ce produit scalaire fait de la famille des vecteurs colonnes de P une base orthonormée.

avec D diagonale (coefficients > 0) et $H' = U^T H U$ symétrique. On fait donc apparaître une collection de N poids strictement positifs, mais ces poids sont afférents aux composantes de v , qui sont des combinaisons linéaires des u_x (les coefficients sont les colonnes de U), il ne s'agit donc pas d'un "charisme" individuels, mais d'un charismes relatifs à des combinaisons d'opinions. Il n'y a pas d'interprétation évidentes de ces nouvelles variables, d'autant que les coefficients de ces combinaisons peuvent être négatifs.

Situations non-gradient typiques

En général, le problème d'évolution (6.19) n'est un flot de gradient pour aucun produit scalaire. Nous décrivons ci-dessous deux situation typiques pour lesquelles on n'a pas cette structure de flots de gradient : (i) la matrice A a des valeurs propres réelles, mais n'est pas diagonalisable ; (ii) la matrice A est diagonalisable, mais ses valeurs propres sont imaginaires non réelles.

Graphes hiérarchiques et matrices non diagonalisables

On considère ici des graphes acycliques, i.e. des graphes sans cycles (voir définition 10.13). Such graphs are called *hierarchical*, because it can be shown that there exists an ordering x_1, x_2, \dots, x_N of the vertices in such a way that $x_i \rightarrow x_j$ implies $i \leq j$. Such an indexing, which is obviously not unique in general, can be built by mean of topological sort. To simplify the situation, we shall assume that $K_{xx} > 0$ only for the influencers (and in this case $K_{xx} = 1$).

With these assumptions, under the chosen ordering of the vertices, the matrix K is upper-triangular, with only 1 and 0 on the diagonal :

$$K = \left(\begin{array}{c|c} B & H \\ \hline 0 & I_m \end{array} \right),$$

where m is the number of influencers, I_m is the identity matrix of size m , B is a $(N - m) \times (N - m)$ upper triangular matrix with only zeros on the diagonal, and H is a $(N - m) \times m$ matrix. Note that B is nilpotent. The matrix K has only two eigenvalues, namely 1 and 0, but the kernel of K , i.e. the eigenspace associated to the eigenvalue 0, has a dimension strictly smaller than $N - m$. Note that, in the archetypal situation of a linear network $x_1 \rightarrow x_2 \rightarrow \dots \rightarrow x_N \rightarrow x_N$, we have that $m = 1$, B is the $(N - 1)$ square matrix with only zeros except for the first upper diagonal filled with 1, and H is a column matrix filled with 0, except for the last entry which is 1. The matrix K then natively identifies with its Jordan normal form.

Let us first consider the discrete model, i.e. the iterative process process (6.2). After at most $N - m$ iterations, the fixed point state will be reached and the opinion of every agent but the influencer will have been flushed out of the network. Indeed, as $B^{N-m} = 0$ (because B is strictly upper triangular), a straightforward induction leads to

$$K^{N-m} = \left(\begin{array}{c|c} 0 & C \\ \hline 0 & I_m \end{array} \right), \quad \text{with } C = \sum_{k=0}^{N-m-1} B^k H,$$

and $K^n = K^{N-m}$ for all $n \geq N - m$. If one denotes by $u^0, u^1, \dots, u^k = K^k u^0, \dots$, the successive opinion fields, the opinions of influencers propagates along the branches of the tree during the first iterates 1, ..., up to $N - m$ at most, and it is then frozen to the fixed point for larger values of k .

Now consider the time continuous evolution problem (6.19). Given an initial condition u^0 , the solution can be expressed as

$$u^t = \exp\left(-\frac{t}{\eta} A\right) u^0 = e^{-t/\eta} \sum_{k=0}^{+\infty} \left(\frac{t}{\eta}\right)^k \frac{K^k}{k!} u^0 = \sum_{k=0}^{+\infty} a_k(t) u^k, \quad (6.22)$$

with

$$a_k(t) = \frac{1}{k!} e^{-t/\eta} \left(\frac{t}{\eta}\right)^k.$$

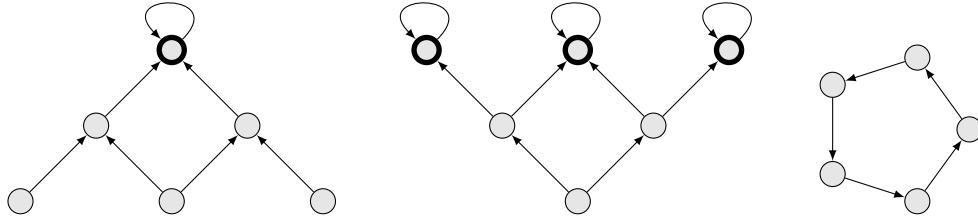


FIGURE 6.3 – Three examples of networks which have not a gradient flow structure : a “dictatorial” hierarchical one, where there is only one influencer (left), a hierarchical “democratic” one, where the vertex in the bottom, thought as the leader, listens (indirectly) to everybody (middle), and a cyclical network of length 5 (right). Influencers are circled.

The exact solution at time t is then an infinite barycentric combination of the discrete iterates (u^k) , weighted by a Poisson distribution of parameter t/η . It makes it possible to distinguish two phases in the evolution : a transitional phase, shorter than $(N-m)\eta$ during which opinions evolves, and a second one, for t larger than $(N-m)\eta$, where most of the mass of the Poisson distribution corresponds to discrete modes u^k equal to the equilibrium state u^{N-m} , where the solution is close to this limit.

Remarque 6.27. Note that the transitional phase can be significantly longer than η , which the characteristic time given by the stability analysis. This is a typical behavior of evolution equations associated to non-diagonalizable operators : flushing out the system of its initial state may take a time much longer than the characteristic damping time.

Remarque 6.28. Hierarchical graph can take two types of extremal, tree-like, forms, depending on the sense of arrows (from the leafs to the root, or the other way around). The first one would be $m = 1$, namely if there exists a unique influencer (see Figure 6.3, left). This corresponds to some kind of dictatorship since, at the end of the process (in no more than $N - 1$ iterations for the discrete version of the model, where N is the number of generations), the opinion of every agent will coincide with the one of the influencer. The other one would be the case when the root of the tree is forwardly connected to all the other vertices (see Figure 6.3, middle). This situation could be thought as a caricatural illustration of democracy, where the leafs are citizens (who are all influencers), and inner nodes correspond to some processes to gather and merge opinions in a hierarchical way. The opinion of the root is then a trade-off between the opinions of citizens.

Cycles purs

La seconde situation correspond au cas de matrices diagonalisables, de valeurs propres non toutes réelles. L’archétype de cette situation correspond aux matrices circulantes (voir figure 6.3, droite, pour le réseau associé), i.e. K est une matrice de permutation avec un cycle unique. On représente l’ensemble des indices $I_N = \{1, \dots, N\}$, de telle sorte que l’orbite de 1 est $1, 2, \dots, N, 1$, etc... La matrice K est circulante : $K_{1,2} = K_{2,3} = \dots = K_{N,1} = 1$, de telle sorte que $K^N = \text{Id}$ (matrice identité). Comme les matrices K^n , pour $n = 0, \dots, N - 1$ sont linéairement indépendantes, le polynôme minimal de K est $X^N - 1$. Ses valeurs propres sont donc $\mu_k = \exp(2ik\pi/N)$, $k = 0, \dots, N - 1$. Noter que cette situation correspond à la saturation de l’inclusion donnée par les cercles de Gershgorin pour une matrice stochastique. les valeurs propres d’une telle matrice sont dans le disque unité de \mathbb{C} , et pour notre matrice circulante elles sont sur le cercle.

Pour tout vecteur initial u^0 , les itérés discrets $u^k = K^k u^0$ sont obtenus par un décalage des valeurs vers la gauche (avec périodicité), i.e. $u_i^{k+1} = u_{i+1}^k$. D’après l’expression (6.22), la solution du problème continu en temps u^t s’exprime comme combinaison convexe des $(u^k)_{k \in \mathbb{N}}$ selon une distribution de Poisson de paramètre t/η :

$$u^t = \exp\left(-\frac{t}{\eta} A\right) u^0 = \sum_{k=0}^{+\infty} \frac{e^{-t/\eta}}{k!} \left(\frac{t}{\eta}\right)^k u^k.$$

L'évolution peut ainsi se voir comme une version diffuse de le phénomène de transport vers les indices décroissants associé à la matrice de shift.

Remarque 6.29. (Additional comments on diagonalizability)

It may sound curious that being diagonalizable or not for a matrix may affect the behavior of the associated evolution model, because diagonalizability in \mathbb{C} is generic for matrices (the set of diagonalizable matrices has full measure). One may be tempted to disregard this pathological character of not being diagonalizable, since it (almost) never happens in practice. Yet, we emphasize here that considering this case does actually make sense. Indeed, for some matrices which are close to the set of non-diagonalizable matrices (and thus far away from the set of *normal* matrices⁹), the behavior is better described by a formula of the type (6.22) rather than by a formula based on eigenvectors. Consider for example the case of a N -linear network detailed previously, with a single influencer, and a series of consecutive influenced agents. The corresponding matrix is “highly” non-diagonalizable, if one may say : 1 is an eigenvalue with multiplicity 1, but 0 has multiplicity $N - 1$, with a one-dimensional eigenspace. Now consider the case when the diagonal elements associated to 1, 2, ..., $N - 1$, are replaced by $K_{11}, \dots, K_{N-1,N-1}$, all positive values of order ε , and pairwise distinct. Since all the eigenvalues are now distinct, the matrix is diagonalizable with eigenvalues $K_{11}, \dots, K_{N-1,N-1}$ and 1, and diagonalizing it makes it possible to express the solution as a linear combination of eigenvectors, with coefficients $e^{-tK_{11}/\eta}, \dots, e^{-t\eta}$. In a modeling or numerical context, this approach must be avoided, because it does not allow a robust description of the behavior of the solution. Indeed, it can be checked that the eigenvectors associated to the K_{ii} 's are almost colinear, so that the change of basis matrix has a very high condition number (its smallest eigenvalue is close to 0), and the eigenvectors depend in a very stiff way upon the coefficients. To sum up, for such matrices which are diagonalizable in theory, but very close to non-diagonalizable matrices, the expected behavior shall be better described by an expression of the type (6.22), rather than on the expression based on eigenvectors. To put it another way, it seems reasonable to consider that diagonalizing the matrix in order to express the exact solution of the evolution problem should be restricted to matrices with a controlled non-normality¹⁰.

6.6 Réseaux charismatiques

Nous nous intéressons ici à des réseaux qui encodent des interactions d'une nature symétrique. Plus précisément, les réseaux que nous considérons ici sont basés sur l'existence d'un paramètre afférent à chaque individus, un poids que nous appellerons *charisme* dans ce contexte, qui conditionne l'influence qu'il exerce sur les autres. Comme nous allons le voir, cette hypothèse rapprochera les réseaux qui la vérifie des réseaux résistifs ou, dans un contexte stochastique, des chaînes de Markov *réversibles*.

Definition 6.30. (Réseaux charismatiques)

On dit que le réseau (VE, K) est *charismatique* s'il existe un champ $m = (m_x) \in]0, +\infty[^V$ tel que, pour tous $x, y \in V$,

$$m_x K_{xy} = m_y K_{yx}. \quad (6.23)$$

Remarque 6.31. On vérifie immédiatement que les réseaux charismatiques sont un cas particulier de flots de gradient. En effet, si l'on introduit la matrice $C = (C_{xy})$, avec $C_{xy} = m_x K_{xy}$, la matrice C est symétrique, et l'on a, avec $M = \text{diag}(m_x)$,

$$C = MK \implies A = I - K = I - M^{-1}C = M^{-1}(M - C) = M^{-1}H,$$

où M est s.d.p. et $H = M - C$ est symétrique, on est donc bien dans le cas d'un flot de gradient pour la métrique induite par M , dans le cas d'une matrice M diagonale. On a de plus

$$m_x K_{xy} = C_{xy} \implies \sum_{y \sim x} m_x K_{xy} = m_x = \sum_{y \sim x} C_{xy}.$$

9. A matrix which commutes with its adjoint or, equivalently, a matrix which can be diagonalized in an orthogonal basis.

10. P. Henrici, Bounds for iterates, inverses, spectral variation and fields of values of non-normal matrices, *Numerische Mathematik* 4, pp. 24–40 (1962).

<https://eudml.org/doc/131513>

Remarque 6.32. Si nous considérons K comme la matrice de transition d'une chaîne de Markov sur l'ensemble V , où K_{xy} est la probabilité de transition de x à y , alors la définition 6.30 correspond à celle d'une chaîne *reversible*, et $m = (m_x)$ joue le rôle d'une mesure invariante.

Remarque 6.33. Si le réseau est charismatique, $K_{xy} > 0$ si et seulement si $K_{yx} > 0$. En conséquence $x \rightarrow y$ si et seulement si $y \rightarrow x$. On considérera néanmoins le graphe sous jacent comme orienté, étant entendu que $(x, y) \in E \iff (y, x) \in E$, mais l'on ne fait pas l'identification entre les deux arêtes. Par ailleurs, tout influenceur $x \in \Gamma$ est isolé, et ne joue donc aucun rôle dans la dynamique d'opinion. Si l'on se restreint à des réseaux connexes, il ne peut donc pas y avoir d'influenceurs.

Remarque 6.34. L'identité (6.23) peut s'écrire

$$K_{xy} = \frac{m_y}{m_x} K_{yx}.$$

L'influence que y exerce sur x dépend donc du rapport des charismes de x et de y , et de l'influence que x exerce sur y . Plus le charisme de y est grand comparé à celui de x , plus y influence x comparé à l'influence de x sur y . C'est ce qui justifie l'appellation *charisma* : plus le charisme est grand, plus l'influence exercée sur les autres est grande.

Proposition 6.35. Soit (V, E, K) un réseau charismatique. Si le réseau est connexe, alors le charisme est défini de façon unique à constante multiplicative près. En particulier il admet un unique charisme $m = (m_x)$ qui est une loi de probabilité sur V , i.e. tel que

$$\sum_V m_x = 1.$$

Démonstration. Notons en premier lieu que, d'après la remarque 6.33, la connexité entraîne la forte connexité. Considérons m et m' deux champs de charisme sur V . Soit $x \in V$ arbitraire. On pose $\lambda = m'_x/m_x > 0$. Pour tout $y \sim x$, on a

$$m'_y = m'_x \frac{K_{xy}}{K_{yx}} = m'_x \frac{m_y}{m_x} = \lambda m_y.$$

Cette relation de proportionnalité se propage de proche en proche, donc en tous les sommets par connexité du graphe, on a donc $m' = \lambda m$. Il existe donc en particulier un unique champ de charisme de masse totale unitaire. \square

Remarque 6.36. A cardinal de V fixé, on peut établir une relation de bijection entre les réseaux charismatiques et l'ensemble des matrices symétriques à coefficient positifs ou nul, à constante positive multiplicative près. En effet, si K et m satisfont les relations (6.23) alors la matrice C définie par $C_{xy} = m_x K_{xy}$, est symétrique. Si l'on note M la matrice $\text{diag}(m_x)$, on peut écrire $K = M^{-1}C$. Réciproquement, si C est une matrice symétrique dont les éléments sont positifs, et que l'on souhaite lui associer une matrice $K = M^{-1}C$ encodant les influences d'un réseau charismatique, le seul choix possible est, étant donnée la contrainte de normalisation des lignes de K ,

$$m_x = \sum_y C_{xy}.$$

Si l'on prend pour M la matrice diagonale de coefficients $(m_x)_{x \in V}$, alors $K = M^{-1}C$ correspond à un réseau charismatique.

Les réseaux charismatiques présentent une propriété de conservation particulière. Nous avons noté (voir remarque 6.4) que l'opinion totale n'est en général pas conservée. Dans le cas des réseaux charismatiques, une certaine quantité est pourtant conservée, il s'agit d'une certaine moyenne de l'opinion, plus précisément de l'espérance de l'opinion relativement à la mesure (m_x) .

Proposition 6.37. (Propriété de conservation)

Soit (V, E, K, m) un réseau charismatique, et (u^k) la suite des opinions associées au modèle (6.2). L'opinion moyenne relativement à la mesure m , définie par

$$\bar{u}^k = \sum_{x \in V} m_x u_x^k,$$

se conserve au cours des itérations.

Démonstration. On a

$$\begin{aligned}\bar{u}^{k+1} &= \sum_{x \in V} m_x u_x^{k+1} = \sum_{x \in V} \sum_{y \sim x} m_x K_{xy} u_y^k = \sum_{x \in V} \sum_{y \sim x} m_y K_{yx} u_y^k \\ &= \sum_{y \in V} m_y u_y^k \sum_{x \sim y} K_{yx} = \sum_{y \in V} m_y u_y^k = \bar{u}^k,\end{aligned}\tag{6.24}$$

qui établit la propriété de conservation annoncée. \square

Cette propriété permet de caractériser les limites possibles du problème d'évolution.

Proposition 6.38. Soit (V, E, K, m) un réseau charismatique. On suppose que la suite des itérés du modèle discret converge vers un consensus associé à la valeur u^∞ . Alors cette valeur u^∞ correspond à la moyenne des opinions initiales relativement au charisme m normalisé à 1 selon la proposition 6.35 :

$$u^\infty = \sum_{x \in V} m_x u_x^0.$$

Démonstration. Si toutes les opinions u_x^k convergent vers u^∞ , on a, d'après la proposition 6.37,

$$\sum_{x \in V} m_x u_x^0 = \sum_{x \in V} m_x u_x^k \rightarrow \sum_{x \in V} m_x u^\infty = u^\infty.$$

qui montre que l'opinion limite commune est bien la combinaison barycentrique des opinions initiales, pondérée par les charismes des agents. \square

Point de vue variationnel, flot de gradient & réseaux résistifs

Une autre particularité du cadre charismatique est que le problème présente une structure variationnelle. Considérons un réseau charismatique (V, E, K) , de charisme normalisé m . Comme décrit dans la section 6.5 (voir aussi la remarque 6.31), nous sommes dans la situation où A s'écrit $M^{-1}B$, avec $M = \text{diag}(m_x)$, et $B = M - C$. Le problème d'évolution continu en temps est donc, d'après la proposition 6.24, un flot de gradient pour la fonctionnelle

$$\Psi(u) = \frac{1}{2} \langle (M - C)u \mid u \rangle = \frac{1}{2} \sum_x u_x \sum_{y \sim x} C_{xy} (u_x - u_y) = \frac{1}{2} \sum_{e \in E} C_{xy} (u_x - u_y)^2,\tag{6.25}$$

pour le produit scalaire défini par

$$\langle u \mid v \rangle_M = \sum_{x \in V} m_x u_x v_x.$$

En conséquence, l'énergie Ψ décroît au cours du temps, et le modèle exprime un principe d'évolution selon la ligne de plus grande pente vis-à-vis de Ψ , pour la métrique définie par M .

Cette énergie permet de faire un lien avec les réseaux résistifs. On peut penser u_x comme un potentiel en x , la quantité $C_{xy} = m_x K_{xy}$ (qui est symétrique en x, y) jouant le rôle d'une *conductance* (inverse d'une résistance) de l'arête (symétrique selon ce point de vue) joignant x et y . La quantité $\Phi(u)$ correspond dans cette analogie électrique à (la moitié de) l'énergie dissipée au sein du réseau aux conductances $m_x K_{xy}$ et aux potentiels u_x . L'évolution tend donc à minimiser cette énergie dissipée, que l'on peut voir comme une estimation de l'écart à l'équilibre en termes d'opinion. Dans cette optique, les paramètres C_{xy} peuvent être interprétés comme des *coefficients de friction*, et les u_x comme des vitesses¹¹.

11. Comme deux objets en contact, allant à des vitesses différentes, sont soumis à une force d'interaction de type friction proportionnelle à leur vitesse relative, dissipant ainsi une énergie proportionnelle au carré de cette vitesse relative.

En poursuivant cette analogie avec les systèmes mécaniques, concevoir l'opinion d'un sommet x comme une quantité scalaire de type vitesse, la quantité obtenue par multiplication par la la "masse" m_x , donne une quantité de mouvement-opinion. On a bien un principe de Newton pour ce système mécanique : d'après la proposition 6.37, la quantité de mouvement-opinion globale pour ce système libre (non forcée de l'extérieur) se conserve. Le carré de la norme naturellement associée au modèle correspond à une énergie cinétique. On prendra garde en revanche que l'énergie globale Φ dont dérive l'équation d'évolution , quadratique en les vitesses, n'a rien d'une énergie cinétique, il s'agit plutôt comme indiqué ci-dessus d'une somme de termes de nature frictionnelle, qui quantifieraient des puissances dissipées au niveau de chaque arête (relation entre deux individus), d'autant plus que les opinions divergent. Cette interprétation est étayée par un pseudo-bilan énergétique que l'on peut obtenir à partir de l'équation de conservation de la quantité de mouvement, on effectue le produit scalaire de

$$M \frac{du}{dt} = -\nabla \Psi(u)$$

avec la "vitesse" u , pour obtenir

$$\frac{d}{dt} \frac{1}{2} \langle Mu | u \rangle = -\langle \nabla \Psi(u) | u \rangle = -\sum_{e \in E} C_{xy} |u_x - u_y|^2,$$

qui peut se lire : la dérivée en temps de l'énergie cinétique est égale à la puissance dissipée par friction entre opinions différentes. S'il s'agissait d'un système mécanique standard, cette énergie serait dissipée sous forme de chaleur au sein du système ou vers le monde extérieur (l'ajout d'un modèle thermique permettrait de préciser le devenir de cette énergie thermique).

On peut aussi interpréter l'équation d'évolution d'un point de vue thermique :

$$m_x \frac{du_x}{dt} = -\frac{1}{\eta} \sum_{y \leftarrow x} C_{xy} (u_x - u_y)$$

6.7 Niveau de certitude

On se propose ici d'intégrer au modèle l'assurance qu'une personne a sur sa propre opinion. Dans cet esprit, l'opinion sera représentée par une loi de probabilité sur l'espace des états. On considérera toujours le réseau d'influence modélisé par la collection de coefficients (K_{xy}) . En toute généralité, si l'on considère que les agents ont une opinion sur une certaine variable u dans un espace mesuré Ω , l'opinion d'un individu x sera décrite par une densité de probabilité ρ_x sur Ω : pour toute partie $A \subset \Omega$ mesurable, $\int_A \rho_x(u) du$ est la probabilité accordée par x que u soit dans A . Si ρ_x est une masse ponctuelle en $\bar{u} \in \Omega$, cela signifie que x est pleinement convaincu que la valeur de u est exactement \bar{u} . Nous nous limiterons dans ce qui suit au cas d'une opinion à valeur réelle, représentée par une densité de probabilité *gaussienne*.

Écrire un modèle, dans l'esprit des sections précédentes, qui intègrerait ce nouvel ingrédient, passe par la définition d'une règle d'interpolation entre les densités de probabilités, qui permettrait d'étendre le modèle (6.2). Dans la cadre de densités gaussiennes réelles, caractérisées par leur couple espérance-écart type (u, σ) , on se convaincra qu'une interpolation linéaire

$$((1 - \theta)u_0 + \theta u_1, (1 - \sigma)u_0 + \theta \sigma_1)$$

n'est pas pertinente. Cela signifierait que, discutant avec deux personnes d'opinions 0 et 1, respectivement, très convaincues ($\sigma \ll 1$), on en ressort avec une forte conviction que la valeur est 1/2.

On se propose de construire ici une métrique sur les densités de probabilité (restreintes ici aux densités gaussiennes), qui conduisent à une interpolation plus adaptée à la réalité modélisée.

L'idée, que nous présentons ici de façon générale, est la suivante : on considère un densité $\rho = \rho(u)$ sur Ω (qui représente l'ensembles des opinions possibles), et une variation $\delta\rho$, de telle sorte que $\rho + \delta\rho$

est encore une densité de probabilité. L'entropie relative de $\rho + \delta\rho$ relativement à ρ (aussi appelée divergence Kullback-Leibler) s'écrit

$$\text{KL}(\rho + \delta\rho | \rho) = \int_{\Omega} (\rho + \delta\rho) \log \left(\frac{\rho + \delta\rho}{\rho} \right) = \int_{\Omega} \frac{\rho + \delta\rho}{\rho} \log \left(\frac{\rho + \delta\rho}{\rho} \right) \rho. \quad (6.26)$$

Remarque 6.39. (Relative entropy and measure of opinion discrepancy)

Measuring the difference between two densities in terms of relative entropy makes some sense in the context of opinions. In particular, if an agent has an opinion described by ρ , with ρ almost vanishing in some zone ω of Ω , it means that this agent considers as very unlikely that the value of u could belong to ω . Any opinion which gives a significant probability that u may belong to ω can then be considered as very remote from ρ , which is exactly expressed by (6.26). More precisely, if $\delta\rho$ has a positive value on ω , even a small value, KL divergence will converge to $+\infty$ when ρ converges to 0 on ω . Let us push this interpretation a little further : the non-symmetry of the KL divergence makes some sense in the context of opinion : if an agent gives to ω some (small) positive probability, the opinion which rules out $u \in \omega$ is not so far away from the initial one, which is again reflected by (6.26). In other words, the way to measure discrepancy provided by the relative entropy is consistent with the asymmetry which is natural in the world of opinions encoded by probabilities : $\epsilon > 0$ is far away from 0, whereas 0 is actually not so far away from ϵ .

L'entropie relative de $\rho + \delta\rho$ relativement à ρ peut être exprimée comme suit (en utilisant le fait que $\delta\rho$ est d'intégrale nulle),

$$\begin{aligned} \text{KL}(\rho + \delta\rho | \rho) &= \int_{\Omega} (\rho + \delta\rho) \log \left(\frac{\rho + \delta\rho}{\rho} \right) \\ &= \int_{\Omega} (\rho + \delta\rho) \left(\frac{\delta\rho}{\rho} - \frac{(\delta\rho)^2}{2\rho^2} + o((\delta\rho)^2) \right) \\ &\sim \frac{1}{2} \int_{\Omega} \frac{(\delta\rho)^2}{\rho} = \frac{1}{2} \int_{\Omega} (\delta \log \rho)^2 \rho. \end{aligned}$$

On considère maintenant une famille de densités de probabilités (ρ_{θ}) paramétrées par une collection de paramètres $\theta = (\theta_1, \dots, \theta_p)$, dans un espace de paramètres Θ . On a

$$\delta \log \rho_{\theta} = \sum_i \frac{\partial \log \rho_{\theta}}{\partial \theta_i} \delta \theta_i,$$

On obtient ainsi

$$\text{KL}(\rho_{\theta+\delta\theta} | \rho_{\theta}) = \langle I\delta\theta, \delta\theta \rangle + o(\delta\theta^2),$$

où I , appelée matrice d'information de Fisher, est définie par

$$I = (I_{ij})_{1 \leq i, j \leq p} = \left(\frac{1}{2} \int_{\Omega} \left(\frac{\partial \log \rho_{\theta}}{\partial \theta_i} \right) \left(\frac{\partial \log \rho_{\theta}}{\partial \theta_j} \right) \rho_{\theta} \right)_{1 \leq i, j \leq p}.$$

Cette expression définit une métrique *riemannienne* sur l'espace des paramètres, ce que l'on note en général

$$ds^2 = \langle Id\theta | d\theta \rangle.$$

On considère maintenant une courbe lisse dans l'espace des densités paramétrées, entre ρ_{θ_0} et ρ_{θ_1} . On l'identifie avec la courbe correspondante dans l'espace des paramètres Θ , que l'on écrit $t \in [0, 1] \mapsto \theta(t)$. La *longueur* de cette courbe selon la métrique considérée est définie par

$$\ell = \int_0^1 \langle I\dot{\theta}, \dot{\theta} \rangle^{1/2} dt.$$

Dans le cas où pour toutes densités ρ_{θ_0} et ρ_{θ_1} il existe au moins une courbe qui les relient, on peut définir une distance *géodésique* comme

$$d_F(\rho_{\theta_0}, \rho_{\theta_1}) = \inf_{\theta \in \Xi} \int_0^1 \langle I\dot{\theta}, \dot{\theta} \rangle^{1/2} dt, \quad (6.27)$$

où Ξ est l'ensemble de toutes les courbes lisses $t \in [0, 1] \mapsto \theta(t) \in \Theta$ qui relient θ_0 à θ_1 . On peut se demander maintenant s'il est possible de définir la notion de barycentre de plusieurs densités. Plus précisément considérons des paramètres $\theta_1, \dots, \theta_p$ et des poids positifs K_1, \dots, K_p , avec $\sum K_i = 1$, le problème

$$\min_{\theta \in \Theta} \sum_{i=1}^p K_i d_F(\rho_\theta, \rho_{\theta_i})^2$$

admet-il une unique solution $\bar{\theta}$? Si c'est bien le cas, on écrira

$$\bar{\theta} = \text{Bar}(\theta_i, K_i)_{1 \leq i \leq p} \quad \text{ou } \bar{\rho} = \rho_{\bar{\theta}} = \text{Bar}(\rho_{\theta_i}, K_i)_{1 \leq i \leq p}.$$

Dans le cas où les barycentres sont définis sans ambiguïté, on pourra définir un problème d'évolution discret en temps, formellement semblable à (6.2), comme

$$\theta_x^{k+1} = \text{Bar} < (\theta_y^k, K_{xy})_{x \rightarrow y} >. \quad (6.28)$$

Considérons maintenant le cas de quantités gaussiennes sur \mathbb{R} :

$$\rho_\theta(u) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(u-m)^2}{2\sigma^2}},$$

paramétrées par $\theta = (m, \sigma) \in \Theta = \mathbb{R} \times (0, +\infty)$. Un calcul direct (impliquant les moments de la gaussienne) conduit à une expression explicite de la métrique :

$$\langle Id\theta, d\theta \rangle = \frac{1}{\sigma^2} (dm^2 + 2d\sigma^2). \quad (6.29)$$

Il s'agit (au facteur 2 près pour le second terme), de la métrique du *demi plan de Poincaré* \mathbb{H}^2 .

6.8 Clivage des opinions au sein d'une population

On s'intéresse ici au phénomène de l'apparition spontanée au sein d'un réseau social de sous-communautés de personnes qui pensent essentiellement la même chose, sans interaction entre les différentes sous communautés. Le point de départ est le modèle linéaire (6.2).

$$u_x^{n+1} = \sum_{x \rightarrow y} K_{xy} u_y^n \quad \forall x \in V,$$

où K_{xy} quantifie l'influence que y exerce sur x . On considérera le modèle discret voisin

$$u_x^{n+1} = u_x^n - \frac{h}{\eta} \sum_{y \leftarrow x} K_{xy} (u_x^n - u_y^n) \quad \forall x \in V$$

où $\eta > 0$ est un temps caractéristique qui conditionne la vitesse d'évolution de l'opinion.

On se propose d'intégrer le fait que les coefficient K_{xy} sont susceptibles d'évoluer au cours du temps, selon le principe qu'on a tendance à moins écouter une personne qui professe des opinions très éloignées de la notre.

On peut modéliser cette tendance de la façon suivante :

$$\begin{cases} u_x^{n+1} = u_x^{n+1} - \frac{h}{\eta} \sum_{y \leftarrow x} K_{xy} (u_x^n - u_y^n), & \forall x \in V \\ \tilde{K}_{xy}^{n+1} = \frac{K_{xy}^n \exp(-|u_x^n - u_y^n|/\bar{u})}{\sum_{y' \leftarrow x} K_{xy'}^n \exp(-|u_x^n - u_{y'}^n|/\bar{u})} & \forall (x, y) \in E, \\ K_{xy}^{n+1} = (1 - h/\eta_K) K_{xy}^n + h/\eta_K \tilde{K}_{xy}^{n+1} \end{cases}$$

où $\eta_K > 0$ est un paramètre qui conditionne la rapidité d'évolution des facteurs d'influence, et \bar{u} un paramètre homogène à une opinion qui conditionne l'écart d'opinion au delà duquel on a tendance à moins écouter son interlocuteur.

Ce modèle est la version discrète du modèle continu en temps :

$$\begin{cases} \frac{du_x}{dt} = -\frac{1}{\eta} \sum_{y \leftarrow x} K_{xy}(u_x - u_y), & \forall x \in V \\ \tilde{K}_{xy} = \frac{K_{xy} \exp(-|u_x - u_y|/\bar{u})}{\sum_{y' \leftarrow x} K_{xy'} \exp(-|u_x - u_{y'}|/\bar{u})} & \forall (x, y) \in E, \\ \frac{dK_{xy}}{dt} = \frac{1}{\eta_K} (\tilde{K}_{xy} - K_{xy}) \end{cases}$$

6.9 Propagation sur des réseaux du type “application”

On considère ici la situation où (V, E, K) s'identifie à une application $\Phi : V \rightarrow V$, i.e. pour tout x , on a $K_{xy} = 1$ pour un unique $y = \Phi(x)$. En d'autres termes

$$K_{xy} = \delta_{\Phi(x), y},$$

où δ est le symbole de Kronecker. Dans cette situation, le problème d'évolution discret s'écrit

$$u_x^{k+1} = u_{\Phi(x)}^k \quad \forall x \in V,$$

de telle sorte que

$$u^k = u^{k-1} \circ \Phi = u^{k-2} \circ \Phi \circ \Phi = \dots = u^0 \circ \Phi^k,$$

qui est le tiré en arrière (*pullback* en anglais) de u^0 par $\Phi^k = \Phi \circ \dots \circ \Phi$.

Graphes bijectifs

On considère ici le cas bijectif : pour tout y , $K_{xy} = 1$ pour un unique x , c'est à dire que Φ est une bijection.

Proposition 6.40. Dans le cas où (K_{xy}) est une matrice de permutation, la suite des itérés (u^k) résultant de (6.2) est périodique pour toute condition initiale u^0 .

Démonstration. C'est une conséquence directe du fait que toute permutation peut être décomposée en le produit de cycles de supports disjoints. En effet, considérons, pour $x \in V$, la suite $(\Phi^k(x))$, avec

$$\Phi^k(x) = \underbrace{\Phi \circ \dots \circ \Phi}_{k \text{ fois}}(x).$$

Comme V est fini, il existe $L \geq 1$ tel que $\Phi^L(x)$ est égal à $(\Phi^j(x))$, pour un $j < L$. considérons le plus petit de ces indices L . Si le j associé n'est pas 0, alors $\Phi(\Phi^{L-1}(x)) = \Phi(\Phi^{j-1}(x))$, ce qui implique (du fait que Φ est une bijection) que $\Phi^{L-1}(x) = \Phi^{j-1}(x)$, ce qui contredit le caractère minimal de L . On a donc $j = 0$, et la suite $(\Phi^k(x))$ est L -périodique. Pour tout x' dans l'orbite de x , i.e. égal à l'un des $(\Phi^k(x))$, la suite est L -périodique, et la restriction de Φ à l'orbite de x est un cycle. Cette orbite et son complémentaire sont stables, on peut donc appliquer la même démarche au complémentaire, jusqu'à épuisement.

La solution de (6.2) est donc périodique, avec une période au plus égale au plus petit commun multiple des ordres des cycles. \square

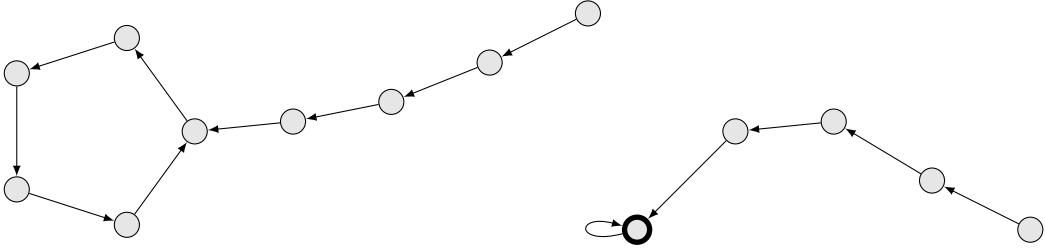


FIGURE 6.4 – Lollypop graph, only one influencer (circled node).

Remarque 6.41. (Fonction de Landau)

Une question naturelle se pose : pour un cardinal N de V donné, quelle est la période maximale d'une solution de (6.2) ? On considère que les opinions initiales sont distinctes deux à deux. Cette question est liée à ce qu'on appelle la *fonction de Landau* $N \mapsto g(N)$, qui représente l'ordre maximal d'un élément dans le groupe symétrique S_N . Cette fonction $g(\cdot)$ est croissante, de façon évidente, avec $g(1) = 1$, $g(2) = 2$, $g(3) = 3$, $g(4) = 4$, $g(5) = 6$, Plus généralement, pour $M \in \mathbb{N}$, avec une décomposition en facteurs premiers

$$M = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_k^{\alpha_k},$$

on introduit $\ell(M) = p_1^{\alpha_1} + p_2^{\alpha_2} + \cdots + p_k^{\alpha_k}$. On a que

$$g(N) = \max_{\ell(M) \leq N} M,$$

qui implique en particulier $\ell(g(N)) \leq N$. On se reportera à [Deleglise¹²] pour un historique des avancées réalisées sur cette fonction de Landau Function, ainsi que pour la description d'un algorithme de calcul effectif de sa valeur pour N grand.

Graph of the mapping type

Let us now assume that the index of each vertex is equal to 1, i.e. for any x , $K_{xy} = 1$ for some $y \in V$. Like in the bijective situation, it is fruitful to consider, for any $x \in V$, the sequence $(\Phi^k(x))$. As previously, since V is finite, there is a $L \geq 1$ such that $(\Phi^L(x))$ is equal to $(\Phi^j(x))$, for some $j \leq L$. If $j = 0$ like in the bijective case, one recovers a j -cycle. If $j = L$, then necessarily the corresponding $y = \Phi^L(x) \in \Gamma$, and the orbit is a straight path to Γ . If $0 < j < L$, then the path connects x to a cycle of length $L - j + 1$. Any solution to (6.2) is then periodic after some time, with a period at most equal to the l.c.m. of the periods $L - j + 1$ exhibited previously. Figure 6.4 presents a graph with two connected components, with two different types of attractors : a cycle (left) and a single vertex (right). In this situation, the period is 6, and number of step after which it is periodic is 4. Note that all the initial opinions lying on the linear parts of the graph have been washed out in the evolution process.

12. M. Deléglise, J.-L. Nicolas, P. Zimmermann, Landau's function for one million billions, *Journal de théorie des nombres de Bordeaux*, Tome 20 (2008) no. 3, pp. 625-671

6.10 Exercices

Exercice 6.1. On considère un réseau (V, E, K) fortement connexe, et l'on suppose que $K_{xx} > 0$ pour au moins un $x \in V$. Montrer que toute solution du modèle linéaire converge vers un consensus.

Exercice 6.2. On considère un réseau (graphe non orienté pondéré) (V, E, K) avec $\Gamma \neq \emptyset$, et $V - \cdot \rightarrow \Gamma$. On note u la solution du problème de Dirichlet associé à $U \in \mathbb{R}^\Gamma$, et $(\mu_{x,y})$ les coefficients barycentriques tels que

$$u_x = \sum_{y \in \Gamma} \mu_{xy} U_y, \quad \sum_{y \in \Gamma} \mu_{xy} = 1.$$

- a) Montrer que si $x - \cdot \rightarrow y \in \Gamma$, alors $\mu_{xy} > 0$.
- b) On considère $(\mu_{x,y})$ comme la matrice d'une application de \mathbb{R}^Γ dans \mathbb{R}^V . Montrer que cette matrice est pleine si et seulement si tout x de V est connecté à tout y de Γ

Exercice 6.3. On rappelle que la solution du problème d'évolution continu (6.19) s'écrit

$$u^t = \sum_{k=0}^{+\infty} a_k(t) u^k, \quad a_k(t) = \frac{1}{k!} e^{-t/\eta} \left(\frac{t}{\eta} \right)^k,$$

où (u^k) est la suite des itérés du modèle discret (6.2).

Établir un lien entre l'itéré k -ième u^k et la solution du problème continu au temps $t = k\eta$.

Exercice 6.4. On considère le problème d'évolution continue associé à un réseau charismatique

$$\frac{du}{dt} = -\frac{1}{\eta} Au,$$

avec $A = I - K = M^{-1}(M - C)$, où C est une matrice symétrique dont les éléments sont positifs, et $M = \text{diag}(m_x)$ la matrice des charismes. On définit l'opinion moyenne (relativement aux charismes) par

$$\bar{u} = \sum m_x u_x = Mu \cdot e,$$

où e est le vecteur qui ne contient que des 1. Montrer que cette opinion moyenne se conserve au cours du temps.

Exercice 6.5. En quel sens peut on dire que, dans un réseau charismatique, l'opinion initiale d'un individu donné ne peut peser au mieux qu'à 50 % dans la valeur de l'éventuel consensus final ? Énoncer précisément et démontrer une propriété correspondant à cette assertion.

Chapitre 7

Modèles de propagation d'épidémies

Sommaire

7.1	Modèle SIR	149
7.2	Modèle stochastique orienté agent	154
7.3	Modèle déterministe portant sur des probabilités d'infection	155
7.4	Développements, extensions	162
7.5	Prise en compte de la perte d'immunité	163
7.6	Exercices	169

7.1 Modèle SIR

On s'intéresse à la propagation d'une épidémie au sein d'une population de N individus. Chaque individu appartient à un instant donné à l'une des trois catégories : Susceptible (d'attraper la maladie), Infected, et Recovered (remis). On se place dans l'hypothèse d'une immunité totale : les personnes qui ont eu la maladie sont immunisées. On note K le nombre de contacts qu'un individu a pendant une journée, p la probabilité qu'une personne infectée contamine une personne susceptible lors d'un contact (voir remarque 7.9). On suppose que le nombre de contacts est le même pour tout le monde (hypothèse très forte et peu réaliste d'homogénéité). Le taux de guérison journalier (inverse de la durée de la maladie) est noté γ . On note S^k la proportion de susceptibles dans la population au jour k , et on définit de même I^k et R^k . Au jour k , chaque S a un nombre de contacts avec un infecté égal à KI^k , sa probabilité d'être infecté est donc de pKI^k . La variation du nombre de susceptible d'un jour à l'autre est donc de

$$NS^{k+1} - NS^k = -NS^k pK I^k.$$

Si l'on note $\beta = pK$, on obtient donc le système suivant (en divisant par N)

$$\begin{cases} S^{k+1} - S^k &= -\beta S^k I^k \\ I^{k+1} - I^k &= +\beta S^k I^k - \gamma I^k \\ R^{k+1} - R^k &= +\gamma I^k \end{cases} \quad (7.1)$$

qui est la version discrète du système d'équations différentielles

$$\begin{cases} \frac{dS}{dt} &= -\beta SI \\ \frac{dI}{dt} &= +\beta SI - \gamma I \\ \frac{dR}{dt} &= +\gamma I \end{cases} \quad (7.2)$$

Le paramètre $\beta = pK$, homogène à l'inverse d'un temps, est appelé taux d'incidence. Il quantifie à la fois le risque de transmission de la maladie lors d'une rencontre entre un sain et un infecté, et la probabilité qu'une telle rencontre puisse se produire,

γ est aussi homogène à l'inverse d'un temps, qui correspond au temps moyen pendant lequel l'individu infecté est contagieux,

$R_0 = \beta/\gamma$ est appelé taux de reproduction de base, ou taux de reproduction initial, il quantifie le nombre d'individus qu'un individu infecté est susceptible de contaminer pendant la durée de sa maladie, s'il se trouve au sein d'une population de susceptibles.

Remarque 7.1. Précisons l'interprétation de R_0 donnée ci-dessus. En premier lieu $\beta S I N$ est le nombre d'infectés par jour, donc βS est le nombre de personnes qu'un infecté contamine par jour. Le terme $-\gamma I$ dans la deuxième équation représente par ailleurs une décroissance exponentielle. Si l'on écarte le terme source, l'équation sur I s'écrit $\dot{I} = -\gamma I$, dont la solution s'écrit $e^{-\gamma t}$ (pour une population initiale unitaire). La durée d'infection pour les individus infectés initialement suit donc une distribution $\gamma e^{-\gamma t}$, la moyenne s'écrit donc

$$D = \int_0^{+\infty} t \gamma e^{-\gamma t} dt = \frac{1}{\gamma}.$$

Le nombre d'infectés moyen par malade est donc $S\beta/\gamma = SR_0$, égal à R_0 si S est égal à 1. On se reportera à la section 7.5 pour une discussion sur le sens de R_0 et sur la manière de l'estimer en pratique.

On notera au passage le caractère peu réaliste de la distribution exponentielle des durées de maladie, qui est faite implicitement lorsque l'on encode la guérison au bout d'un temps $1/\gamma$ par le terme $-\gamma I$. Cela signifie en particulier que pour une part significative des individus, la maladie dure très peu de temps. La prise en compte d'une durée de maladie plus réaliste ne peut se faire qu'en sortant du cadre du modèle SIR classique (voir section 7.4)

Remarque 7.2. Pour aller un peu plus sur les questions d'adimensionnement, notons que le changement de variable en temps

$$\tilde{t} = \gamma t,$$

qui revient à prendre pour unité de temps la durée moyenne d'infection, conduit au système (sous forme réduite à deux équations)

$$\begin{cases} \frac{d\tilde{S}}{d\tilde{t}} = -\frac{\beta}{\gamma} \tilde{S} \tilde{I} \\ \frac{d\tilde{I}}{d\tilde{t}} = +\frac{\beta}{\gamma} \tilde{S} \tilde{I} - \tilde{I}. \end{cases} \quad (7.3)$$

On en déduit que le comportement du système, en particulier la forme des courbes (monotonie, convexité / concavité, ...), l'ensemble des valeurs prises par les différentes variables (en particulier le max et le min de chacune des variables), et donc en particulier la fraction totale de personnes infectées *in fine* (égale à 1 moins le nombre de Susceptibles résiduels), ne dépendent que du rapport β/γ appelé taux de reproduction, et noté R_0 . À R_0 fixé le paramètre γ , que nous avons éliminé par adimensionnement (comme nous aurions pu le faire pour β), conditionne lui la *cinétique* d'évolution.

Question 7.1. Proposer une manière d'intégrer des campagnes de vaccination à ce modèle.

Points d'équilibre, stabilité

La troisième équation n'influençant pas les deux autres, on considère le système constitué par les deux premières équations de (7.2). Ce système en (S, I) admet une famille de points d'équilibre $(S_0, 0)$, avec $S_0 \in [0, 1]$.

On peut estimer la stabilité informelle du système en considérant la seconde équation de (7.2), qui s'écrit

$$\frac{dI}{dt} = \gamma I \left(\frac{\beta}{\gamma} S - 1 \right) = \gamma I (R_0 S - 1).$$

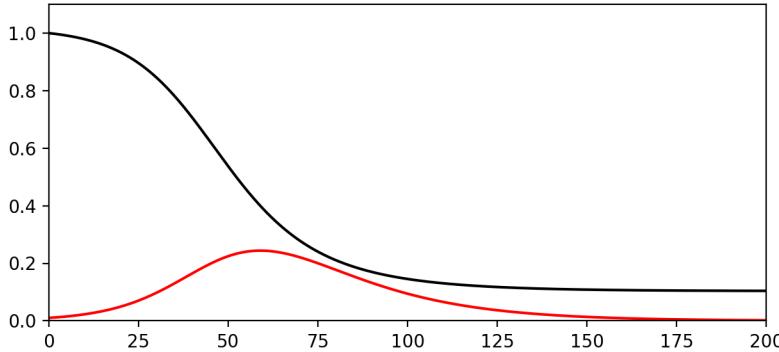


FIGURE 7.1 – Courbes $t \mapsto S(t)$ (noir) et $t \mapsto I(t)$ (rouge) pour $R_0 = 2.5$, $\beta = 1/18$.

Au voisinage du point fixe, cette équation représente une évolution exponentielle de I au taux $R_0 S_0 - 1$. Si l'on part d'une condition initiale sans immunité ($S_0 = 1$), on a donc instabilité, c'est à dire croissance de l'épidémie à partir d'un petit nombre de cas, si le taux de reproduction de base est R_0 est plus grand que 1. De façon plus générale, le nombre d'infectés commencera à décroître quand $R_0 S$ est inférieur à 1. On appelle parfois cette quantité le *taux de reproduction effectif*, qui prend en compte la fraction des personnes immunisées (voir section 7.5).

Remarque 7.3. On retrouve bien sûr le même résultat de stabilité en considérant, de façon plus rigoureuse, le système linéarisé au point $(S_0, 0)$, qui s'écrit

$$\begin{pmatrix} 0 & -\beta S_0 \\ 0 & \beta S_0 - \gamma \end{pmatrix}$$

dont le polynôme caractéristique est $-\lambda(\beta S_0 - \gamma - \lambda)$. Le point d'équilibre est instable (i.e. la maladie peut se développer à partir d'un petit nombre de gens infectés) dès que la valeur propre non nulle est strictement positive, i.e. si

$$\beta S_0 - \gamma = \gamma(R_0 S_0 - 1) > 0.$$

Pour résumer, le système admet un continuum de points d'équilibres $(S_0, 0)$, qui sont instables pour $S_0 > 1/R_0$, et stables pour $S_0 < 1/R_0$. Comme on le verra plus loin, une épidémie “naturelle” (i.e. sans mesures sanitaires particulières) peut être vue comme une trajectoire d'un point fixe instable vers un point fixe stable.

Temps caractéristique

Dans la situation instable, le temps caractéristique correspond au temps qu'il faut pour multiplier la population des I par $e \approx 2.7$ (au début de l'épidémie, i.e. pour I petit), il s'exprime comme l'inverse de la valeur propre non triviale :

$$\tau = \frac{1}{\gamma} \frac{1}{R_0 S_0 - 1}, \tag{7.4}$$

il correspond donc à la durée moyenne de l'infection d'un individu corrigée par un facteur qui dépend de R_0 et de la population initialement susceptible. Si l'on sait que la population est entièrement constituée de Susceptibles (i.e. $S_0 = 1$), et que l'on connaît la durée moyenne $1/\gamma$ de l'infection par observation de cas cliniques, on notera que l'observation du taux de croissance exponentielle de l'épidémie dans ses premiers temps permet a priori d'identifier le paramètre R_0 .

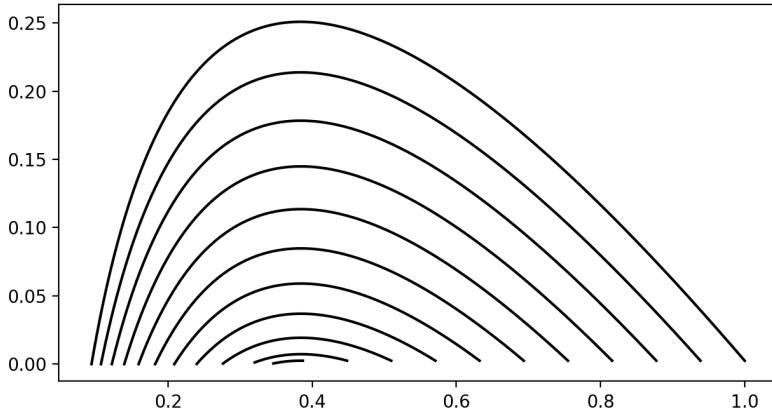


FIGURE 7.2 – Portrait de phase (S, I) (évolution de la droite vers la gauche) pour $R_0 = 2.6$, et différentes valeurs de S_0 entre $1/R_0$ et 1.

Comportement des solutions

On considère le système (7.2) réduit (2 premières équations), dans le cas $R_0 > 1$. Si des premiers cas se déclarent, cela revient à considérer une condition initiale du type $(S_0, I_0) = (1-\varepsilon, \varepsilon)$. L'évolution de I suit l'équation

$$\frac{dI}{dt} = +\beta SI - \gamma I = I\gamma(R_0S - 1).$$

Dans les premiers temps, I va rester petit, et S proche de 1, on a donc une croissance exponentielle au taux $\gamma(R_0 - 1)$. Lorsque la fraction de I devient significative, et la fraction de S s'éloigne conjointement de 1, le taux de croissance diminue, jusqu'à ce que la dérivée de I s'annule. Ce point correspond concrètement (exprimé en langage courant) au moment où le nombre de nouveaux cas s'équilibre avec le nombre de guérisons (et / ou de décès). À ce moment-charnière $S_c = 1/R_0$. Cela ne signifie pas que l'épidémie est terminée, mais qu'elle s'épuise d'une certaine manière, ne trouvant plus assez de S à contaminer pour continuer sa progression. Une fraction significative de personnes va néanmoins encore être infectée dans la phase terminale de l'épidémie, surtout si les individus, considérant que le pire est passé, relâchent leur attention, ce qui est susceptible d'augmenter la valeur du β , donc du R_0 , et potentiellement de faire repartir l'épidémie. Pour le cas d'un R_0 autour de 2.5, ce point correspond à $S_c = 0.4$, soit une proportion de 60 % de personnes qui ont été infectées, soit donc, pour une population totale de 67 millions, autour de 40 millions. Ensuite le nombre d'infectés diminue, jusqu'à converger exponentiellement vers 0. Le système passe ainsi d'un point d'équilibre instable à un point d'équilibre stable (voir figure 7.2).

Liens avec les modèles proies-prédateurs

Il est assez éclairant de voir le modèle SIR, ou simplement SI si l'on ne considère que les deux premières équations, comme un modèle proie-prédateur amputé du terme de naissance des proies. Dans ce modèle, appelé modèle de Lotka-Volterra¹, la variable S (on garde les notations du modèle SI) désigne la population de proies, et I la population de prédateurs. Le modèle complet s'écrit

$$\begin{cases} \frac{dS}{dt} = -\beta SI + \alpha S \\ \frac{dI}{dt} = +\delta SI - \gamma I \end{cases} \quad (7.5)$$

1. On trouvera des développements très complets autour de ce type de modèle dans *J.D Murray, Mathematical biology (Springer 1993)*.

Le terme δSI représente une *consommation* : les prédateurs doivent se nourrir de proies pour se développer. Ce mécanisme conduit à une diminution du nombre de proies, représenté par le terme βSI dans le système, avec en général $\delta \neq \beta$ (il n'y a pas de raison que la consommation d'une proie permette au prédateur d'engendrer exactement un individu). Le terme γI encode la mort naturelle des prédateurs (durée de vie $1/\gamma$), et le terme supplémentaire αS représente simplement la naissance de nouvelles proies, sous l'hypothèse, pour les proies, de ressources illimitées, ou d'un développement suffisamment raisonnable (grâce aux prédateurs) pour que la limitation des ressources ne se fasse jamais sentir. On prendra garde au fait que S et I représentent ici des populations constituées d'entités différentes, par exemple des requins et des sardines, alors que dans le modèle SI/SIR il s'agit d'individus semblables affectés d'une caractéristique qui se propage au sein de la population. Le système a néanmoins une structure proche, jusqu'à être parfaitement identique si l'on prend $\delta = \beta$ et $\alpha = 0$ (pas de génération spontanée des proies / pas de processus de reproduction des susceptibles à une échelle de temps qui est celle de l'épidémie). Le système de Lotka-Volterra (7.5) admet une intégrale première, i.e. il existe une fonction du couple (S, I) qui reste constante le long des trajectoires. Il s'agit de la fonction

$$(S, I) \mapsto F(S, I) = \delta S + \beta I - \gamma \log(S) - \alpha \log(I).$$

Dans le cas présent, on peut ainsi vérifier que la fonction

$$(S, I) \mapsto F(S, I) = \beta(S + I) - \gamma \log(S)$$

reste constante le long des trajectoires du modèle SI. On a en effet

$$\frac{d}{dt} F(S, I) = \beta(\dot{S} + \dot{I}) - \gamma \frac{\dot{S}}{S} = -\beta\gamma I - \gamma\beta I = 0.$$

Les trajectoires dans l'espace des phases (voir figure 7.2) sont donc des courbes isovaleurs de cette fonction. Dans ce cas dégénéré on n'a bien sûr plus existence de solutions périodiques, car ces courbes rencontrent en l'axe des S des points d'équilibre. Noter que l'existence de solutions périodiques reste assurée en théorie pour n'importe quel $\alpha > 0$, aussi petit soit-il. Si l'on prend en compte les naissances au sein de la population des susceptibles², on a en théorie un phénomène périodique, qui exprime le fait que, lorsque la maladie à essentiellement épuisé le capital de personnes susceptibles, il faut attendre un renouvellement de cette population pour que la maladie puisse se développer à nouveau. Ce phénomène n'est pas lié au caractère saisonnier de la grippe, qui lui est dû à des facteurs exogènes (température de l'air).

De façon plus informelle, on pourra s'inspirer de cette analogie pour considérer le virus comme un prédateur qui ne peut survivre qu'en se nourrissant de proies qu'il n'a pas encore dévorées (les individus ayant déjà été infectés ne sont plus *comestibles*). On peut ainsi espérer le détruire en l'affamant (i.e. en lui coupant l'accès aux proies par des mesures de cloisonnement), mais en gardant à l'esprit que redonner l'accès d'une population massive de proies à quelques I résiduels peut faire repartir la population des prédateurs.

Instabilités gelées

La notion d'instabilité de points d'équilibre dans le contexte de l'épidémiologie prend un sens différent de celui qu'il a par exemple en mécanique. Un stylo posé à la verticale sur une table est en position instable, et toute expérience visant à explorer la stabilité de cette configuration conduira à la chute du stylo, de telle sorte que les positions d'équilibre instable ne se voient jamais dans la réalité. Dans un contexte d'épidémiologie, une petite variation autour du point d'équilibre correspond à l'apparition de quelques individus infectés, voire un seul, de telle sorte que des points d'équilibre instables peuvent parfaitement être viables et donc observés, et sont d'une certaine manière *stables* puisque leur activation nécessite cette nucléation originelle qui peut ne pas se produire. Nous pouvons ainsi considérer une situation sans épidémie comme un point d'équilibre instable vis à vis d'affections

2. Il faudrait en toute rigueur revenir au modèle SIR, et considérer un terme de naissance du type $\alpha(S + R)$, puisque les personnes ayant contracté la maladie gardent la capacité après guérison de se reproduire.

comme la peste, le choléra, voire de toute affection virtuelle que l'on pourrait imaginer. Nous avons vu que l'évolution spontanée d'une épidémie fait passer d'un point d'équilibre instable à un point d'équilibre stable (au prix d'une réduction significative de la population des S , donc d'un nombre important de personnes infectées, et donc d'un nombre de morts importants), mais une politique sanitaire agressive peut aussi viser à éradiquer la maladie en amenant (en modifiant en particulier le paramètre β par des mesures de confinement provisoires, par exemple) le couple (S, I) à un point $(S_0, 0)$ qui serait instable pour la valeur finale (élevée) du paramètre β , qui correspond à la situation où toutes les mesures sanitaires ont été levées. Cette stratégie est évidemment périlleuse puisque quelques cas (qui pourraient venir de lieux où l'épidémie s'est déclarée plus tard) peuvent suffire à faire redémarrer l'épidémie.

7.2 Modèle stochastique orienté agent

On s'intéresse ici à un modèle basé sur un suivi individuel de la situation de chaque individu³. On note V l'ensemble des individus, vus comme les sommets d'un graphe, et l'on considère que chaque sommet peut être dans l'état S , I , ou R . On note $u_x^n \in \{S, I, R\}$ l'état de x au jour $n \in \mathbb{N}$. On note K_{xy} le nombre de contacts un jour donné entre x et y (on suppose pour simplifier les notations que ce nombre de contact ne varie pas d'un jour à l'autre), et l'on note $p > 0$ la probabilité de contamination entre un infecté et un susceptible lors d'un contact. Si x est S et y est I , la probabilité que I contamine x lors de K_{xy} contacts est

$$1 - (1 - p)^{K_{xy}}.$$

Un jour donné, on considèrera les contaminations éventuelles de x par un y infecté comme des événements indépendants. On définit le graphe pondéré non orienté $G = (V, E, K)$ en considérant chaque individu comme un sommet du graphe, avec $(x, y) \in E \iff K_{xy} \neq 0$, ce que l'on peut aussi écrire $E = \text{supp}(K)$ (E est le support de K , i.e. l'ensemble des (x, y) tels que $K_{xy} \neq 0$).

L'état de la population au jour n est donc décrit par la vecteur aléatoire

$$U^n = (U_x^n)_{x \in V}, \quad U_x^n \in \{S, I, R\}.$$

On suppose que, partant d'un état donné au jour n , les changements d'états pour les individus sont des événements indépendants⁴, qui ne dépendent que des personnes avec qui l'on a été en contact au jour $n + 1$, et de son propre état au jour n , ce qui s'écrit

$$\begin{aligned} \mathbb{P}(U^{n+1} = u^{n+1} | U^n = u^n, U^{n-1} = u^{n-1}, \dots, U^0 = u^0) = \\ \mathbb{P}(U^{n+1} = u^{n+1} | U^n = u^n) = \prod_V \mathbb{P}(U_x^{n+1} = u_x^{n+1} | U^n = u^n). \end{aligned}$$

L'évolution du système est donc entièrement déterminée par les probabilités conditionnelles sui-

3. Il est évident qu'un tel modèle à l'échelle de la population d'un pays ou du monde est destiné à rester essentiellement théorique du fait de l'impossibilité d'accéder à l'ensemble des paramètres qui le constitue, et de la difficulté qu'il y aurait à confronter ses résultats avec des mesures individuelles. Nous le proposons néanmoins comme exercice de réflexion autour des mécanismes qui conditionnent la propagation d'un maladie à l'échelle individuelle, ou d'un petit groupe d'individus.

4. Les contaminations peuvent être vues comme s'effectuant en parallèle. On exclut donc en particulier le fait qu'un individu contaminé le jour $n + 1$ puisse contaminer une autre personne plus tard dans la même journée.

vantes :

$$\begin{aligned}
\mathbb{P}(U_x^{n+1} = S | U^n = u^n) &= \prod_{y \sim x, u_y^n = I} (1-p)^{K_{xy}} && \text{si } U_x^n = S, \\
\mathbb{P}(U_x^{n+1} = I | U^n = u^n) &= 1 - \prod_{y \sim x, u_y^n = I} (1-p)^{K_{xy}} && \text{si } U_x^n = S, \\
\mathbb{P}(U_x^{n+1} = R | U^n = u^n) &= 0 && \text{si } U_x^n = S, \\
\mathbb{P}(U_x^{n+1} = S | U^n = u^n) &= 0 && \text{si } U_x^n = I, \\
\mathbb{P}(U_x^{n+1} = I | U^n = u^n) &= 1 - \gamma && \text{si } U_x^n = I, \\
\mathbb{P}(U_x^{n+1} = R | U^n = u^n) &= \gamma && \text{si } U_x^n = I, \\
\mathbb{P}(U_x^{n+1} = S | U^n = u^n) &= 0 && \text{si } U_x^n = R, \\
\mathbb{P}(U_x^{n+1} = I | U^n = u^n) &= 0 && \text{si } U_x^n = R, \\
\mathbb{P}(U_x^{n+1} = R | U^n = u^n) &= 1 && \text{si } U_x^n = R,
\end{aligned} \tag{7.6}$$

où γ est le taux journalier de guérison (ou de décès), que l'on peut assimiler à l'inverse de la durée de la maladie.

Dans ce qui précède nous avons implicitement considéré le graphe sous-jacent comme statique. Une approche plus réaliste consisterait à intégrer la possibilité que ce graphe puisse évoluer d'un jour à l'autre. Une personne voyageant d'un pays à l'autre va ainsi voir son réseau de contacts complètement remis à jour. Par ailleurs, même dans le cas d'un comportement routinier, par exemple prise d'un train aux mêmes horaires tous les jours, l'ensemble des personnes que l'on est amené à côtoyer est renouvelé.

Question 7.2. Le fait de décrire de façon markovienne le passage de I à R n'est pas très réaliste, cela revient à considérer que la durée de la maladie suit une loi exponentielle. On peut affiner cet aspect du modèle de la façon suivante : on peut prendre en compte la période écoulée depuis que la maladie a été contractée en rajoutant une variable réelle (ou discrète) pour suivre le temps écoulé depuis le début de l'infection, pour les individus de type I . Proposer un nouveau modèle d'évolution prenant en compte cet aspect (on pourra considérer que la durée de la maladie suit une loi connue, ou plus simplement qu'elle est la même pour tout le monde).

Question 7.3. Dans ce qui précède nous avons implicitement considéré le graphe sous-jacent comme statique. Une approche plus réaliste consisterait à intégrer la possibilité que ce graphe puisse évoluer d'un jour à l'autre. Une personne voyageant d'un pays à l'autre va ainsi voir son réseau de contacts complètement remis à jour. Par ailleurs, même dans le cas d'un comportement routinier, par exemple prise d'un train aux mêmes horaires tous les jours, l'ensemble des personnes que l'on est amené à côtoyer est renouvelé. On pourra réfléchir à des moyens d'intégrer au modèle ces considérations, en proposant éventuellement des idées pour définir des graphes aléatoirement à chaque étape du processus (i.e. chaque jour), de façon à traiter telle ou telle problématique particulière (transport en commun, individus voyageant d'un pays / ville à l'autre, réunion de travail ponctuelle, ...).

7.3 Modèle déterministe portant sur des probabilités d'infection

Nous présentons ici un modèle déterministe portant sur les probabilités pour un individus d'être dans tel ou tel état. Cette approche permet de construire formellement le modèle (7.2). On considère une population de N individus. On cherche à construire un modèle d'évolution discret (de jour en jour pour fixer les idées) sur les probabilités que les agents soient dans tel ou tel état. On note S_x^k , I_x^k , $R_x^k = 1 - S_x^k - I_x^k$ les probabilité que x soit, à l'étape k dans l'état Susceptible, Infecté, ou Remis, respectivement. Les événements susceptibles de modifier le statut d'un individus (de S à I) sont des contacts avec une ou des personnes infectées. On ne prend pas en compte la modification de statut des entités au sein d'une même journée (plus généralement d'une étape du modèle), comme si à la fin de chaque jour on fait le bilan sur les contacts subis, et que l'on modifie à ce moment la probabilité

d'être infecté. Ainsi une personne qui est sûrement S au début d'un jour donné ne peut commencer à infecter d'autres personnes qu'à partir du jour suivant.

Construction du modèle

Supposons que x soit de type S . Lors d'un contact avec un y infecté, sa probabilité que la maladie se transmette est notée $p > 0$.

Si y a une probabilité I_y d'être infecté, la probabilité que x ne soit pas infecté s'écrit $1 - pI_y$. La probabilité que x soit infecté lors de K_{xy} contacts avec y est donc le complémentaire de la probabilité de n'avoir été infecté par aucun des contacts est donc

$$1 - (1 - pI_y)^{K_{xy}}.$$

Si x est infecté, sa probabilité de guérir est $\gamma > 0$, ou $1/\gamma$ est la durée moyenne de la maladie.

Remarque 7.4. Précisons que le modèle ci-dessous n'est pas obtenu rigoureusement à partir du modèle stochastique individuel de la section précédente. En particulier, la probabilité que x ne soit infecté par aucun des y rencontrés est écrite comme produit des probabilités que y n'infecte pas x pour tout y . Or cette probabilité dépend de la probabilité que y soit lui-même infecté, et ces probabilités ne sont *pas indépendantes*.

Modèle global

On note K_{xy}^n le nombre de contacts entre x et y le jour n . Le modèle d'évolution discret s'écrit

$$\begin{aligned} S_x^{n+1} &= S_x^n \prod_{y \sim x} (1 - pI_y^n)^{K_{xy}^n} \\ I_x^{n+1} &= S_x^n \left(1 - \prod_{y \sim x} (1 - pI_y^n)^{K_{xy}^n} \right) + (1 - \gamma)I_x^n \\ R_x^{n+1} &= R_x^{n+1} + \gamma I_x^n \end{aligned} \tag{7.7}$$

Proposition 7.5. On se donne des conditions initiales S^0 , I^0 , et R^0 vérifiant

$$S_x^0, I_x^0, R_x^0 \in [0, 1], \quad S_x^0 + I_x^0 + R_x^0 = 1 \quad \forall x \in V.$$

La solution du problème (7.7) vérifie alors, pour tout $n \in \mathbb{N}$, les mêmes propriétés, i.e.

$$S_x^n, I_x^n, R_x^n \in [0, 1], \quad S_x^n + I_x^n + R_x^n = 1 \quad \forall x \in V.$$

Démonstration. D'après l'hypothèse et la définition, on a $S_x^1 \in [0, 1]$ pour tout x . De même $I_x^1 \geq 0$, et de la forme $aS_x^0 + bI_x^0$, avec a et b dans $[0, 1]$, et $a + b \leq 1$, d'où $I_x^1 \leq 1$. Enfin $R_x^1 \in [0, 1]$ grâce aux mêmes arguments. La conservation de la somme des $S_x + I_x + R_x$ se vérifie immédiatement en sommant les relations de récurrence. La propriété pour tout n s'établit par récurrence. \square

Modèle semi-linéarisé

On peut obtenir formellement à partir de ce modèle discret un système discret plus simple, en considérant que l'on a fait un choix d'unités pour les contacts tel que p est petit devant 1, et que le pas de temps choisi pour le modèle (des jours en l'occurrence) est tel que l'on a aussi pK_{xy} petit devant 1 pour tous (x, y) . On a

$$1 - \prod_{y \sim x} (1 - pI_y^n)^{K_{xy}^n} \approx p \sum_{y \sim x} K_{xy} I_y^n$$

On obtient ainsi formellement le système

$$\begin{aligned} S_x^{n+1} - S_x^n &= -p S_x \sum_{y \sim x} K_{xy} I_y^n \\ I_x^{n+1} - I_x^n &= p S_x \sum_{y \sim x} K_{xy} I_y^n - \gamma I_x^n \\ R_x^{n+1} - R_x^n &= \gamma I_x^n. \end{aligned} \tag{7.8}$$

que l'on peut écrire sous forme continue en temps :

$$\begin{aligned} \frac{dS_x}{dt} &= -p S_x \sum_{y \sim x} K_{xy} I_y \\ \frac{dI_x}{dt} &= p S_x \sum_{y \sim x} K_{xy} I_y - \gamma I_x \\ \frac{dR_x}{dt} &= \gamma I_x. \end{aligned} \tag{7.9}$$

Matrices d'incidence

Nous introduisons ici un objet de type matrice permettant de donner une représentation globale de la propagation d'une épidémie au sein d'une population, décrite par un modèle de type (7.7). Le principe en est simple, pour tout sommet $y \in V$, on considère la condition initiale $I^0 = \delta_y$ (i.e. $I_y = 1$, et $I_x = 0$ pour tout $x \neq y$). On résout numériquement le modèle d'évolution (7.7) jusqu'à un temps T pré-fixé (par exemple une dizaine de jours). On obtient alors un vecteur de probabilités d'infections pour toute la population⁵, que l'on note $I^y = (I_x^y)_{x \in V}$. On construit alors la matrice \mathcal{J} dont la y -ième colonne est I^y . Cette matrice représente l'univers des conséquences épidémiologiques des différents scénarios possibles en termes de patient zéro. Elle permet d'estimer des risques individuels d'exposition et des niveaux de dangerosité. Plus précisément, si l'on D_y la moyenne des éléments de la y -ième colonne de la matrice $D_y \in [0, 1]$ quantifie le risque causé potentiellement par y (dangerosité), puisqu'il représente la probabilité moyenne qu'un individu de la population soit infecté directement ou indirectement dans les T jours suivant à l'arrivée de la maladie par y . Dans l'autre sens, la moyenne E_x des éléments de la x -ème ligne représente le risque que x lui-même soit infecté, moyenné sur l'ensemble des scénarios possibles en termes de patient 0.

Taux de reproduction individuel

Nous proposons ici de donner un sens, au travers du modèle (7.7), à la notion de taux de reproduction *individuel*. L'idée consiste simplement, pour un x fixé, à évaluer dans le cadre du modèle (7.7) la quantité de personnes contaminées par x . Conformément à la définition usuelle, ces contaminations doivent être *directes*. Un moyen de se limiter à ces contaminations directes consiste à désactiver les contacts entre y et z dès que les deux sont différents de x . On résout alors le modèle (7.7) avec cette nouvelle matrice, jusqu'à un temps T de l'ordre de la durée de la maladie, à partir de la condition initiale $I^0 = \delta_x$, et l'on définit le taux de reproduction associé R_x comme la somme des coefficients de I^T . Ce nombre peut être interprété comme l'espérance du nombre de personnes contaminées par x .

Équation de réaction-diffusion sous jacente

Considérons la deuxième équation du système (7.10) :

$$\frac{dI_x}{dt} = p S_x \sum_{y \sim x} K_{xy} I_y - \gamma I_x.$$

5. Comme il s'agit pour chaque x de mesurer le risque qu'il ait été infecté, on pourra considérer $I_x + R_x$ plutôt que I_x . Nous conservons néanmoins la notation I pour désigner le vecteur correspondant.

Si l'on se place au voisinage de l'émergence de l'épidémie (S_x proche de 1 pour tout x), on peut approcher cette équation par

$$\frac{dI_x}{dt} = p \sum_{y \sim x} K_{xy} I_y - \gamma I_x = -p \sum_{y \sim x} K_{xy} (I_x - I_y) + p K_{x\bar{y}} I_x - \gamma I_x,$$

que l'on peut écrire comme une équation sur le vecteur I

$$\frac{dI}{dt} + pLI = pK_{\bar{y}}I - \gamma I,$$

où $K_{\bar{y}}$ est la matrice diagonale dont les coefficients sont les

$$K_{x\bar{y}} = \sum_{y \sim x} K_{xy},$$

et L est la matrice du Laplacien associé au réseau pondérés par les contacts K . On obtient l'analogue discret d'une équation aux dérivées partielles très classique, dite de réaction-diffusion (dans le cas simplement d'un terme de réaction linéaire) :

$$\frac{\partial u}{\partial t} - p\Delta u = Ku - \gamma u.$$

Cette analogie permet d'interpréter le terme pLI comme un terme de diffusion au travers du réseau, un terme source de croissance de l'épidémie $pK_{\bar{y}}I$ (avec des taux de croissance individuels proportionnels aux nombres de contact des individus), et un terme de disparition qui encode le mécanisme de guérison.

Taux de reproduction comme *spectre*

On se place toujours ici dans le cas d'une épidémie émergente, avec S_x proche de 1 pour tout x (et donc I_x proche de 0), on considère l'équation linéarisée

$$\frac{dI}{dt} = \gamma \left(\frac{p}{\gamma} K - \text{Id} \right) I,$$

où K est la matrice des contacts par unité de temps, p la probabilité d'être infecté lors d'un contact, $\gamma = 1/D$ l'inverse de la durée de la période contagieuse.

Le taux de reproduction de base défini à partir de la version scalaire de l'équation précédente apparaît maintenant comme une matrice pDK . Il est alors naturel d'assimiler le taux de reproduction de base \mathcal{R}_0 au spectre de la matrice symétrique réelle pDK

$$\mathcal{R}_0 = \text{Sp}(pDK) \subset \mathbb{R}.$$

Il suffit que l'une des valeurs propres de cette matrice soit strictement supérieure à 1 pour que le système soit instable. On a stabilité asymptotique de ce système dès que toutes les valeurs propres de la matrice

$$pK - \gamma \text{Id} = \gamma(pDK - \text{Id})$$

sont strictement négatives, comme l'exprime la proposition suivante.

Proposition 7.6. Les solutions de l'équation

$$\frac{dI}{dt} = \gamma \left(\frac{p}{\gamma} K - \text{Id} \right) I,$$

tendent exponentiellement vers 0 quelle que soit la condition initiale si et seulement toutes les valeurs propres de la matrice pDK sont strictement inférieures à 1.

Démonstration. La matrice $R = pDK$ est symétrique, toutes ses valeurs (notées (λ_i)) propres sont donc réelles, et elle admet une base de vecteurs propres orthogonaux (w_i) . Soit $I^0 = (I_x^0)_V$ une condition initiale, et $I(t)$ la solution associée. On peut décomposer cette solution pour tout t sur la base des vecteurs propres

$$I = \sum_{j=1}^N \alpha_j(t) w_j.$$

Les w_i étant orthogonaux entre eux et normés, on a

$$\frac{d\langle I | w_i \rangle}{dt} = \frac{d\alpha_i}{dt} = w_i \cdot \frac{dI}{dt} = w_i \cdot \gamma \left(\frac{p}{\gamma} K - \text{Id} \right) \sum_{j=1}^N \alpha_j(t) w_j = \gamma w_i \cdot \sum_{j=1}^N (\lambda_j - 1) \alpha_j w_j = \gamma(\lambda_i - 1) \alpha_i.$$

On obtient ainsi une collection d'équations différentielles ordinaires indépendantes

$$\dot{\alpha}_i = \gamma(\lambda_i - 1) \alpha_i \implies \alpha_i = e^{\gamma(\lambda_i - 1)t} \alpha_i^0,$$

d'où la propriété annoncée. \square

Dans le cas d'un réseau général, il peut être naturel d'appeler taux de reproduction de base la plus grande valeur propre de la matrice pKD .

Remarque 7.7. On notera que le vecteur propre associé à la plus grande valeur propre est *réaliste* en tant que collection de probabilités (qui doivent être positives). En effet, dans le cas d'un réseau connexe (mais le résultat s'étend au cas général) le théorème de Perron-Frobenius (théorème ??, page ??) assure que le vecteur propre associé à la plus grande valeur propre d'une matrice ne contenant que des coefficients positifs peut être choisi de telle sorte que tous ses coefficients soient de même signe.

Phénomène d'advection-réaction-diffusion

On peut pousser plus loin cette démarche consistant à identifier dans l'équation sur I des mécanismes de propagation courants dans le domaine des équations aux dérivées partielles. Comme précédemment, le point de départ est la deuxième équation du système (7.10) :

$$\frac{dI_x}{dt} = p S_x \sum_{y \sim x} K_{xy} I_y - \gamma I_x,$$

que l'on considère à un moment arbitraire de l'évolution, sans supposer que S_x est proche de 1, ni bien sûr I_x proche de 0. On écrit la somme

$$p S_x \sum_{y \sim x} K_{xy} I_y = -p S_x \sum_{y \sim x} K_{xy} (I_x - I_y) + p K_{x\bar{y}} S_x I_x,$$

de telle sorte que l'équation d'évolution devient

$$\frac{dI_x}{dt} = -p S_x \sum_{y \sim x} K_{xy} (I_x - I_y) + p K_{x\bar{y}} S_x I_x - \gamma I_x.$$

Reconnaissons que le terme de réaction du type *SI*, qu'il est satisfaisant de voir apparaître ici, comme dans le modèle de départ, n'a pas une signification très claire dans le présent modèle qui porte sur des probabilités individuelles⁶. On verra néanmoins que ce terme prendra tout son sens dans le contexte de modèles agrégés, où un sommet du réseau représentera non plus un individu unique, mais un groupe d'individus : le terme $S_x I_x$ correspondra alors bien à une contamination des S du groupe par les I du même groupe. Le terme que nous avons identifié comme encodant un phénomène de diffusion au travers du Laplacien discret, est maintenant multiplié par les valeurs locales du champ S .

Dans le contexte des Équations aux Dérivées Partielles, cela correspondrait à un terme du type $-S\Delta I$, qui peut s'écrire (les variables S et I correspondent ci-dessous à des champs continus sur un domaine)

$$-S\Delta I = -\nabla \cdot S \nabla I + \nabla S \cdot \nabla I.$$

On fait apparaître un terme de diffusion standard, avec un coefficient de diffusion non uniforme. Ce terme correspond à phénomène conservatif, qui ne fait qu'encoder une propagation des I ; en effet, si on l'intègre sur un sous-domaine ω , on obtient (en notant ‘1’ la fonction constante égale à 1 sur le domaine)

$$-\int_{\omega} \nabla \cdot S \nabla I = -\int_{\omega} \nabla \cdot S \nabla I \times 1 = \int_{\omega} S \nabla I \cdot \nabla 1 - \int_{\partial\omega} S \frac{\partial I}{\partial n} = -\int_{\partial\omega} S \frac{\partial I}{\partial n}$$

qui se réduit donc à un terme d'échange avec le monde extérieur à ω , selon un flux qui s'écrit $J = -S \nabla I$ (loi de Ficke). Le second terme, du type $v \cdot \nabla I$, correspond à un transport non conservatif de I par le champ de vitesse $v = \nabla S$. On peut interpréter ce terme comme encodant un phénomène de *chimiotaxie*, i.e. qui correspond à un transport vers les zones où S est le plus important, comme si les I étaient *attirés* par les S . L'équation globale continue correspondant au modèle discret s'écrit ainsi

$$\frac{\partial I}{\partial t} - \nabla \cdot S \nabla I + \nabla S \cdot \nabla I = \beta SI - \gamma I.$$

Retournons au modèle sur graphe. On peut, en forçant un peu les choses, mener une version discrète de la démarche précédente. En effet

$$-p S_x \sum_{y \sim x} K_{xy} (I_x - I_y) = -p \sum_{y \sim x} K_{xy} \frac{S_x + S_y}{2} (I_x - I_y) + p \sum_{y \sim x} \frac{S_x - S_y}{2} (I_y - I_x).$$

On fait bien apparaître un opérateur de Laplacien sur graphe pondéré, avec un poids pour l'arête (x, y) égal à $K_{xy}(S_x + S_y)/2$. Le dernier terme peut s'interpréter comme un terme de transport chimiotactique. Cette interprétation est particulièrement limpide en dimension 1 : on se place dans le cas d'un réseau “linéaire” $\dots j-1 \longleftrightarrow j \longleftrightarrow j+1 \longleftrightarrow \dots$, en considérant que tous les poids K_{xy} sont égaux. On obtient pour le point $x = j$

$$\sum_{y \sim x} \frac{S_x - S_y}{2} (I_y - I_x) = -\frac{1}{2} ((S_{j+1} - S_j)(I_{j+1} - I_j) + (S_j - S_{j-1})(I_j - I_{j-1})).$$

Si l'on considère que les I_j et S_j sont les interpolées de fonctions régulières en des points de la droite réelle $x_j = j\Delta x$, on a

$$-\frac{1}{2} ((S_{j+1} - S_j)(I_{j+1} - I_j) + (S_j - S_{j-1})(I_j - I_{j-1})) = -S'(x_j)I'(x_j) + \mathcal{O}(\Delta t),$$

qui correspond bien (le signe ‘-’ disparaît quand on repasse le terme au membre de gauche), à un transport non conservatif de I à la vitesse S' .

Modèle Homogène

On se propose ici de retrouver le modèle SIR scalaire classique à partir du modèle discret (7.8), sous certaines conditions d'homogénéité. On se place dans le cas où tous les individus ont le même nombre de contacts, plus précisément la même durée de contact totale par jour (le nombre d'individus rencontrés peut varier), c'est à dire que

$$\sum_{y \sim x} K_{xy}$$

ne dépend pas de x . On note simplement \bar{K} cette valeur commune. On suppose les probabilités initiales uniformes (indépendantes de i), on vérifie immédiatement qu'elles le restent, on obtient donc un système simplifié, que l'on écrit ici simplement pour la classe I :

$$I^{n+1} - I^n = S^n p \sum_{y \neq x} K_{xy} I^n - \gamma I^n = p \bar{K} S^n I^n - \gamma I^n. \quad (7.10)$$

6. C'est un peu comme si la “partie” probablement saine d'une personne était infectée par sa partie probablement malade

qui est la forme discrète de l'équation différentielle

$$\frac{dI}{dt} = p\bar{K}SI - \gamma I = \beta SI - \gamma I. \quad (7.11)$$

On retrouve ainsi le système scalaire de départ (7.2).

Modèles à compartiments

Le modèle SIR et ses extensions sont appelés couramment modèles à compartiments, les compartiments correspondant aux différents états des individus (S , I , ou R). On peut aller plus loin dans cette décomposition de la population en compartiments, à partir du modèle centré sur les individus, en regroupant des sous-groupe d'individus selon certains critères statiques (qui caractérisent un état pérenne), par exemple une classe d'âge, ou une catégorie socio-professionnelle. Nous nous plaçons comme précédemment dans le cas d'une population de N individus, identifiée à un ensemble de sommets V , et un ensemble d'arêtes (non orientées) encodant la présence de contacts entre les individus-sommets. La matrice K_{xy} , dont E est le support, contient le nombre de contacts entre x et y . Considérons une partition de l'ensemble V , et la relation d'équivalence associée : deux sommets sont en relation s'ils appartiennent à la même partie. On notera \bar{x} la classe de x . On se propose d'élaborer un modèle “à gros grain”, qui porte sur les probabilités d'être dans tel ou tel état au sein de chaque classe, en partant du modèle d'évolution discret en temps qui facilite l'approche. On considère une situation où les probabilités d'être dans les différents états sont initialement uniformes au sein de chaque classe, on les notera ainsi $S_{\bar{x}}$, $I_{\bar{x}}$ et $R_{\bar{x}}$. On suppose que le nombre total de contacts entre un individu donné $x \in \bar{x}$ et des individus de \bar{y} ne dépend pas de x , y compris lorsque $\bar{y} = \bar{x}$, et l'on note $K_{\bar{x}\bar{y}}$ la quantité (qui ne dépend donc pas du représentant choisi)

$$K_{\bar{x}\bar{y}} = K_{x\bar{y}} = \sum_{y \in \bar{y}} K_{xy}.$$

On peut alors écrire, pour tout $x \in \bar{x}$

$$I_x^{n+1} - I_x^n = p S_x^n \sum_{\bar{y}} \sum_{y \in \bar{y}} K_{xy} I_y^n - \gamma I_x^n$$

Si l'on suppose qu'à l'étape k les probabilités sont uniformes au sein de chaque classe, on a

$$I_x^{k+1} - I_x^k = p S_x^k \sum_{\bar{y}} I_{\bar{y}}^k \sum_{y \in \bar{y}} K_{xy} - \gamma I_x^k = p S_{\bar{x}}^k \sum_{\bar{y}} I_{\bar{y}}^k K_{\bar{x}\bar{y}} - \gamma I_x^k.$$

La quantité I_x^{k+1} ne dépend donc pas du représentant x choisi dans la classe \bar{x} , on peut donc l'écrire $I_{\bar{x}}^{k+1}$. On montre de la même manière que S_x^{k+1} et R_x^{k+1} ne dépendent pas du représentant choisi, et ces propriétés d'uniformité au sein de chaque classe sont vérifiées immédiatement par récurrence. On obtient donc le système “condensé” (que l'on écrit maintenant sous forme continue en temps)

$$\begin{aligned} \frac{dS_{\bar{x}}}{dt} &= -p S_{\bar{x}} \sum_{\bar{y}} K_{\bar{x}\bar{y}} I_{\bar{y}} \\ \frac{dI_{\bar{x}}}{dt} &= p S_{\bar{x}} \sum_{\bar{y}} K_{\bar{x}\bar{y}} I_{\bar{y}} - \gamma I_{\bar{x}} \\ \frac{dR_{\bar{x}}}{dt} &= \gamma I_{\bar{x}}. \end{aligned}$$

Noter que la somme qui encode les interactions contient maintenant un terme diagonal, qui prend en compte les contacts entre individus au sein d'une même classe.

On notera que l'élaboration de ce modèle à compartiments généralisés s'est basée sur des hypothèses très fortes d'homogénéité.

7.4 Développements, extensions

Prise en compte de la période d'incubation : le modèle SEIR

Cette extension du modèle initial est basée sur l'introduction d'une variable supplémentaire quantifiant le nombre de personnes E (pour *Exposed*) ayant contracté la maladie, mais sans être contagieux. On introduit un nouveau paramètre α représentant le taux de passage des E vers les I : α est l'inverse du temps moyen de la période d'incubation.

$$\begin{cases} \frac{dS}{dt} = -\beta SI \\ \frac{dE}{dt} = +\beta SI - \alpha E \\ \frac{dI}{dt} = +\alpha E - \gamma I \\ \frac{dR}{dt} = +\gamma I \end{cases} \quad (7.12)$$

Comme précédemment, on s'intéressera au système réduit sans la dernière équation (elle n'est que faiblement couplée au système). Ce système admet une infinité de points fixes $(S_0, 0, 0)$ pour $S_0 \in [0, 1]$. La matrice du système linéarisé en un tel point fixe s'écrit

$$\begin{pmatrix} 0 & 0 & -\beta S_0 \\ 0 & -\alpha & \beta S_0 \\ 0 & \alpha & -\gamma \end{pmatrix},$$

de polynôme caractéristique

$$P(\lambda) = -\lambda (\lambda^2 + (\alpha + \gamma)\lambda + \alpha(\gamma - \beta S_0)).$$

Les valeurs propres sont 0 et

$$\lambda^\pm = \frac{1}{2}(\alpha + \gamma) \left(-1 \pm \sqrt{1 - 4 \frac{\alpha(\gamma - \beta S_0)}{(\alpha + \gamma)^2}} \right).$$

SEIR \rightarrow SIR

Remarquons en premier lieu que, lorsque α tend vers l'infini (i.e. la durée de la période d'incubation tend vers 0), la valeur propre de plus grande partie positive s'écrit

$$\begin{aligned} \lambda &= \frac{1}{2}(\alpha + \gamma) \left(-1 + \left(1 - 2 \frac{(\gamma - \beta S_0)}{\alpha} + o(1/\alpha) \right) \right) \\ &= \beta S_0 - \gamma + o(1/\alpha). \end{aligned}$$

On retrouve bien au premier ordre en $1/\alpha$ la valeur propre correspondant au modèle SIR, ce qui était attendu puisque faire tendre α vers $+\infty$ revient à considérer les E sont *éphémères* : ils se transforment immédiatement en I dès leur apparition.

Critère de stabilité pour le modèle SEIR

On a une valeur propre à partie réelle strictement positive si et seulement si

$$\beta S_0 - \gamma = \gamma(R_0 S_0 - 1) > 0.$$

On retrouve donc la même condition de développement de l'épidémie que dans le cas SIR : pour $S_0 = 1$, on a instabilité dès que $R_0 > 1$.

Temps caractéristique

Si, comme nous l'avons vu, le critère de développement de l'épidémie est le même que pour le modèle SIR, le temps caractéristique de développement de l'épidémie est modifié. Étudions en premier lieu le cas extrême d'une période d'incubation très longue par rapport à la phase post incubation (i.e. $1/\alpha$ grand devant $1/\gamma$, dans le cas instable). En faisant tendre α vers 0^+ , on obtient pour la valeur propre la plus instable

$$\lambda = \frac{1}{2}(\gamma + \alpha) \left(-1 \pm \sqrt{1 - 4\frac{\alpha(\gamma - \beta S_0)}{\gamma^2} + o(\alpha)} \right) = \alpha(R_0 S_0 - 1) + o(\alpha).$$

On trouve donc au premier ordre

$$\tau = \frac{1}{\alpha} \frac{1}{R_0 S_0 - 1}$$

c'est à dire une expression analogue au cas SIR (7.4), où la durée d'infection $1/\gamma$ est remplacée par la durée de la période d'incubation $1/\alpha$.

Modélisation affinée de la durée d'infection

Comme évoqué dans la remarque 7.1, page 150, les modèles différentiels présentés précédemment reposent implicitement sur une distribution exponentielle de la durée d'infection, ce qui signifie que chaque patient infecté a une même probabilité de guérir dans le jour qui vient, indépendamment du temps pendant lequel il a déjà été malade. Si l'on considère à l'autre extrême que la maladie a une durée fixe $D = 1/\gamma$, identique pour tous les patients, on peut simplement écrire que les gens qui guérissent à instant t donné sont exactement ceux qui sont tombés malades au temps $t - D$:

$$\begin{cases} \Phi(t) &= \beta S(t) I(t) \\ \frac{dS}{dt} &= -\Phi(t) \\ \frac{dI}{dt} &= +\Phi(t) - \Phi(t - D) \\ \frac{dR}{dt} &= +\Phi(t - D) \end{cases} \quad (7.13)$$

avec la convention que Φ est nulle sur les temps négatifs. Il s'agit d'un système différentiel avec retard, dont l'étude générale sort du cadre de ce cours, mais qui peut être approché numériquement très facilement.

Une approche intermédiaire consiste à considérer que les durées de maladie suivent une certaine loi $\delta(\cdot)$ sur \mathbb{R}^+ , et à écrire le modèle sous la forme d'un système intégré-différentiel, dont nous n'écrivons ici que la deuxième équation :

$$\frac{dI}{dt} = +\Phi(t) - \int_0^{+\infty} \delta(s) \Phi(t - s) ds \quad (7.14)$$

avec $\Phi(t) = \beta S(t) I(t)$ comme précédemment. Le système (7.13) correspond au cas limite où la distribution δ est un Dirac en D .

7.5 Prise en compte de la perte d'immunité

Il a été implicitement supposé précédemment qu'une personne qui a été malade ne le sera plus jamais. Si l'on considère maintenant qu'un ex malade est susceptible de perdre son immunité au bout

d'un temps $D_i = 1/\delta$, i.e. on autorise un retour de la classe R à la classe S , la démarche précédente conduit au système

$$\left\{ \begin{array}{lcl} \frac{dS}{dt} & = & -\beta SI + \delta R \\ \frac{dI}{dt} & = & +\beta SI - \gamma I \\ \frac{dR}{dt} & = & +\gamma I - \delta R \end{array} \right. \quad (7.15)$$

Exercice 7.4. Écrire un modèle dans l'esprit de (7.13) permettant de modéliser une durée d'immunité fixe $D_i > 0$, ainsi qu'une durée de maladie fixe. Dans un second temps, on pourra proposer un modèle plus réaliste, dans l'esprit de (7.14), basé sur une modélisation des durées d'infection et d'immunité sous forme de distributions sur \mathbb{R}_+ .

Matrices de contacts / distances

On note $K = (K_{ij})$ la matrice de contacts associés à une certaine population pendant une période de temps donnée. Dans une optique de structuration des individus, il peut être fécond de définir une distance entre eux qui soit respectueuse des aspects épidémiologiques. On cherche donc une correspondance

$$F : K \longmapsto D = F(K)$$

qui à un nombre de contact K associe une distance D , d'autant plus petite que le nombre de contacts est important : F est décroissante. Il semble illusoire de trouver une telle fonction F qui conduirait à une distance (qui vérifie en particulier l'inégalité triangulaire) sur l'espace discret des individus. On procède donc en 2 temps, en recherchant une fonction F décroissante qui permettra de définir des poids (= longueurs) d'arêtes sur les arêtes du graphe des contacts, et l'on construit la métrique globale comme la métrique des plus courts chemins associés au graphe ainsi construit.

Pour des raisons précisées plus loin, on choisit

$$F(K) = \sqrt{-\log(K/\bar{K})}, \quad (7.16)$$

où \bar{K} est un majorant strict des valeurs prises par K (de telle sorte que K/\bar{K} reste strictement inférieur à 1). Avec la convention $F(0) = +\infty$, cette application permet de construire une collection de longueurs pour toutes les arêtes du graphes. On obtient donc une métrique globale sur l'espace discret des individus, ou sur chaque composante connexe du graphe des contacts s'il n'est pas connexe.

Justification de $F(K) = \sqrt{-\log(K/\bar{K})}$

Le choix effectué résulte⁷ des 2 remarques suivantes :

1. En premier lieu, remarquons qu'une matrice de contacts présente une certaine analogie avec une matrice de conductances associée à un réseau électrique. En effet, le regroupement de plusieurs entités en un noeuds unique se fait selon les mêmes lois de sommation. Considérons un ensemble I d'entités, $K = (K_{ij})_{i,j \in I}$ une matrice de contacts associée à cette population. Soit $X \subset I$. On souhaite construire la matrice de contacts associée à la population $\bar{X} \cup (I \setminus X)$, où \bar{X} est maintenant considéré comme une entité unique, qui regroupe les entités individuelles dans X . Les nouveaux coefficients de la matrice réduite K' s'écrivent

$$\forall j \notin X, K'_{\bar{X},j} = \sum_{i \in X} K_{ij}. \quad (7.17)$$

7. Il s'agit plus de considérations informelles que d'une preuve de quoi que ce soit...

Le groupage fait apparaître un coefficient diagonal (qui encode les contact entre éléments de X) :

$$K'_{\overline{X}, \overline{X}} = \sum_{i,j \in K} K_{ij}.$$

La formule (7.17) correspond à la loi de sommation des conductances en parallèle : si les éléments de I sont considérés comme les points d'un réseau résistif, de matrice de conductances K , et que l'on rassemble tous les points de X en un unique nœuds (qui auront donc tous le même potentiel), alors la matrice de conductance du nouveau réseau est précisément la matrice définie ci-dessus.

2. Le deuxième ingrédient porte sur la correspondance entre conductances et distances. Nous considérons une matrice de contacts, selon le point de vue ci-dessus, comme une matrice de conductances, de telles sorte que la population est vue comme un réseau résistif. Il s'agit maintenant de transformer les conductances de chaque arête en longueur, en respectant le fait qu'un grande conductance (i.e. un grand nombre de contacts) correspond à une petite distance. On cherche donc une fonction strictement décroissante de \mathbb{R}^+ dans \mathbb{R}^+ . Un résultat théorique⁸ suggère un choix naturel pour cette correspondance. Il s'agit d'un résultat de convergence d'un opérateur de Laplacien discret associé à collection de points sur une variété, tirés aléatoirement selon une loi uniforme sur la variété, vers l'opérateur dit de Laplace-Beltrami associé à cette variété. Nous le présentons ici dans le cas simplifié d'un domaine de l'espace euclidien \mathbb{R}^d . Considérons donc un domaine $\Omega \subset \mathbb{R}^d$ régulier, et X_1, X_2, \dots des points de Ω tirés indépendamment selon la loi uniforme sur Ω . Pour $n \geq 1$, $h_n > 0$, on considère l'opérateur aléatoire discret Δ_{n,h_n} défini par, pour tout fonction f continue sur $\overline{\Omega}$,

$$x \in \Omega \mapsto \Delta_{n,h_n} f(x) = \frac{1}{nh_n^{1/d}} \sum_{i=1}^n \Phi\left(\frac{x - X_i}{h_n}\right) (f(x) - f(X_i))$$

où Φ est un noyau Gaussien

$$\Phi(p) = \frac{1}{(4\pi)^{d/2}} \exp(-|p|^2/4).$$

Si l'on restreint son action aux points (X_i) , cet opérateur peut être vu comme le Laplacien discret associé à un réseau résistif construit sur graphe complet, avec des conductances égales à

$$c_{ij} = \frac{1}{(4\pi)^{d/2}} \exp(|X_j - X_i|^2/4).$$

Il est montré que la quantité

$$\sup_{x \in \Omega} \sup_{f \in B} \left| \Delta_{n,h_n} f(x) - \frac{1}{|\Omega|} \Delta f(x) \right|$$

converge presque sûrement vers 0, où B désigne la boule unité de $C^3(\Omega)$, pour une certaine plage de comportements de h_n , plus précisément dès que

$$nh_n^{d+2} \log(1/h_n) \rightarrow +\infty, \quad nh_n^{d+4} \log(1/h_n) \rightarrow 0.$$

On a donc en particulier convergence pour

$$h_n = \frac{1}{n^{\frac{1}{d+2}-\varepsilon}}.$$

Ce résultat établit une correspondance canonique entre conductances et distances (euclidiennes en l'occurrence) basée sur une fonction F définie comme la réciproque de la Gaussienne. Si l'on assimile les nombres de contacts à des conductances, ramenée dans $[0, 1[$ par division par un majorant strict des valeurs, on obtient (à une constante multiplicative près), l'expression (7.16).

⁸ Evarist Giné, Vladimir Koltchinskii, Empirical graph Laplacian approximation of Laplace–Beltrami operators : Large sample results, IMS Lecture Notes–Monograph Series High Dimensional Probability Vol. 51 (2006) 238–259 , https://projecteuclid.org/download/pdf_1/euclid.lnms/1196284116

Remarque 7.8. Le lien exprimé ci-dessus entre conductances et distances est aussi utilisé plus informellement dans le contexte des *diffusion maps*⁹. L'approche est la suivante : on considère comme précédemment un ensemble fini X , et une matrice de “poids” K_{xy} associée, avec $K_{xy} = K_{yx}$, et $K_{xy} \geq 0$ pour tous $x, y \in X$. On peut associer à cette matrice un réseau résistif $\mathcal{N} = (V, E, r)$ construit sur l'ensemble de sommets $V = X$ (selon la définition 2.1, page 39), avec la convention que $(x, y) \in E \Leftrightarrow K_{xy} > 0$, et $r_{xy} = 1/K_{xy}$. On associe à ce réseau (la démarche est détaillée dans la section 6.2, page 128) une matrice stochastique Π définie par

$$\pi_{xy} = K_{xy}/K_x, \quad K_x = \sum_y K_{xy}.$$

Noter qu'avec la convention $K_{xy} = 0$ dès que $(x, y) \notin E$, on peut garder la convention de sommation sur l'ensemble des points $y \in X$. La matrice $\Pi^k = (\pi_{xy}^k)$ (attention, le premier k représente une puissance, alors que le second est un indice en position d'exposant) est une matrice de transition pour k étapes de la marche aléatoire initiale, i.e. π_{xy}^k est la probabilité d'atteindre y en k étapes, partant de la position initiale x . En d'autres termes, pour x fixé, $(\pi_{xy})_y$, que l'on notera aussi $\pi_{x,\cdot}$, est la loi de probabilité sur X décrivant la position de la particule issue de x , après k pas. On construit alors des *distances de diffusion* (une distance pour chaque valeur de k) de la façon suivante :

$$(x, x') \mapsto D_{xy}^k = |\pi_{x,\cdot}^k - \pi_{x',\cdot}^k| = \sqrt{\sum_y \frac{1}{K_x} |\pi_{xy}^k - \pi_{x'y}^k|^2}.$$

Retour sur le taux de reproduction

Le taux de reproduction R_0 associé à une épidémie est défini comme le nombre de personnes qu'un malade contamine en moyenne au cours de sa période contagieuse. Plus précisément, on parle du taux de reproduction *initial*, ou taux de reproduction *de base*, le nombre défini ci-dessus dans l'hypothèse où le malade évolue au sein d'une population entièrement (ou quasi-entièrement) susceptible, c'est à dire constituée d'individus qui n'ont pas été vaccinés, ni immunisés de quelque manière que ce soit.

Cette section aborde d'une manière critique certaines difficultés liées à la définition même de cette valeur, aux manières de l'estimer en pratique, à ses généralisations possibles...

Decomposition de R_0

Le nombre R_0 se définit a priori comme le produit de 3 facteurs :

$$R_0 = p \times K \times D,$$

où D est la durée (en jours) pendant laquelle un malade est contagieux, K le nombre de contacts qu'il a par jour avec d'autres personnes (au sein d'une population susceptible), et p la probabilité que la maladie se transmette lors d'un contact.

Remarque 7.9. La définition de ce qu'on appelle un “contact” est ici imprécise, mais cette ambiguïté n'est pas trop problématique, tant que l'on se situe dans un régime où la probabilité d'être infecté lors d'un contact est essentiellement proportionnelle à la durée de ce contact, ce qui est utilisé dans la formule définissant R_0 . Rappelons que cet propriété n'est pas générale : si la probabilité d'être infecté lors d'un contact est p , la probabilité d'être infecté lors de K contacts, est égale¹⁰ à

$$1 - (1 - p)^K,$$

9. Ronald R. Coifman, Stéphane Lafon, Diffusion maps, Appl. Comput. Harmon. Anal. 21 (2006) 5–30,
<https://cis.temple.edu/~latecki/Courses/RobotFall108/Papers/DiffusionMaps06.pdf>

10. Pour ne pas être infecté, il faut “gagner” (c'est à dire ne pas être infecté) lors de tous les contacts. On peut voir assimiler ça à un jeu de pile ou face biaisé, avec une probabilité p de perdre à chaque partie (i.e. chaque contact). Pour être infecté, il suffit de perdre une seule fois, et pour ne pas l'être il faut gagner toutes les parties.

qui n'est proche de pK que si p et pK sont petits devant 1. Considérons par exemple que le nombre de contacts soit estimé en minutes, de telle sorte qu'un contact de 8 minutes sera par exemple compté comme 8 contacts. La valeur p représente alors la probabilité que la maladie se transmette lors d'un contact d'une minute. Si l'on cherche à estimer la probabilité que la maladie se transmette en un contact de 8 minutes, on a $p_{10} = 1 - (1 - p)^8 \approx 8p$ tant que $8p$ est petit devant 1. Un changement d'unité sur la notion de contact se traduit donc essentiellement par le même changement sur p , de telle sorte qu'il n'y a pas lieu de préciser la durée d'un contact élémentaire, tant que celle-ci est telle que la probabilité d'être infecté durant un tel contact reste petite devant 1, et que le nombre d'infectés par unité de temps (ici des jours) est petit devant 1.

Le terme de *taux de reproduction effectif*, noté R_{eff} est en général utilisé pour désigner ce que l'on estime à partir des chiffres des nouvelles infections jour après jour (voir plus loin). En termes de modélisation, il est naturel de le définir comme une correction de R_0 par la fraction de susceptibles dans la population des personnes rencontrées par le malade générique considéré pour cette définition théorique, i.e.

$$R_{\text{eff}} = SR_0 = Sp \times K \times D.$$

Rôle de R_0 dans le modèle SIR

Comme précisé dans la section 7.2, le taux de reproduction de base R_0 intervient dans la condition de stabilité $R_0 < 1$ au moment de l'émergence possible d'une épidémie. Si $R_0 > 1$, on a dans les premiers temps croissance exponentielle du nombre d'infecté avec un taux $\gamma(R_0 - 1)$ (qui est bien homogène à l'inverse d'un temps). Dans le cas d'une épidémie pleinement développée, dont on suppose tous les paramètres constants, on aura plateau puis infléchissement du nombre d'infectés lorsque le nombre de S devient inférieur à $1/R_0$, de telle sorte que le taux de reproduction effectif $R_{\text{eff}} = R_0 S$ devienne inférieur à 1.

Estimation du taux de reproduction effectif

À la lumière de ce qui précède, on pourrait être tenté d'évaluer le taux de reproduction effectif en estimant séparément chacun des termes du produit $Sp \times K \times D$ (proportion de susceptibles dans la population, probabilité de contamination lors d'un contact, nombre de contacts par jour, durée de la période contagieuse). Les stratégies d'estimations de ce taux de reproduction ne sont pas basées sur cette approche décomposée, mais sur un retour à la définition initiale, à savoir le nombre de personnes que chaque malade infecte à son tour. Considérons dans un premier temps, pour simplifier, une situation extrême : on suppose que la maladie considérée est telle qu'une personne infectée n'est contagieuse qu'une seule journée, par exemple le 7ème jour qui suit le jour où la maladie s'est déclarée. Supposons que l'on soit capable d'estimer, au sein d'une population, le flux journalier de nouveau infectés. On note N_j ce nombre. Au vu des hypothèses formulées, toutes les personnes infectées aujourd'hui l'ont été par des personnes infectées il y a exactement 7 jours, qui étaient au nombre de N_{j-7} . Le nombre moyen de malades dus à chacun d'elles est donc N_j/N_{j-7} , qui est donc une estimation du R_{eff} .

Avant de considérer des hypothèses plus réalistes, notons que cette approche présente nativement une certaine robustesse. Ainsi, dans l'hypothèse où une fraction (inconnue) de personnes ne développent pas de symptômes ni ne sont contagieuses, il sera impossible de déterminer à partie de la simple observation des infectés symptomatiques quelle est par exemple la fraction des personnes immunisées (partiellement ou totalement), mais l'estimation du R_{eff} selon la méthode décrite ci-dessus reste envisageable, du fait que la fraction inconnue d'infectés symptomatiques et contagieux se trouve au numérateur et au dénominateur.

L'approche utilisée en 2020 pour estimer le R_{eff} du Corona virus prend en compte le fait que la contagion par un malade donné peut se faire pendant une certaine période qui suit son infection. On précise cette approche en introduisant des coefficients d'infectivité w_1, w_2, \dots , (*infectivity functions* en anglais) qui quantifient la contagiosité du patient à $J + 1, J + 2, \dots$, respectivement, de telle sorte qu'un patient infecte en moyenne $w_1 R_{\text{eff}}$ à $J + 1, w_2 R_{\text{eff}}$ à $J + 2, \dots$, pour un total de R_{eff} dès

que l'on suppose que la somme des w_j est égale à 1. Si l'on inverse le point de vue, en partant des infectés et non pas des infectants, une part des N_J nouveaux infectés au jour j a été infectée par des personnes elles-mêmes infectées depuis 1 jour, soit une contribution de $R_{\text{eff}}w_1N_{j-1}$, une autre part par des personnes elles-mêmes infectées deux jours avant, soit $R_{\text{eff}}w_2N_{j-2}$, etc ... On arrive finalement à l'expression

$$N_J = R_{\text{eff}} \sum_{i=1}^N w_i N_{J-i},$$

où N est la durée de la maladie, plus précisément le nombre de jours au delà duquel un malade n'est plus contagieux. On en déduit la formule

$$R_{\text{eff}} = \frac{N_J}{\sum_{i=1}^N w_i N_{J-i}},$$

qui peut être utilisée comme estimateur dès que l'on se donne la collection de poids (w_i).

Noter que ce modèle est une stricte généralisation du précédent (prendre $w_j = \delta_{j,7}$). Par ailleurs, on peut vérifier que le modèle précédent dépasse le cadre de ses propres hypothèses (qui semblent assez irréalistes), sous certaines conditions. Considérons par exemple qu'un patient est contagieux de façon constante pendant les 13 jours qui suivent son infection, i.e.

$$w_1 = w_2 = \dots = w_{13} = \frac{1}{13},$$

alors si N_t varie de façon affine¹¹ sur la période $[j-13, j]$, alors la moyenne des valeurs sur cet intervalle s'identifie à la valeur en $j-7$, de telle sorte que l'on retrouve la première approche.

¹¹. Cela signifie que N_t varie de façon quadratique sur le même intervalle. Noter que cette hypothèse est au moins approximativement vérifiée dans une très grande généralité : si I_t varie de façon "lisse", il est approchable avec une bonne précision par un polynôme de degré 2 en t .

7.6 Exercices

Exercice 7.5. On considère le système différentiel en (S, I) qui correspondant aux deux premières équations de (7.2), page 149. Montrer que, pour toute condition initiale dans $[0, 1] \times [0, 1]$, ce système admet une solution unique globale à valeurs dans $[0, 1] \times [0, 1]$.

Exercice 7.6. (Modèle SEIR)

1) Proposer un modèle dans l'esprit du modèle SIR (de (7.2), page 149), qui prenne en compte un compartiment E (exposed), correspondant à des personnes contaminées, mais pas encore contagieuse.

2) Préciser les point d'équilibre de ce système, et étudier leur stabilité.

Exercice 7.7. ()

On considère le modèle discret (7.10), page 157.

1) Montrer que, si l'on a $S_x^0 + I_x^0 + R_x^0 = 1$, avec les 3 quantités positives ou nulles, alors on a, pour tout n , $S_x^n + I_x^n + R_x^n = 1$.

2) On part d'une condition initiale $I^0 = \delta_x$. Décrire la suite d'indices

$$\text{supp}(I^n) = \{x, I_x^n > 0\}$$

en fonction des propriétés du graphe comme espace métrique (pour la métrique combinatoire).

3) On choisit (sans se préoccuper du réalisme) de prendre $\gamma = 0$ (une personne malade le reste indéfiniment). Dans les conditions de la questions précédentes, décrire le comportement de la suite I^n .

Exercice 7.8. (Développements limités)

Formuler et démontrer une propriété permettant de formaliser rigoureusement la démarche permettant d'aboutir au système simplifié (7.8), à l'aide de développement limités.

Exercice 7.9. (Modèle homogène)

Montrer rigoureusement que, sous des hypothèses adaptées d'homogénéité de K et d'uniformité des champs initiaux, le système (7.10) peut se réduire en le système SIR homogène (7.2), page 149.

Chapitre 8

Modèles de trafic routier ou piéton

Sommaire

8.1	Le modèle FTL	170
8.1.1	Points d'équilibres, stabilité, propagation des perturbations	172
8.1.2	Cas périodique	176
8.1.3	Extensions, développements	180
8.1.4	Exercices	183
8.2	Modèles d'ordre 2	184
8.2.1	Stabilité	185
8.2.2	Extensions, développements	189
8.3	Modèles granulaires de foules	191
8.3.1	Modèle monodimensionnel	191
8.3.2	Modèle en dimension 2 (disques rigides)	193
8.4	Modèles macroscopiques de trafic routier	197
8.4.1	Modèle d'évolution	197
8.4.2	Solutions faibles	198
8.5	Modèles granulaires de foules	200
8.5.1	Modèle monodimensionnel	200
8.5.2	Modèle en dimension 2 (disques rigides)	202

8.1 Le modèle FTL

Modèle 8.1. (Modèle *Follow the Leader*)

Le modèle dit *Follow the Leader*¹ 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

$$x_1(t) < x_2(t) < \cdots < x_{n+1}(t). \quad (8.1)$$

1. 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,
Cette dénomination est cependant partiellement impropre dans le cas qui nous intéresse : chaque entité suit de fait l'entité qui la précède, mais la présence de cette dernière est plus une gêne (qui conduit à une diminution de la vitesse) qu'une incitation positive, comme le suggèreraient la dénomination choisie.

La vitesse du véhicule i est supposée ne dépendre que de la distance au véhicule précédent, c'est-à-dire $x_{i+1} - x_i$. Le système s'écrit alors

$$\dot{x}_j = \varphi(x_{j+1} - x_j) \quad 1 \leq j \leq n. \quad (8.2)$$

Ce modèle peut se décliner en une version *séquentielle*, on se donne alors la trajectoire du véhicule de tête, ou *périodique*, en considérant que l'on est sur une route circulaire².

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

$$w \mapsto \varphi(w) = U \left(1 - \exp \left(-\frac{w - w_m}{w_s} \right) \right)_+, \quad (8.3)$$

où w_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 autoroute), et w_m la taille des véhicules, de telle sorte que $w = w_m$ correspond à la situation pare-choc contre pare-choc. Cette quantité conditionne la raideur (*stiffness* en anglais) du modèle.

Remarque 8.2. On peut représenter le graphe de dépendance du modèle de la façon suivante : si l'on note $V = \{1, 2, \dots, n\}$, on peut définir un ensemble A d'arêtes :

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

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*.

Proposition 8.3. On se donne des positions initiales vérifiant la relation d'ordre (8.1). On suppose que la trajectoire $t \mapsto x_{n+1}(t) = X(t)$ de l'entité de tête est une fonction continue du temps, croissante. On se donne une fonction de comportement φ Lipschitzienne nulle en 0 (prolongée par 0 en deçà), et prenant ses valeurs dans l'intervalle $[0, U]$. Le système (8.2) admet une unique solution maximale, qui est globale.

Démonstration. L'application ainsi construite est Lipschitzienne. On peut appliquer le théorème de Cauchy-Lipschitz 11.10 sur $[0, +\infty[\times \mathbb{R}^n$, ce qui assure l'existence et l'unicité d'une solution maximale. Cette solution est globale car la vitesse est bornée (donc a fortiori sous-linéaire à l'infini) d'après la proposition 11.15. \square

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 8.4. On se place dans les hypothèses de la proposition précédente, avec des conditions initiales telles que les distances sont strictement positives. Les distances restent alors strictement positives pour tout temps.

Démonstration. On note $T^* > 0$ le plus temps en lequel $X - x_n$ est nul. Par continuité $X - x_n$ tend vers 0 quand t tend vers T^* . La distance $z - x_n$, tend donc aussi vers 0, avec $z = X(T^*)$. Or on a

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

avec $L = \|\varphi'\|_\infty$, d'où

$$\frac{d}{dt} (z - x_n) \geq -L(z - x_n)$$

et ainsi $z - x_n \geq C e^{-Lt}$. On procède de même avec $w_{n-1} = x_n - x_{n-1}$, puis w_{n-2} , etc ...

\square

2. On se reportera à <https://www.youtube.com/watch?v=RYgQJdm7f0E> pour la description d'une telle situation.

Remarque 8.5. Le caractère lipschitzien de φ est essentiel pour éviter les accidents. Prenons par exemple une fonction φ qui se comporte comme w^α au voisinage de 0, avec $\alpha \in]0, 1[$. On considère que le véhicule de tête est arrêté en $a \in \mathbb{R}$. L'équation s'écrit

$$\dot{x} = (a - x)^\alpha, \quad x(0) < a,$$

ce qui conduit à

$$x(t) = a - ((a - x(0))^{1-\alpha} - (1-\alpha)t)^{1/(1-\alpha)}.$$

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

8.1.1 Points d'équilibres, stabilité, propagation des perturbations

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

$$w_{eq} = -w_s \ln \left(1 - \frac{V_{eq}}{U} \right),$$

ils vont tous à la vitesse V_{eq} 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_{eq} .

On introduit les variables de distances entre véhicules :

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

Le système s'écrit, pour ces nouvelles variables

$$\dot{w}_i = \varphi(w_{i+1}) - \varphi(w_i), \quad i = 1, \dots, n, \quad \text{ou } \dot{w} = F(u).$$

et $W_{eq} = (w_{eq}, \dots, w_{eq})$ est point d'équilibre du système.

Proposition 8.6. Le point d'équilibre défini ci-dessus est asymptotiquement stable.

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

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

On a donc une unique valeur propre $-\varphi'(w_{eq}) < 0$, donc stabilité asymptotique avec un temps caractéristique de retour à l'équilibre³ égal à $1/\varphi'(w_{eq})$. \square

Remarque 8.7. On notera (cette remarque dépasse largement le cas de ce modèle particulier) le lien entre le “support” de la matrice du gradient (ensemble des positions des éléments non nuls), et la matrice d'adjacence M du graphe d'influence défini dans la remarque 8.2. Plus précisément, si l'on rajoute explicitement dans la définition du graphe qu'un sommet pointe sur lui-même (la vitesse d'un individu dépend aussi de sa propre position), et avec le choix fait de créer l'arête (i, j) lorsque j influence i , le support de ∇F est exactement le support de M^T . Le fait que le graphe soit acyclique (en dehors des boucles) est exprimé par le caractère triangulaire supérieur de la matrice du gradient

3. 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 $1/\varphi'(w_{eq})$, 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 $1/\varphi'(w_{eq})$ n'en dépend pas.

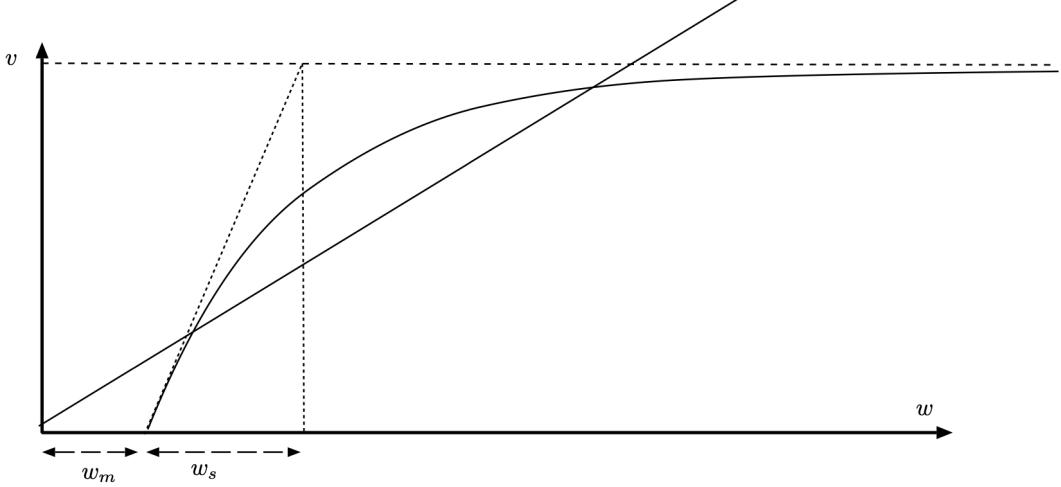


FIGURE 8.1 – Vitesse fonction de la distance

(sans qu'il soit même nécessaire, ici, d'effectuer une renumérotation). On notera en particulier que, dans un tel cas (graphe acyclique), toutes les valeurs propres sont réelles. Par ailleurs, si les éléments diagonaux sont identiques, la matrice n'est pas diagonalisable, sauf dans le cas trivial de n équations indépendantes. De façon plus générale, dès que certaines valeurs propres sont dégénérées, on aura un bloc de Jordan non réductible. Comme on le verra, en termes de système dynamique, cette situation correspond à une propagation de l'information à partir du véritable mode propre vers les modes dégradés du sous-espace stable.

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 w_{eq} , qui avancent à la vitesse $v_e = \varphi(w_{eq})$. En se plaçant dans le référentiel qui suit le train, à la vitesse w_{eq} , on peut décrire les petites évolutions du modèle en considérant que les distances sont du type $w_{eq} + h_i$, où h_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

$$\dot{w}_j = \dot{h}_j = \varphi(w_{eq} + h_{j+1}) - \varphi(w_{eq} + h_j) \approx \varphi'(w_{eq})(h_{j+1} - h_j) = w_{eq}\varphi'(w_{eq})\frac{h_{j+1} - h_j}{w_{eq}}.$$

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

$$\frac{\partial h}{\partial t} - w_{eq}\varphi'(w_{eq})\frac{\partial h}{\partial x} = 0.$$

Il s'agit d'une équation de transport à la célérité $c = -w_{eq}\varphi'(w_{eq})$. 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 $\varphi(w_{eq})$. On aura effectivement propagation vers l'arrière⁴ (pour l'observateur extérieur) si

$$w_{eq}\varphi'(w_{eq}) > \varphi(w_{eq}) \iff \varphi'(w_{eq}) > \frac{\varphi(w_{eq})}{w_{eq}}.$$

4. 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.

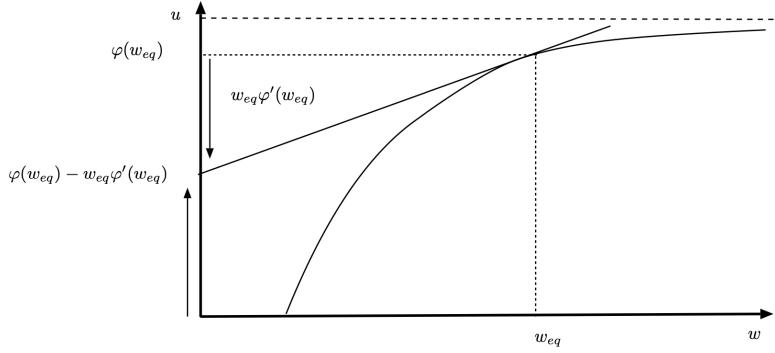


FIGURE 8.2 – Vitesse de propagation des perturbation

Dans le cas où l'on a négligé la taille des entités ($w_m = 0$), la fonction φ est nulle en 0. Si on la suppose concave (par exemple φ donné par (8.3)), toute corde intersecte la courbe en un point unique, et la pente de la courbe est inférieure à la pente de la corde, i.e. $\varphi'(w_{eq}) < \frac{\varphi(w_{eq})}{w_{eq}}$. Dans ce cas l'information ne va pas suffisamment vite pour remonter le courant. Si la taille des entités est prise en compte en revanche (voir figure 8.1, avec $w_m > 0$), on a deux régimes possibles pour une même pente de corde, i.e. pour un même flux (le flux d'entités par unité de temps est $\varphi(w_{eq})/w_{eq}$). Le premier est 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.

Noter également que la vitesse apparente (selon le point de vue *eulérien*, i.e. dans le référentiel fixe) de propagation des perturbations peut être représentée graphiquement (voir figure 8.2) : elle correspond à l'intersection entre la tangente à la courbe au point d'équilibre w_{eq} avec l'axe vertical des vitesses. La figure représente une situation où cette vitesse est positive, mais elle peut être négative (point d'intersection sous l'axe des x) pour des valeurs plus petites de w_{eq} , lorsque la courbe de comportement, concave sur son support, présente un plateau à 0 pour des distances petites.

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 :

$$\dot{w} = Mw,$$

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

On garde la notation w pour désigner le vecteur inconnu, mais les w_i correspondent maintenant à des variations autour du point d'équilibre, qui évoluent au voisinage de 0 (et non pas de w_{eq}).

La solution du problème ci-dessus s'écrit

$$w(t) = e^{tM}w_0,$$

où w_0 est une perturbation initiale. La matrice M s'écrit

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

avec $\beta = \varphi'(w_{eq})$, et N une matrice nilpotente

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

L'exponentielle s'écrit donc

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

Montrons que la forme particulière de cette matrice rend compte d'une propagation des perturbations vers les index des véhicules 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

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

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

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

On reconnaît, pour tout instant t , une loi de Poisson de paramètre βt . Dans les premiers instants, cette fonction va avoir un maximum glissant qui correspond au véhicule affecté par la perturbation au temps considéré.

On peut par exemple rechercher à quel moment la perturbation ressentie par l'entité $n - k$ est maximale. On a

$$p_{n-k}(t) = e^{-\beta t} \frac{(\beta t)^k}{k!}, \quad p'_{n-k}(t) = e^{-\beta k} \frac{\beta^k t^{k-1}}{k!} (-\beta t + k)$$

qui s'annule pour $t = k/\beta$. L'information se propage donc vers l'arrière sur le train de véhicules à la vitesse de $1/\beta = \varphi'(w_{eq})$ véhicules par seconde. Si l'on prend en compte la distance entre véhicules, qui est w_{eq} , on retrouve une célérité au sens usuel de $w_{eq}\varphi'(w_{eq})$ (en ms^{-1}).

Exercice 8.1. Donner un équivalent du (maximum de) l'intensité de la perturbation ressentie par l'entité $n - k$ pour k grand.

Exercice 8.2. 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 w_s , et concave sur $[w_s, +\infty]$) 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 8.8. 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 :

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

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^∞ . 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.

Stabilité non linéaire

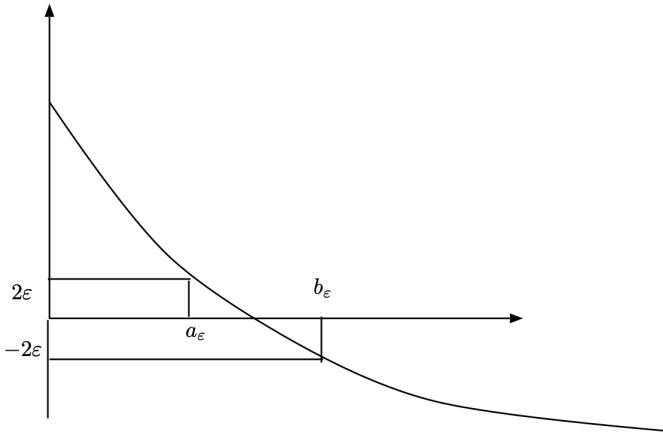


FIGURE 8.3 – Stabilité non linéaire

Reprendons la situation d'un véhicule de tête avançant à la vitesse $V_{eq} = \varphi(w_{eq})$. L'état d'équilibre (en distances) correspond à des véhicules équi-répartis espacés de w_{eq} . Le fait que les valeurs propres du gradient soient strictement négatives assure la stabilité asymptotique de cet équilibre, c'est à dire que, pour une perturbation suffisamment petite de l'état d'équilibre, on a retour exponentiel à l'équilibre. Mais cette stabilité est locale, et l'on peut se demander si, partant d'un état initial quelconque, on aura convergence vers l'équilibre. C'est effectivement le cas, comme le précise la proposition suivante.

Proposition 8.9. On considère le modèle (8.2), où φ est une fonction C^1 qui croît strictement de 0 à la valeur limite $U > 0$. On suppose que la vitesse de l'entité $n + 1$ est constante égale à $V_{eq} = \varphi(w_{eq})$, avec $w_{eq} > 0$. La solution globale converge alors vers l'état d'équilibre, au sens où toutes les distances u_i convergent vers w_{eq} .

Démonstration. On peut en fait montrer une propriété un peu plus générale, qui nous permettra de montrer la propriété annoncée par récurrence. On suppose que le véhicule de tête avance à la vitesse $V(t)$ qui converge vers V_{eq} quand t tends vers $+\infty$. On a

$$\dot{w}_n = V(t) - \varphi(w_n) = V_{eq} - \varphi(w_n) + \varepsilon(t) = f(u_n, t),$$

avec $\varepsilon(t) \rightarrow 0$ quand $t \rightarrow +\infty$. Pour tout $\varepsilon > 0$, il existe un temps T_ε tel que $|\varepsilon(t)| < \varepsilon$ pour tout temps $t > T_\varepsilon$. Au delà de T_ε , $f(u, t)$ est inférieur ou égal à $V_{eq} - \varphi(b_\varepsilon) + 2\varepsilon < 0$ pour tout $u \geq b_\varepsilon$ (voir figure 8.3). De la même manière, on a une vitesse positive minorée à gauche de a_ε . La trajectoire est donc nécessairement dans l'intervalle $[a_\varepsilon, b_\varepsilon]$ pour un temps assez grand. Quand ε tend vers 0, a_ε et b_ε tendent donc vers u_e du fait de la stricte croissance de φ , et l'on vient de démontrer que la trajectoire était, au delà d'un certain temps, dans l'intervalle $[a_\varepsilon, b_\varepsilon]$. On a donc montré que u_n tendait vers w_{eq} quand t tend vers $+\infty$. On en déduit que \dot{x}_n tend vers V_{eq} , on l'on peut démontrer exactement de la même manière que u_{n-1} , puis u_{n-2} , etc ..., tendent vers w_{eq} . \square

On notera que l'on a utilisé de façon essentielle la stricte croissance de φ . Il est évident que cet effet attractif du point d'équilibre sera très faible dans le cas de grandes distances (qui correspondent à une zone où φ est presque constante), voire inexistant si l'on considère (ce qui est pertinent en termes de modélisation, que φ est constante au delà d'une certaine distance).

Exercice 8.3. Que se passe-t-il si le véhicule de tête se déplace à la vitesse maximale U ?

8.1.2 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

$$\dot{x}_j = \varphi(x_{j+1} - x_j), \quad j = 1, \dots, n \quad (n+1 \equiv 1),$$

ou, exprimé sur les variables de distance $w_j = x_{j+1} - x_j$ (avec la convention $w_n = x_1 - x_n$)

$$\dot{w}_j = \varphi(w_{j+1}) - \varphi(w_j), \quad j = 1, \dots, n \quad (n+1 \equiv 1), \quad (8.4)$$

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

Remarque 8.10. Comme dans le cas linéaire, on peut définir un graphe orienté (V, A) , 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 cycle⁵.

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

Proposition 8.11. On suppose que φ est une fonction C^1 strictement croissante sur $[0, +\infty[$. Le point d'équilibre $y_{eq} = (w_{eq}, \dots, w_{eq})$, $w_{eq} = L/n$, solution stationnaire de (8.4) est alors asymptotiquement stable.

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

$$\nabla F(y_{eq}) = \varphi'(w_{eq}) \begin{pmatrix} -1 & 1 & 0 & \cdot & 0 \\ 0 & -1 & 1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & -1 & 1 \\ 1 & \cdot & \cdot & 0 & -1 \end{pmatrix} = \varphi'(w_{eq}) A_{per} = \varphi'(w_{eq}) (-\text{Id} + C).$$

où C est une matrice circulante, matrice de permutation particulière qui réalise le shift à droite périodique. Cette dernière vérifie $C^n = \text{Id}$ et la famille $(C^k)_{0 \leq k \leq n-1}$ 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

$$\mu_k = -1 + \exp\left(\frac{2ik\pi}{n}\right), \quad k = 0, \dots, n-1. \quad (8.5)$$

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 partie réelle strictement négative⁶. \square

5. Ce cycle 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).

6. 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

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

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

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

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

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

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

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

Corollaire 8.12. 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 8.13. On considère n entités avançant sur un chemin circulaire et fermé, on suppose l'évolution régie par

$$\dot{x}_j = \varphi(x_{j+1} - x_j), \quad j = 1, \dots, n \quad (n+1 \equiv 1),$$

où φ est une fonction croissante. On note $w_j = x_{j+1} - x_j$, et l'on considère une solution du système (8.4). Pour toute fonction g continûment différentiable et convexe, la quantité

$$S(w(t)) = \sum_j g(w_j)$$

est décroissante.

Si l'on suppose φ strictement croissante et g strictement convexe, cette décroissance est stricte tant que l'on n'a pas $w_j = L/n$ pour tout j .

Démonstration. Les distances vérifient

$$\dot{w}_j = \varphi(w_{j+1}) - \varphi(w_j), \quad j = 1, \dots, N.$$

On a donc

$$\begin{aligned} \frac{d}{dt} \left(\sum_j g(w_j) \right) &= \sum_j g'(w_j) \dot{w}_j = \sum_j g'(w_j) (\varphi(w_{j+1}) - \varphi(w_j)) \\ &= \sum_j \varphi(w_j) (g'(w_{j-1}) - g'(w_j)). \end{aligned}$$

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

$$\sum_j \varphi \circ (g')^{-1}(y_j) (y_{j-1} - y_j),$$

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

$$\psi(y_j) + \psi'(y_j)(y_{j-1} - y_j) \leq \psi(y_{j-1}),$$

de telle sorte que

$$\frac{d}{dt} \left(\sum_j g(u_j) \right) \leq \sum_j (\psi(y_{j-1}) - \psi(y_j)) = 0.$$

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

Dans le cas où φ est strictement croissante et g strictement convexe, au moins l'une des inégalités ci-dessus est stricte, sauf dans le cas où toutes les distances sont les mêmes. \square

Remarque 8.14. 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 1.10, page 25)

$$S(u) = \sum_j u_j \log u_j.$$

Remarque 8.15. 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 $u_1 = u_2 = \dots = u_N = L/N$. On converge alors nécessairement 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 8.16. 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

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

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 $-w_{eq}\varphi'(w_{eq})$. 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

$$\frac{du}{dt} = \varphi'(w_{eq})(-\text{Id} + C)u.$$

La matrice est diagonalisable, d'éléments propres

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

Les parties réelles des valeurs propres,

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

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

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

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 (grandes longueurs d'onde, les plus lentes à relaxer vers 0) une célérité de l'ordre de $-\varphi'(w_{eq})$ (en s^{-1} , ou entités par seconde), ou, si l'on prend en compte le fait que les entités sont séparées de w_{eq} , d'une vitesse effective de $-w_{eq}\varphi'(w_{eq})$ (en ms^{-1}). On notera la mise en évidence d'un phénomène de *dispersion* : la vitesse de propagation des ondes dépend de leur fréquence. Par exemple pour la haute fréquence qui jouera un rôle clé pour le modèle d'ordre 2 en temps, qui correspond à $k = n/6$, on a une vitesse de propagation légèrement inférieure (facteur $3/\pi \sin(\pi/3)$).

8.1.3 Extensions, développements

Aspects énergétiques

On s'intéresse ici, dans le cadre du modèle précédent, à l'estimation de la consommation d'essence et au bilan carbone des véhicules virtuels impliqués. La consommation en essence d'une voiture se déplaçant sur une route est liée essentiellement à trois phénomènes, dont l'importance relative peut varier selon le contexte :

1. Augmentation de l'énergie cinétique du véhicule dans les phases d'accélération.
2. Puissance des forces exercées par l'air environnant, qui s'opposent au mouvement.
3. Variations d'altitudes au cours du parcours (augmentation de l'énergie potentielle du véhicule dans les montées).

Nous ne considérerons ici que les deux premiers points, le troisième nécessitant une connaissance précise de l'altitude le long du parcours.

1. *Accélération*. Avec les notations utilisées précédemment, l'énergie cinétique du véhicule j , de masse m , est $1/2m\dot{x}_j^2$. La puissance nécessaire pour faire varier cette énergie s'écrit (on ne prend en compte que la puissance associée à une accélération du véhicule) :

$$\left[\frac{d}{dt} \left(\frac{1}{2}m\dot{x}_j^2 \right) \right]_+ = \frac{m}{2} [\dot{x}_j \ddot{x}_j]_+.$$

L'accélération s'exprime

$$\ddot{x}_j = \frac{d}{dt} \varphi(x_{j+1} - x_j) = (\dot{x}_{j+1} - \dot{x}_j) \varphi'(x_{j+1} - x_j).$$

Si l'on suppose φ croissante, et du fait que les véhicules vont vers les abscisses croissantes (φ est positive), on a finalement une puissance totale du train de véhicules (cas périodique avec $n+1 \equiv 1$) égale à

$$\mathcal{P}_{acc} = \sum_{j=1}^n [\dot{x}_{j+1} - \dot{x}_j]_+ \varphi'(x_{j+1} - x_j) \varphi(x_{j+1} - x_j).$$

Si l'on s'intéresse à une perturbation $x+h$ du régime stationnaire périodique $x_{j+1} - x_j \equiv w_{eq}$, vitesses égales à $\varphi(w_{eq})$, on a

$$\mathcal{P}_{acc} = \frac{m}{2} \sum_{j=1}^n [\dot{h}_{j+1} - \dot{h}_j]_+ \varphi'(w_{eq}) \varphi(w_{eq}).$$

2. *Résistance de l'air*. La force exercée par l'air environnant sur un véhicule avançant à vitesse V est estimée par

$$F = \frac{1}{2} \rho S C_x V^2,$$

où ρ est la densité de l'air ($\approx 1.2 \text{ kg m}^{-3}$), S la surface efficace du véhicule (autour de 2m^2 pour une citadine standard), et C_x un nombre sans dimension appelé *coefficient de trainée* qui dépend

des propriétés aéro-dynamiques du véhicule (de l'ordre de 0.5 pour les véhicules courants). Pour une vitesse de 100 km h⁻¹, on trouve une force de l'ordre de 500 N.

On obtient la puissance en multipliant par la vitesse (de l'ordre de 30 ms⁻¹) : $\mathcal{P} \approx 15$ kW. Le véhicule parcourt 100 km en une heure, il a donc besoin d'une énergie de 15 kWh aux 100 km. Un litre d'essence contenant approximativement une énergie de 10 kWh, dont un moteur peut récupérer environ le tiers en énergie mécanique, on retrouve l'ordre de grandeur attendu : 5 L aux 100 km pour compenser la résistance de l'air.

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 $v_{eq} = \varphi_{n+1}(w_{eq})$, et que l'on se donne des courbes de comportement φ_i toutes strictement croissantes (pour $w \geq w_m$), 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.

$$v_e < \max_w \varphi_i(w) \quad \forall i.$$

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

$$\nabla F = \begin{pmatrix} -\beta_1 & \beta_2 & 0 & \cdot & 0 \\ 0 & -\beta_2 & \beta_3 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & -\beta_{n-1} & \beta_n \\ 0 & \cdot & \cdot & 0 & -\beta_n \end{pmatrix}, \quad \beta_i = \varphi'_i(w_{eq}^i) \quad i = 1, \dots, n. \quad (8.6)$$

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 génériquement⁷ différents deux à deux. Mais alors la matrice est diagonalisable, et l'étude du comportement de la solution du système linéarisé $e^{tA} w_{pert}$, 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

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

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 extraite près) vers un vecteur propre de la matrice $A = -\text{Id} + N$, 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

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

7. 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.

Si l'on considère maintenant la solution du problème d'évolution linéaire, avec une perturbation sur la 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 conditionnées⁸ : 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 (8.6), pour des β_i proches les uns des autres.

On peut se convaincre que le modèle possède une certaine stabilité structurelle au voisinage du point considéré (toutes les entités ont le même comportement) sans utiliser le caractère génériquement diagonalisable de la matrice du gradient. On considère pour cela le système linéaire

$$\frac{dw}{dt} = (A + \varepsilon B) w,$$

avec

$$A + \varepsilon B = \begin{pmatrix} -1 & 1 & 0 & \cdot & 0 \\ 0 & -1 & 1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & -1 & 1 \\ 0 & \cdot & \cdot & 0 & -1 \end{pmatrix} + \varepsilon \begin{pmatrix} -b_1 & b_2 & 0 & \cdot & 0 \\ 0 & -b_2 & b_3 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & -b_{n-1} & b_n \\ 0 & \cdot & \cdot & 0 & -b_n \end{pmatrix}$$

où B est la matrice qui représentent les écarts à l'uniformité en termes de comportement des conducteurs. Les b_i n'ayant pas de raison d'être identiques, la matrice $A + \varepsilon B$ est en général diagonalisable (les valeurs propres sont distinctes), mais de façon très instable comme le suggère le développement précédent. Notons w^0 la solution du problème de référence $\dot{w}^0 = Aw^0$. Cherchons alors la solution du système perturbé sous la forme $\dot{w}_\varepsilon = w^0 + \varepsilon w^1$. On obtient alors l'équation sur w^1 :

$$\dot{w}^1 = Aw^1 + Bw^0 + \varepsilon Bw^1,$$

qui converge bien vers une solution bornée lorsque ε tend vers 0, ce qui assure que le comportement effectif du système correspond bien à celui associé à la matrice de Jordan, avec un terme correctif

$$w^1(t) = \int_0^t e^{-(t-s)A} Bw^0(s) ds,$$

qui traduit les effets dus aux disparités entre conducteurs (disparités supposées légères).

8. 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.

8.1.4 Exercices

Exercice 8.4. On s'intéresse ici à un modèle de trafic d'ordre 2 en temps

$$\ddot{x}_j = \frac{1}{\tau}(\varphi(x_{j+1} - x_j) - \dot{x}_j), \quad (8.7)$$

où τ est un temps caractéristique, et φ une fonction globalement lipschitzienne de \mathbb{R} dans \mathbb{R} , nulle sur \mathbb{R}_- , C^1 sur $]0, +\infty[$, et à valeurs dans $[0, U]$, avec $U > 0$.

- 1) Écrire un résultat d'existence et d'unicité pour ce problème dans le cas périodique, et expliquer pourquoi certaines de ces solutions pourraient ne pas être réalistes.
- 2) On se place dans une configuration périodique : N véhicules sur une route de longueur L . Écrire l'équation vérifiée par les distances $w_j = x_{j+1} - x_j$, et mettre ce système sous la forme d'une équation différentielle ordinaire dans \mathbb{R}^{2N} (on pourra noter $u_j = \dot{w}_j$), du type $\dot{y} = f(y)$.
- 3) On suppose φ strictement croissante sur \mathbb{R}_+ . Montrer que le système en $y = (w, u)$ admet un point fixe unique du type $y_{eq} = (w_{eq}, \dots, w_{eq}, 0, \dots, 0)$.
- 4) Écrire le gradient de f sous la forme d'une matrice M par blocs 2×2 , en fonction de la matrice $A_{per} = -\text{Id} + C \in \mathcal{M}_n(\mathbb{R})$, où C est la matrice de shift à gauche avec périodicité, et de $\eta = \varphi'(w_{eq})^{-1}$.
- 5) Préciser le spectre $(\mu_k)_{1, \leq k \leq n}$ de A_{per} .
- 6) Montrer qu'à toute valeur propre μ_k sont associées deux valeurs propres λ_k^\pm de M .
- 7)a) Montrer que le caractère instable du système est conditionné au fait que, dans le plan complexe, la racine carré du cercle centré en $1 - \alpha$ et de rayon α , noté C_α , contient des points de partie réelle strictement supérieure à 1, avec $\alpha = 4\tau/\eta$.
- b) Écrire l'équation sur (x, y) exprimant que le carré de $x + iy$ est dans C_α , et en déduire une équation sur $(X, Y) = (x^2, y^2)$ qui en résulte.
- c) Montrer que X peut s'exprimer en fonction de $Y \geq 0$, au voisinage de $(X, Y) = (1, 0)$, et que $\partial X / \partial Y > 0$ dès que $\alpha > 2$.
- d) (••) Estimer le mode asymptotiquement le plus instable lorsque α tend vers $+\infty$.
- 8) Conclure quand à l'émergence d'instabilités en fonction des valeurs de τ et η , et commenter la signification en termes de modélisation de cette condition.
- 9) Que donnerait cette analyse de stabilité pour un modèle de type masses-ressorts vérifiant le principe de l'action-réaction, plus précisément dans le cas où la matrice A_{per} est remplacée par une matrice toujours à diagonale dominante, avec des -1 sur la diagonale, mais *symétrique* ?

8.2 Modèles d'ordre 2

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

$$\ddot{x}_j = \frac{1}{\tau}(\varphi(x_{j+1} - x_j) - \dot{x}_j), \quad (8.8)$$

où τ est un temps caractéristique d'accès à une vitesse souhaitée. Pour des voitures, τ représente le temps caractéristique 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

$$w \mapsto U(1 - \exp(-w/w_s)), \quad (8.9)$$

où w_s représente l'ordre de grandeur de la distance considérée par le conducteur comme étant de sécurité (pour une vitesse égale à $1 - 1/e \approx 0.6$ fois la vitesse maximale. Pour un conducteur agressif peu scrupuleux des distances de sécurité, w_s sera donc petit. Nous supposerons pour simplifier les conducteurs tous identiques, ce qui conduit bien au modèle (8.8), avec une fonction φ qui ne dépend pas de i .

Modèle alternatif : prise en compte du temps de réaction

Une approche alternative consiste à enrichir le modèle d'ordre 1 en temps proposé dans la section ?? de la façon suivante : on considère que chaque conducteur ou piétons module sa vitesse en fonction de ce qu'il estime être la distance à l'entité précédente, distance psychologique en quelque sorte, qu'il estime par rapport à la distance réelle instantanée avec un certain retard. On modélise ce retard en considérant que la distance psychologique relaxe vers la vraie distance avec un temps caractéristique $\tau > 0$, pour obtenir

$$\frac{dx_j}{dt} = \varphi(u_j) \quad (8.10)$$

$$\frac{du_j}{dt} = \frac{1}{\tau}(x_{j+1} - x_j - u_j). \quad (8.11)$$

La prise en compte de ce retard est assez similaire à la prise en compte d'une inertie mécanique. On peut en particulier vérifier que l'étude de stabilité au voisinage des points d'équilibre est parfaitement semblable. En revanche la philosophie est différente. Le modèle avec inertie exprime en particulier qu'il est impossible à une entité de s'arrêter brusquement. Pour le modèle avec retard, un arrêt brusque n'est a priori pas exclu : on peut considérer qu'une entité est subitement sortie du modèle du fait d'une perturbation extérieure (ou d'une volonté propre interne), qui la conduit à s'arrêter brusquement.

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

$$\left| \begin{array}{lcl} \frac{dx_j}{dt} & = & v_j \\ \frac{dv_j}{dt} & = & \frac{1}{\tau}(\varphi(x_{j+1} - x_j) - v_j), \end{array} \right. \quad (8.12)$$

assure l'existence d'une unique solution maximale, qui est globale d'après la proposition 11.15, page 239. 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 en effet 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 $x \leq 0$, vérifie

$$\ddot{x} = \frac{1}{\tau}(\varphi(-x) - \dot{x}),$$

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'accidents (et donc la durée de vie finie de la solution en tant que trajectoire viable).

8.2.1 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 $w_{eq} = L/N$). La configuration où tous les véhicules roulent à la même vitesse $V = \varphi(w_{eq})$, correspond au régime stationnaire.

Pour étudier la stabilité de cette situation, on travaille sur les variables de distance $w_j = x_{j+1} - x_j$. Le modèle s'écrit pour cette nouvelle variable

$$\ddot{w}_j = \frac{1}{\tau}(\varphi(w_{j+1}) - \varphi(w_j) - \dot{w}_j), \quad (8.13)$$

pour lequel le vecteur $(w_{eq}, w_{eq}, \dots, w_{eq})$ est point fixe. On peut écrire ce modèle $(\dot{w}, \dot{v}) = \Psi(w, v)$, avec $v = \dot{w}$.

$$\left| \begin{array}{lcl} \dot{w}_j & = & v_j \\ \dot{v}_j & = & \frac{1}{\tau}(\varphi(w_{j+1}) - \varphi(w_j) - v_j) \end{array} \right. \quad (8.14)$$

La stabilité du point d'équilibre est conditionnée par les propriétés de la matrice

$$\nabla \Psi|_{y=y_f} = \begin{pmatrix} 0 & \text{Id} \\ \frac{1}{\tau}\varphi'(u_e)A_{\text{per}} & -\frac{1}{\tau}\text{Id} \end{pmatrix}, \quad \text{avec } A_{\text{per}} = \begin{pmatrix} -1 & 1 & 0 & \cdot & 0 \\ 0 & -1 & 1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & \cdot & \cdot & -1 & 1 \\ 1 & 0 & \cdot & 0 & -1 \end{pmatrix} \quad (8.15)$$

La matrice A_{per} est somme de $-\text{Id}$ et d'une matrice circulante C . Cette dernière vérifie $C^n = \text{Id}$, 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

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

le problème aux valeurs propres pour la matrice globale s'écrit donc

$$v = \lambda w, \quad \frac{\varphi'(w_{eq})}{\tau} Aw - \frac{1}{\tau} v = \lambda v \implies \left(\lambda^2 + \frac{\lambda}{\tau} - \frac{\varphi'(u_e)}{\tau} A\right) w = 0$$

Pour tout couple propre $z_k, \mu_k = -1 + \exp\left(\frac{2ik\pi}{n}\right)$ 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

$$\lambda_k^\pm = \frac{1}{2\tau} \left(-1 \pm \sqrt{1 - 4\varphi'(u_e)\tau \left(1 - \exp\left(\frac{2ik\pi}{N}\right) \right)} \right) \quad (8.16)$$

Notons $\alpha = 4\varphi'(u_e)\tau$. Le lieu des λ_k^\pm est donc l'ensemble image du cercle unité par la transformation (bivaluée) dans le plan complexe

$$z \mapsto \left(-1 \pm \sqrt{1 - \alpha(1 - z)} \right) / 2\tau.$$

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 $\text{Re}(z) \leq 1$?

On peut préciser la réponse à cette question :

Lemme 8.17. La racine carrée du cercle centré (sur l'axe réel) en $1 - \alpha$ et de rayon α intersecte le demi espace $\text{Re}(z) > 1$ si et seulement si $\alpha > 2$.

Démonstration. Une première approche consiste à poser le problème à l'envers, en remarquant qu'il y aura des points de l'ensemble recherché qui sont à droite de la droite $\text{Re}(z) = 1$ 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 $x + iy$ tels que $(x + iy)^2 = x^2 - y^2 + 2ixy$ appartienne au cercle centré en $1 - \alpha$ et de rayon α . Il s'agit donc d'une courbe quartique d'équation

$$(x^2 - y^2 - 1 + \alpha)^2 + 4x^2y^2 = \alpha^2,$$

qui contient le point $(1, 0)$. On va chercher à exprimer x en fonction de y , plus précisément x^2 fonction de y^2 , au voisinage de ce point. Nous allons montrer que, pour certaines valeurs de α , $X = x^2$ est fonction strictement croissante de $Y = y^2$. On pose donc $X = x^2$, $Y = y^2$, pour obtenir

$$(X - Y - 1 + \alpha)^2 + 4XY = \alpha^2, \quad \text{soit } \Psi(X, Y) = 0.$$

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

$$\frac{dX}{dY}_{|(1,0)} = -\frac{\partial\Psi/\partial_Y}{\partial\Psi/\partial_X}_{|(1,0)} = \frac{\alpha - 2}{\alpha},$$

qui est > 0 (ie. les abscisses dépassent strictement 1) dès que $\alpha > 2$. □

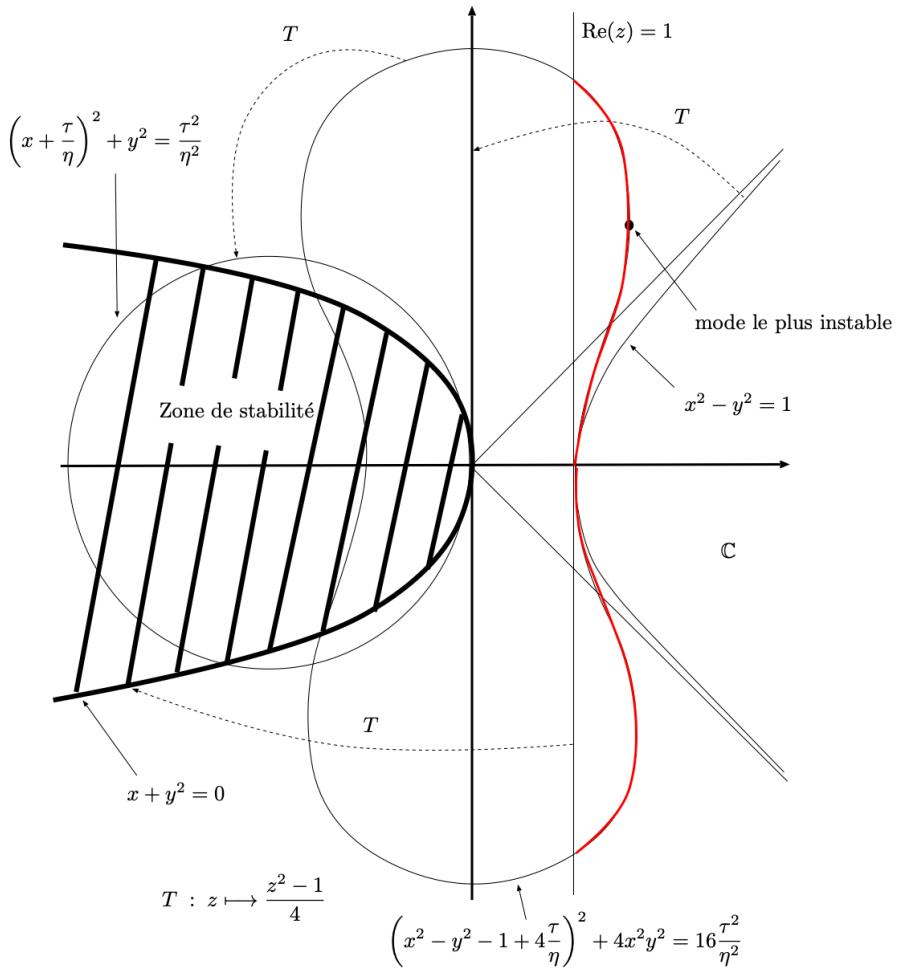


FIGURE 8.4 – Étude spectrale du système linéarisé

Remarque 8.18. 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 8.19. Pour $\alpha = 1/2$, le lieu des valeurs propres est une *lemniscate de Bernoulli* (voir figure 8.5), 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

$$\frac{dX}{dY} = -\frac{X + Y + 1 - \alpha}{X + Y - 1 + \alpha}.$$

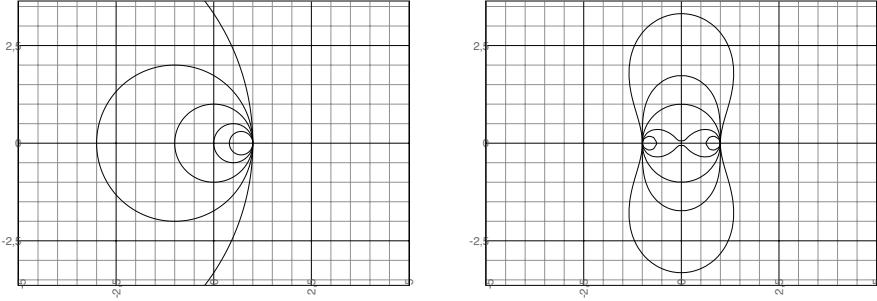


FIGURE 8.5 – Cercles (gauche) et quartiques associées (droite), pour $\alpha = 0.3, 0.5, 1, 2, 6$.

La variable X est donc maximale pour $Y = -X - 1 + \alpha$. En ré-injectant 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 correspondant au cercle centré en $1 - \alpha$, de rayon α . Les abscisses des points de ce cercle s'écrivent $\bar{x}^2 - \bar{y}^2 = X - Y$. L'abscisse, relativement au centre $1 - \alpha$ du cercle, du rayon vecteur correspondant au mode le plus instable est donc

$$x - 1 + \alpha = X - Y - 1 + \alpha = 2X.$$

Comme ce rayon vecteur est de norme α , l'angle s'écrit

$$\theta = \arccos\left(\frac{2X}{\alpha}\right) = \arccos\left(\frac{\alpha}{2(\alpha - 1)}\right).$$

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

$$u_k = \left(e^{2i\pi k \ell/n}\right)_\ell,$$

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 8.6). 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 pratique⁹.

Exercice 8.5. Estimer l'ordre de grandeur de la vitesse de propagation vers l'amont du mode le plus instable, pour $\alpha = 4$.

9. 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 $\eta = \varphi'(u_e)$.

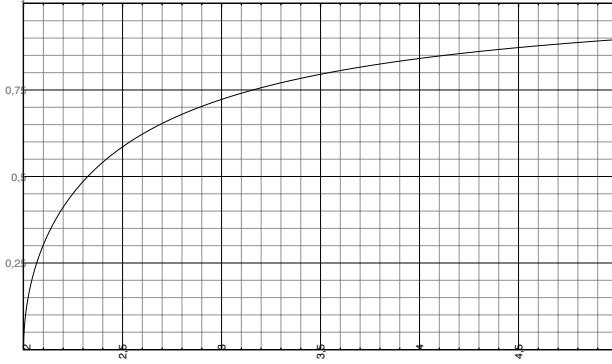


FIGURE 8.6 – Angle θ (mode le plus instable) fonction de α .

Remarque 8.20. 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 8.6, page 172). 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 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) ? Si nous considérons que ce temps prend en compte à la fois le temps de réaction (entre 1 et 3 secondes selon l'état du conducteur) et le temps mis pour décélérer significativement (de l'ordre de 2 secondes), il est raisonnable de considérer un ordre de grandeur de 4 s. La condition indique que l'on aura donc un système plus stable dans le cas d'une bonne réactivité du conducteur et d'une bonne capacité de freinage du véhicule (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. Considérons par exemple le modèle le plus simple utilisé pour approcher le diagramme fondamental en trafic routier, qui correspond à une vitesse exprimée $U(1 - \rho/\rho_{max})$, où ρ est la densité linéique de véhicules. Comme la densité est l'inverse de la distance entre les centres des véhicules ($\rho = 1/w$), cela correspond à une fonction φ qui s'exprime

$$\varphi(w) = U \left(1 - \frac{w_m}{w} \right),$$

d'où $1/\eta = \varphi'(w) = U w_m / w^2$. Avec $w_m = 4$ m, $U = 36 \text{ ms}^{-1}$, et une distance courante de $U = 36$ m (trafic dense), on trouve $\eta \approx 11$ s. On obtient donc avec ces estimations grossières un rapport $\tau/\eta \approx 0.36$ proche de la valeur critique, ce qui suggère qu'il peut être très délicat en pratique de savoir si l'on est dans une situation instable, i.e. susceptible de conduire à une circulation “en accordéon”.

8.2.2 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

$$\ddot{w}_i = \frac{1}{\tau} (\varphi(w_{i+1}) - \varphi(w_i) - \dot{w}_i).$$

La situation $w_i \equiv w_{eq}$ est point d'équilibre du système¹⁰. 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

$$\ddot{w}_i = \frac{1}{\tau} (\varphi'(w_{eq})(w_{i+1} - w_i) - \dot{w}) = \frac{1}{\tau} \left(w_{eq} \varphi'(w_{eq}) \frac{w_{i+1} - w_i}{w_{eq}} - \dot{w} \right)$$

Si l'on considère que les w_i sont les valeurs d'une fonction lisse w aux points équidistants de w_{eq} , on obtient formellement

$$\partial_{tt} w + \frac{1}{\tau} (\partial_t w - c \partial_x w) = 0,$$

avec $c = w_{eq} \varphi'(w_{eq})$.

Exercice 8.6. 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.

Modèle avec retard et inertie.

Comme on peut le vérifier par le calcul, les modèles (8.8) et (8.10) présentent des comportements analogues au voisinage d'une solution d'équilibre (exprimée en distance). Néanmoins les phénomènes modélisés sont distincts, et l'on peut se demander ce qui se passe si les deux sources de retard sont prises en compte simultanément. Le modèle correspondant s'écrit

$$\begin{aligned}\ddot{x}_i &= \frac{1}{\tau} (\varphi(w_i) - \dot{x}_i) \\ \dot{w}_i &= \frac{1}{\tau'} (x_{i+1} - x_i - w_i),\end{aligned}$$

ou, exprimé à l'aide des distances $u_i = x_{i+1} - x_i$,

$$\begin{aligned}\dot{u}_i &= v_i \\ \dot{w}_i &= \frac{1}{\tau'} (u_i - w_i) \\ \dot{v}_i &= \frac{1}{\tau} (\varphi(w_{i+1} - w_i) - \dot{u}_i).\end{aligned}$$

Ce système (avec conditions périodiques) admet un point d'équilibre unique, pour lequel toutes les distances (subjectives et réelles) sont les mêmes.

L'analyse de stabilité est basée sur le spectre de

$$\nabla \Psi = \begin{pmatrix} 0 & 0 & I_d \\ \frac{1}{\tau} \text{Id} & -\frac{1}{\tau} \text{Id} & 0 \\ 0 & \frac{1}{\tau} \varphi'(u_e) A & -\frac{1}{\tau} \text{Id} \end{pmatrix}, \quad \text{avec } A = \begin{pmatrix} -1 & 1 & 0 & \cdot & 0 \\ 0 & -1 & 1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & -1 & 1 \\ 1 & \cdot & \cdot & 0 & -1 \end{pmatrix} = -I + C.$$

La recherche d'éléments propres (u, w, v, λ) peut se faire en considérant successivement les éléments propres (w_k, μ_k) de A . Pour chaque $k = 0, \dots, n-1$, on aura alors 3 valeurs propres, solutions de

$$\lambda^3 + \left(\frac{1}{\tau'} + \frac{1}{\tau} \right) \lambda^2 + \frac{1}{\tau \tau'} \lambda - \frac{1}{\tau \tau'} \frac{1}{\eta} \mu_k = 0.$$

On notera le rôle symétrique joué par les temps caractéristiques τ et τ' . Dans l'asymptotique où tous deux tendent vers 0, on retrouve d'ailleurs l'étude de stabilité effectuée dans la section 8.2, pour un temps effectif qui est la somme des deux temps. Plus précisément, on peut écrire l'équation

$$\tau \tau' \lambda^3 + \left((\tau + \tau') \lambda^2 + \left(\lambda - \frac{1}{\eta} \mu_k \right) \right) = 0.$$

10. On pourra considérer le cas périodique, avec $w_{eq} = L/n$, ou la situation d'entités sur une voie rectiligne, derrière une entité de tête à vitesse fixée égale à $v_e = \varphi(w_{eq})$.

Pour τ et τ' petits, on voit bien sous cette forme la cascade de perturbations singulières de l'identité associée au modèle d'ordre 1, qui donne (voir équation (8.5), page 177) les valeurs propres $\lambda_k = \mu_k/\eta = (-1 + \exp(2ik\pi/n))/\eta$. Le terme en $\tau + \tau'$ rajoute une perturbation d'ordre 2, qui va faire germer deux valeurs propres pour chaque μ_k (selon l'expression (8.16)), avec un temps de relaxation qui est simplement la somme des 2. Le terme en $\tau\tau'$ ajoute une dernière perturbation qui va conduire à l'apparition de 3 valeurs propres pour chaque μ_k . L'identification du lieu des valeurs propres est laissé en exercice ...

8.3 Modèles granulaires de foules

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.

8.3.1 Modèle monodimensionnel

On considère N individus assujettis à se déplacer en ligne droite (comme dans un couloir étroit). Les positions sont notées q_1, \dots, 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 segments de longueur $2r$). On considérera comme admissibles les configurations de

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

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 par¹¹

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

Le problème s'écrit donc

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

Formulation point-selle

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

$$J(v) = \frac{1}{2} |v - U|^2, \tag{8.17}$$

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 dont 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

$$C_q = \{v \in \mathbb{R}^N, Bv \leq 0\}. \tag{8.18}$$

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

Proposition 8.21. Le problème consistant à minimiser la fonctionnelle J (définie par (8.32)) sur C_q (défini par (8.29)) est équivalent à la formulation point-selle suivante

$$\begin{cases} u + B^* p &= U, \\ Bu &\leq 0, \\ p &\geq 0, \\ Bu \cdot p &= 0. \end{cases} \quad (8.19)$$

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 13.23, page 267). La proposition ?? assure donc l'existence d'un vecteur p de multiplicateurs de Lagrange tel que le système (8.33) ci-dessus soit vérifié. Réciproquement, si (u, p) est solution du système, le théorème 13.33, page 271 assure que ce couple est point-selle du Lagrangien

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

et donc que u minimise la fonctionnelle quadratique sous la contrainte $Bu \leq 0$ (d'après la proposition 13.32, page 270). \square

Si l'on considère une rangée de personnes $1, \dots, N$ 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^* 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

$$BB^*p = BU.$$

La matrice BB^* , d'ordre $N - 1$, est exactement la matrice du Laplacien discret en dimension 1 avec conditions de Dirichlet aux extrémités (matrice donnée par (19.16), page 395). 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 8.22. Les remarques précédentes (sur le fait que B encode l'opposé d'une divergence discrète) renforcent l'analogie formelle entre le problème (8.33) et le problème de Darcy, telle qu'elle apparaît pour modéliser les écoulements en milieux poreux (équation (9.11), page 216, ou sous forme plus abstraite dans le cadre des réseaux résistifs (équation (2.10), page 46).

Remarque 8.23. 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 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 changement de vitesse souhaitée d'un individu particulier va changer instantanément les vitesses réelles de tous les individus.

8.3.2 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

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

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

$$C_q = \{v, D_{ij}(q) = |q_j - q_i| - 2r = 0 \Rightarrow G_{ij} \cdot v \geq 0\}. \quad (8.20)$$

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 18.7, page 359) et stable (proposition 18.10).

Formulation point-selle

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

$$J(v) = \frac{1}{2} |v - U|^2, \quad (8.21)$$

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

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

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 8.11)

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

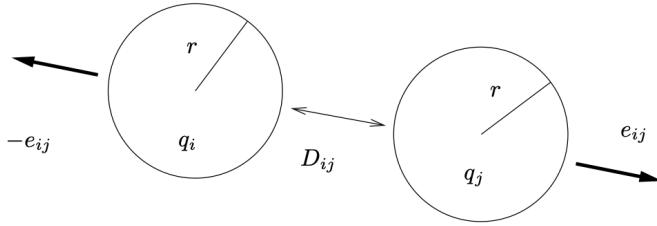


FIGURE 8.7 – Notations.

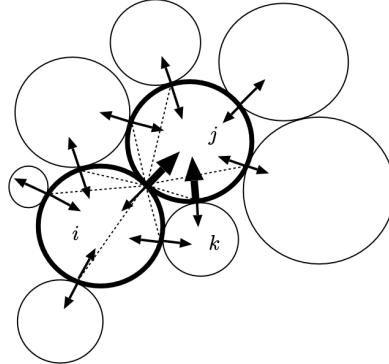


FIGURE 8.8 – Stencil non structuré

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

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

Proposition 8.24. Le problème consistant à minimiser la fonctionnelle J (définie par (8.32)) sur C_q (défini par (8.31)) est équivalent à la formulation point-selle (8.33), qui peut s'exprimer sous la forme suivante

$$\left| \begin{array}{lcl} u - \sum_{i \sim j} p_{ij} G_{ij} & = & U, \\ -G_{ij} \cdot u & \leq & 0 \quad \forall i \sim j, \\ p & \geq 0, \\ G_{ij} \cdot u > 0 & \implies & p_{ij} = 0. \end{array} \right. \quad (8.22)$$

Démonstration. La démonstration est parfaitement analogue à celle de la proposition 8.32. □

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

Considérons une configuration $q \in K$ (voir figure 8.12), 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 $q = (q_1, \dots, q_N)$. L'opérateur discret B^* 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 $-B^*$ réalise l'action de ces forces d'interaction sur le réseau primal de degré de liberté associés aux centres des particules. dans le cas

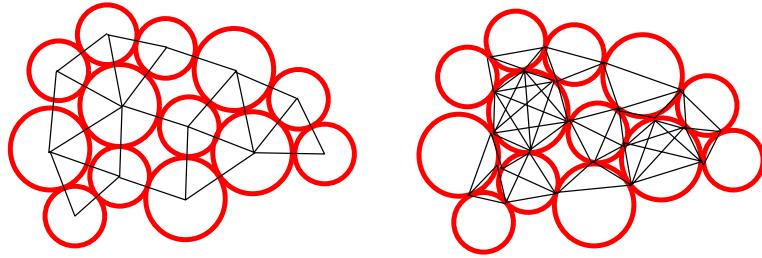


FIGURE 8.9 – Réseaux primal (gauche) et dual (droite)

d'une configuration structurée, (par exemple réseau cartésien, ou réseau triangulaire comme représenté sur la figure 8.14) un champ de pression p uniforme est de gradient discret nul sur les points intérieurs au réseau¹². 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 8.12 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 8.14. Le nombre de disques est 14, donc le nombre 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 $B^* \in \mathcal{M}_{29,28}(\mathbb{R})$ 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 grains¹³. Une conséquence de ces comportements pathologiques est que l'opérateur discret BB^* , 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 8.13) ne vérifie pas le principe du maximum : il peut exister des champs de pression p tels que $BB^* \geq 0$ (i.e. les pressions contribuent à l'augmentation de toutes les distances entre centre), alors que certaines composantes de p sont strictement négatives.

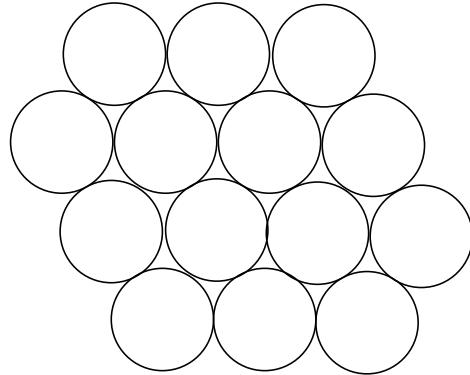


FIGURE 8.10 – Situation hyperstatique (28 degrés de liberté pour 29 contraintes)

L'opérateur discret BB^* 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^*p est un vecteur qui vit lui même sur

12. 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.

13. 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 8.14 : 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

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 8.12. 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 $(q_j - q_i) / \|q_j - q_i\|$. Des exemples de tels vecteurs sont représentés sur la figure 8.12 en gras. Cette propriété est générique pour des collections de disques congestionnées. Certains éléments extra diagonaux de la matrice BB^* sont alors *strictement positifs*, et ainsi la matrice BB^* n'est *pas* une M -matrice¹⁴. Le réseau résistif associé à cet opérateur possède donc des résistances *négatives* : on retrouve la situation de certaines matrices résultant de la discrétisation du Laplacien par éléments finis, sur un maillage contenant des triangles *amblygones* (voir section 17.4, page 356).

14. 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 $p \geq 0$.

8.4 Modèles macroscopiques de trafic routier

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

8.4.1 Modèle d'évolution

On considère l'évolution d'une population de piétons ou de véhicules sur une voie rectiligne, population représentée par une densité linéaire $\rho(x, t)$. 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 chapitre 4)

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

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

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x} f(\rho) = 0, \quad (8.23)$$

où f est le *flux*.

Propagation des perturbations

Si l'on considère une solution stationnaire ρ_{eq} de l'équation, et une solution perturbée $\rho_{eq} + h$, on obtient formellement une équation de transport sur la perturbation :

$$\partial_t h + f'(\rho_{eq}) \partial_x h = 0 \quad (8.24)$$

qui exprime que les perturbations sont transportées à la vitesse $f'(\rho_{eq})$.

Supposons que $\rho(x, t)$ est une solution régulière de cette équation. On appelle courbe caractéristique une courbe $t \mapsto x(t)$ 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 caractéristique, la célérité (fonction de cette seule densité) elle-même est constante, et l'on a

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

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

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

Si l'on suppose que la fonction f est C^2 , on peut calculer le jacobien 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 8.25. 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 8.1, en notant que la densité linéique (nombre de véhicules ou de piétons par mètre) est l'inverse de la distance entre les personnes : $\rho = 1/w$. Si l'on reprend la fonction φ qui définit la vitesse comme fonction de la distance, on a

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

qui, exprimée en distance locale $w_{eq} = 1/\rho_{eq}$, donne

$$f'(\rho) = \varphi(w_{eq}) - w_{eq} \varphi'(w_{eq}).$$

Si l'on s'intéresse à l'évolution d'une perturbation autour d'une densité uniforme ρ_{eq} , l'équation (8.24), exprime un transport à la vitesse $f'(\rho_{eq})$. On retrouve au niveau macroscopique la vitesse de propagation vers l'amont $-w_{eq} \varphi'(w_{eq})$ trouvée dans la section 8.1. La vitesse macroscopique contient naturellement le terme de vitesse des entités $\varphi(w_{eq})$, 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 microscopique qui est naturellement *lagrangienne* (les variables sont afférentes aux entités en mouvement).

Remarque 8.26. Il est très facile 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 nous 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) :

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

8.4.2 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 sont susceptibles d'apparaître spontanément, on définit la notion de solution faible :

Definition 8.27. On dit que $\rho \in L^1_{loc}(\mathbb{R} \times]0, T[)$ est une solution faible de (8.23) sur $\mathbb{R} \times]0, T[$ si $f(\rho) \in L^1_{loc}(\mathbb{R} \times]0, T[)$ et si, pour tout φ , fonction C^1 à support compact dans $\mathbb{R} \times]0, T[$, on a

$$\int_{\mathbb{R}} \int_0^T \partial_t \varphi \rho(x, t) dx dt + \int_{\mathbb{R}} \int_0^T \partial_x \varphi f(\rho(x, t)) dx dt = 0. \quad (8.25)$$

On peut intégrer une condition initiale à cette définition. On dira que ρ est solution faible associée à la condition initiale $\rho|_{t=0} = \rho^0 \in L^1_{loc}(\mathbb{R})$ si

$$\int_{\mathbb{R}} \int_0^T \partial_t \varphi \rho(x, t) dx dt + \int_{\mathbb{R}} \int_0^T \partial_x \varphi f(\rho(x, t)) dx dt + \int_{\mathbb{R}} \varphi(x, 0) \rho^0(x) dx = 0$$

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

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

est solution faible stationnaire de (8.23), de même que la densité obtenue en intervertissant ρ_- et ρ_+ (ces propriétés sont des cas particuliers de la proposition 8.27 démontrée ci-après). 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 8.28. (Relation de Rankine-Hugoniot)

On suppose la fonction flux f continue sur son intervalle de définition, et ρ_L et ρ_R deux valeurs sur cet intervalle. La densité

$$\rho = \rho_L \mathbf{1}_{]-\infty, s(t)[} + \rho_R \mathbf{1}_{]s(t), +\infty[} \quad (8.26)$$

est solution faible de (8.23) si et seulement si la discontinuité s progresse à la vitesse constante

$$\dot{s} = \frac{f(\rho_L) - f(\rho_R)}{\rho_L - \rho_R}. \quad (8.27)$$

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

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

On obtient donc finalement

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

qui est identiquement nul pour toute fonction test φ si et seulement si la condition (8.27) est identiquement vérifiée. \square

Remarque 8.29. Si l'on admet que la quantité définie par 8.26 est solution, on peut retrouver la relation (8.27) en écrivant simplement un bilan de masse au voisinage de la discontinuité. Plus précisément, considérant \bar{s} la position du front à un certain temps t , la dérivée de la masse contenue dans $[x - \eta, x + \eta]$ peut s'écrire de deux manières (en intégrant ρ sur l'intervalle considérer et en dérivant, on en faisant le bilan des flux à gauche et à droite)

$$\frac{d}{dt} \int_{\bar{s}-\eta}^{\bar{s}+\eta} \rho(x, t) dx = \dot{s} (\rho_R - \rho_L) = f(\rho_R) - f(\rho_L).$$

Remarque 8.30. On peut voir cette formule comme la généralisation de la formule donnant la vitesse de propagation de perturbations au voisinage d'une densité uniforme, en prenant $\rho_R = \rho_L + \varepsilon$, ce qui donne $\dot{s} \approx f'(\rho_L)$.

On peut vérifier que, sous sa forme faible, l'équation n'est pas bien posée, au sens où elle admet en général plusieurs solutions. On pourra considérer par exemple la donnée initiale $\rho = \mathbf{1}_{]-\infty, 0[}$, que l'on peut voir comme une quantité infinie de véhicules à l'arrêt à un feu en 0, qui passe au vert à l'instant initial. On vérifie immédiatement (c'est une conséquence de la proposition 8.28) que c'est une solution faible stationnaire de l'équation.

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 physique¹⁵ 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. Le développement précédent donnant la vitesse de propagation de la discontinuité en fonction des états à gauche et à droite, on peut exprimer le fait que les caractéristiques vont vers la discontinuité de la façon suivante :

Definition 8.31. Soit $\rho(x, t)$ une solution faible de l'équation de conservation (8.23), avec $f(\cdot)$ une fonction C^1 , au sens de la définition 8.27. On suppose que ρ présente localement (au voisinage d'un point de l'espace temps) une discontinuité entre les valeurs ρ_L et ρ_R . On dit que cette discontinuité vérifie la condition d'entropie de Lax si

$$f'(\rho_L) > \frac{f(\rho_R) - f(\rho_L)}{\rho_R - \rho_L} > f'(\rho_R).$$

On notera que, dans le cas où f est convexe (ou f concave), la condition ci-dessus peut se limiter à l'inégalité entre les bornes.

8.5 Modèles granulaires de foules

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.

8.5.1 Modèle monodimensionnel

On considère N individus assujettis à se déplacer en ligne droite (comme dans un couloir étroit). Les positions sont notées q_1, \dots, 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 considérera comme admissibles les configurations de

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

¹⁵. 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.

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 par¹⁶

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

Le problème s'écrit donc

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

Formulation point-selle

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

$$J(v) = \frac{1}{2} |v - U|^2, \tag{8.28}$$

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 dont 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

$$C_q = \{v \in \mathbb{R}^N, Bv \leq 0\}. \tag{8.29}$$

Proposition 8.32. Le problème consistant à minimiser la fonctionnelle J (définie par (8.32)) sur C_q (défini par (8.29)) est équivalent à la formulation point-selle suivante

$$\begin{cases} u + B^* p &= U, \\ Bu &\leq 0, \\ p &\geq 0, \\ Bu \cdot p &= 0. \end{cases} \tag{8.30}$$

Plus précisément, u étant la solution du problème de minimisation sous 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 13.23, page 267). La proposition ?? assure donc l'existence d'un vecteur p de multiplicateurs de Lagrange tel que le système (8.33) ci-dessus soit vérifié. Réciproquement, si (u, p) est solution du système, le théorème 13.33, page 271 assure que ce couple est point-selle du Lagrangien

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

et donc que u minimise la fonctionnelle quadratique sous la contrainte $Bu \leq 0$ (d'après la proposition 13.32, page 270). \square

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

Si l'on considère une rangée de personnes $1, \dots, N$ 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^* 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

$$BB^*p = BU.$$

La matrice BB^* , d'ordre $N - 1$, est exactement la matrice du Laplacien discret en dimension 1 avec conditions de Dirichlet aux extrémités (matrice donnée par (19.16), page 395). 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 8.33. Les remarques précédentes (sur le fait que B encode l'opposé d'une divergence discrète) renforcent l'analogie formelle entre le problème (8.33) et le problème de Darcy, telle qu'elle apparaît pour modéliser les écoulements en milieux poreux (équation (9.11), page 216, ou sous forme plus abstraite dans le cadre des réseaux résistifs (équation (2.10), page 46).

Remarque 8.34. 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 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 changement de vitesse souhaitée d'un individu particulier va changer instantanément les vitesses réelles de tous les individus.

8.5.2 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

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

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

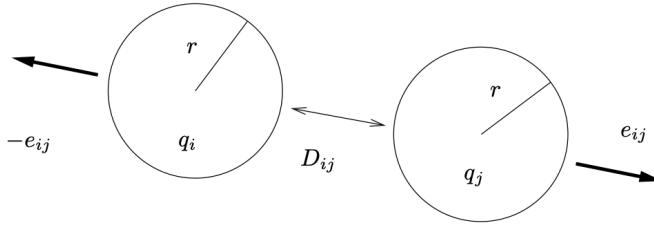


FIGURE 8.11 – Notations.

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

$$C_q = \{v, D_{ij}(q) = |q_j - q_i| - 2r = 0 \Rightarrow G_{ij} \cdot v \geq 0\}. \quad (8.31)$$

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 18.7, page 359) et stable (proposition 18.10).

Formulation point-selle

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

$$J(v) = \frac{1}{2} |v - U|^2, \quad (8.32)$$

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

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

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 8.11)

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

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

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

Proposition 8.35. Le problème consistant à minimiser la fonctionnelle J (définie par (8.32)) sur C_q (défini par (8.31)) est équivalent à la formulation point-selle (8.33), qui peut s'exprimer sous la forme suivante

$$\left| \begin{array}{lcl} u - \sum_{i \sim j} p_{ij} G_{ij} & = & U, \\ -G_{ij} \cdot u & \leq & 0 \quad \forall i \sim j, \\ p & \geq 0, \\ G_{ij} \cdot u > 0 & \implies & p_{ij} = 0. \end{array} \right. \quad (8.33)$$

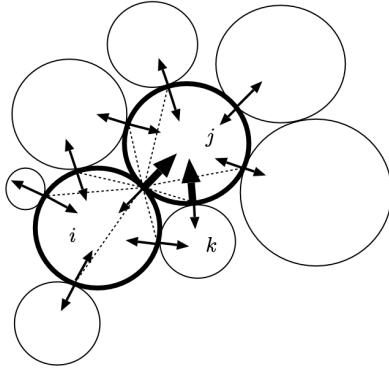


FIGURE 8.12 – Stencil non structuré

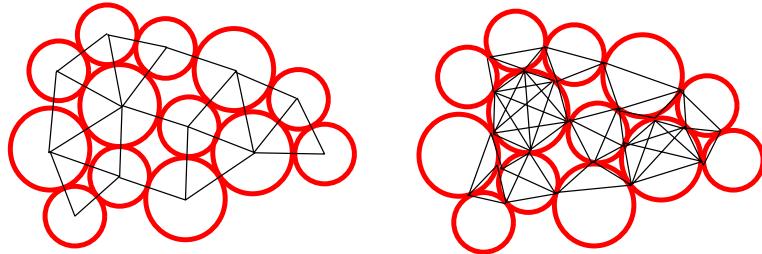


FIGURE 8.13 – Réseaux primal (gauche) et dual (droite)

Démonstration. La démonstration est parfaitement analogue à celle de la proposition 8.32. \square

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

Considérons une configuration $q \in K$ (voir figure 8.12), 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 $q = (q_1, \dots, q_N)$. L'opérateur discret B^* 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 $-B^*$ 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 8.14) un champ de pression p uniforme est de gradient discret nul sur les points intérieurs au réseau¹⁷. 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 8.12 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 8.14. Le nombre de disques est 14, donc le nombre de degrés de liberté primaux est 28, et le nombre de contacts (nombre de degrés de liberté duals) est 29. En conséquence, le noyau de $B^* \in M_{29,28}(\mathbb{R})$ 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

17. 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.

sur les grains¹⁸. Une conséquence de ces comportements pathologiques est que l'opérateur discret BB^* , 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 8.13) ne vérifie pas le principe du maximum : il peut exister des champs de pression p tels que $BB^* \geq 0$ (i.e. les pressions contribuent à l'augmentation de toutes les distances entre centre), alors que certaines composantes de p sont strictement négatives.

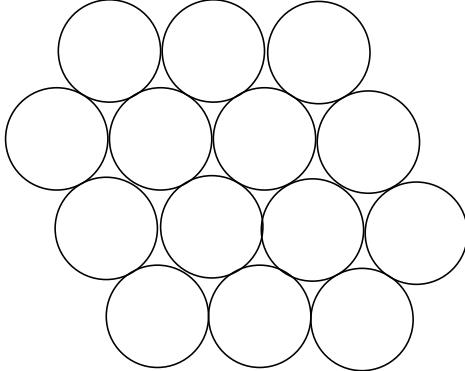


FIGURE 8.14 – Situation hyperstatique (28 degrés de liberté pour 29 contraintes)

L'opérateur discret BB^* 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^*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 8.12. 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 $(q_j - q_i)/|q_j - q_i|$. Des exemples de tels vecteurs sont représentés sur la figure 8.12 en gras. Cette propriété est générique pour des collections de disques congestionnées. Certains éléments extra diagonaux de la matrice BB^* sont alors *strictement positifs*, et ainsi la matrice BB^* n'est *pas* une M -matrice¹⁹. Le réseau résistif associé à cet opérateur possède donc des résistances *négatives* : on retrouve la situation de certaines matrices résultant de la discrétisation du Laplacien par éléments finis, sur un maillage contenant des triangles *amblygones* (voir section 17.4, page 356).

18. 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 8.14 : 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

19. 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 $p \geq 0$.

Chapitre 9

Éléments de mécanique des fluides

Sommaire

9.1	Tenseur des contraintes, équation générale du mouvement	206
9.2	Fluides parfaits	208
9.2.1	Fluide parfait incompressible	209
9.2.2	Fluide parfait barotrope	210
9.3	Fluides newtoniens	211
9.4	Bilan d'énergie pour les équations de Navier-Stokes	213
9.5	Écoulements en milieu poreux	215
9.6	Cadre mathématique pour le problème de Darcy	216
9.7	Cadre mathématique pour les équations de Stokes	217
9.8	Ecoulement de Poiseuille, notion de résistance	219
9.9	Ecoulement autour d'une sphère	221

9.1 Tenseur des contraintes, équation générale du mouvement

Definition 9.1. (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 point¹ 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 $n \mapsto F(n)$ est linéaire, on appelle tenseur² 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 vitesse³, et par f une force en volume agissant sur le fluide (typiquement la gravité $f = \rho g$). On considère

1. 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.

2. 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.

3. 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 4.1, page 84), 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

un système matériel $\omega(t)$, c'est à dire à ensemble de particules que l'on suit dans leur mouvement⁴. 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 :

$$\frac{d}{dt} \int_{\omega(t)} \rho u = \text{somme des forces extérieures.} \quad (9.1)$$

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 9.1,

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

Le membre de gauche de 9.1 s'écrit donc

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

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

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

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 9.2. (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

$$\frac{\partial}{\partial t} (\rho u) + \nabla \cdot (\rho u \otimes u) - \nabla \cdot \sigma = f. \quad (9.2)$$

La conservation de la masse s'écrit par ailleurs

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

Modèle 9.3. (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 9.4. 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 inertuelle) 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

“points” sont en fait dans le vide, et cela n'a pas de sens de définir 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.

4. 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.

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

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

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

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

$$\begin{aligned} & -\int_0^T m \partial_t \varphi(x(t), t) u_t - \int_0^T \partial_x \varphi(x(t), t) u(t)^2 \\ &= -\int_0^T m u(t) \left(\underbrace{\partial_t \varphi(x(t), t) u_t + \partial_x \varphi(x(t), t) u_t}_{d\varphi(x(t), t)/dt} \right) = \int_0^T mg \varphi(x(t), t). \end{aligned}$$

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 Eulerienne par une solution faible de l'équation (dite d'Euler sans pression) ci-dessus. En prenant par exemple un forçage extérieur nul, et

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

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 énergétique. 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.

9.2 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 9.5. (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

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

où Id est le tenseur identité.

Pour un tel fluide, on a

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

ce qui conduit à l'équation d'Euler

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

9.2.1 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 9.6. (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 pression), homogène, et incompressible. Le triplet (ρ, u, p) vérifie alors les Équation d'Euler incompressibles

$$\left| \begin{array}{lcl} \rho \frac{\partial u}{\partial t} + \rho (u \cdot \nabla) u + \nabla p & = & f \\ \nabla \cdot u & = & 0 \end{array} \right. \quad (9.3)$$

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*⁵ : 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 9.7. (“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

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

se conserve le long des lignes de courant.

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 produit scalaire avec u de la première ligne de (9.3), sans le terme de dérivée en temps (supposé nul),

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

d'où la propriété annoncée. □

5. 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://www.numdam.org/item/AST_2010__332__101_0.pdf

9.2.2 Fluide parfait barotrope

Une autre manière de fermer⁶ les équations d'Euler est de supposer un lien univoque entre la densité et la pression. On obtient alors le

Modèle 9.8. (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

$$\left| \begin{array}{l} \frac{\partial \rho}{\partial t} + \nabla \cdot (\rho u) = 0, \\ \frac{\partial}{\partial t} (\rho u) + \nabla \cdot (\rho u \otimes u) + \nabla p = f \\ p = p(\rho). \end{array} \right. \quad (9.4)$$

Équations de l'acoustique

Le modèle précédent 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

$$\partial_t \rho + \nabla \cdot (\rho u) = 0, \quad (9.5)$$

$$\frac{\partial}{\partial t} (\rho u) + \nabla \cdot (\rho u \otimes u) + \nabla p = 0, \quad (9.6)$$

avec $p = p(\rho)$. 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 hypothèse 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 :

$$\begin{aligned} \partial_t \rho + \rho_0 \nabla \cdot u &= 0, \\ \rho_0 \partial_t u + \nabla p &= 0. \end{aligned} \quad (9.7)$$

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

$$\frac{\partial^2 \rho}{\partial t^2} - \beta \Delta \rho = 0,$$

avec $\beta = p'(\rho_0)$, 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

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

6. 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).

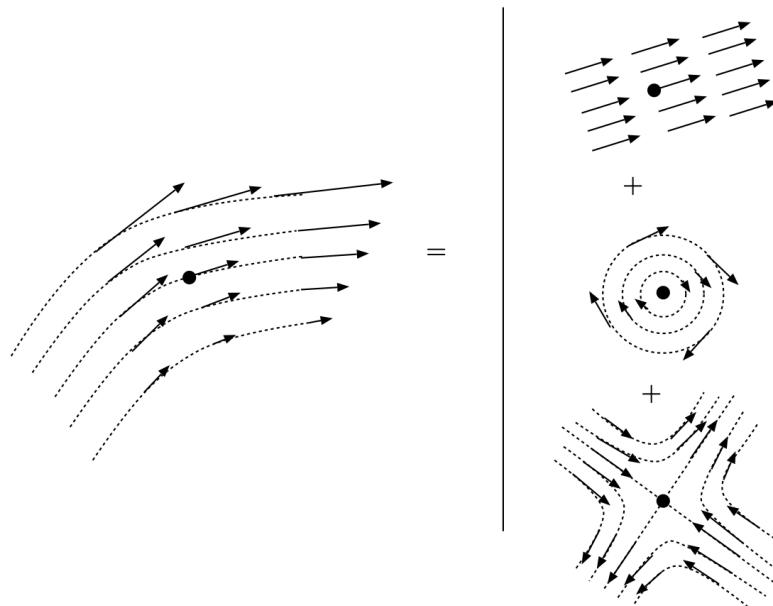


FIGURE 9.1 – Décomposition locale d'un champ de vitesse

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

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

9.3 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'un point x . La vitesse au voisinage de x s'écrit

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

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 9.1 pour une illustration (en dimension 2 d'espace) de ces trois contributions.

Definition 9.9. (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 9.10. (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'écrit

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

où $p = p(x, t)$ 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

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

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 9.11. (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

$$\begin{cases} \rho \left(\frac{\partial u}{\partial t} + (u \cdot \nabla) u \right) - \mu \Delta u + \nabla p = f \\ \nabla \cdot u = 0. \end{cases}$$

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

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

En notant ∇^* (resp. Δ^*) le gradient (resp. le Laplacien) relativement à la variable d'espace adimensionnée, on obtient

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

où $p^* = p/(\rho U^2)$ est la pression adimensionnée, et $f^* = f L/(\rho U^2)$ le terme de forçage adimensionné.

Definition 9.12. Le nombre $\text{Re} = \rho U L / \mu$ 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 9.13. (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

$$\begin{cases} -\mu \Delta u + \nabla p = f \\ \nabla \cdot u = 0 \end{cases} \tag{9.8}$$

Remarque 9.14. 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 domaine délimité par des murs physiques imperméables, on considère en général⁷ que le fluide accroche à la paroi, ce qui s'exprime sous la forme de *conditions de Dirichlet homogènes* $u = 0$ sur la frontière $\partial\Omega$.

9.4 Bilan d'énergie pour les équations de Navier-Stokes

Nous établissons ici formellement (en supposant les champs suffisamment réguliers pour que toutes les dérivées puissent être définies au sens classique) un bilan d'énergie pour un fluide newtonien incompressible, dans différentes situations.

Conditions de Dirichlet homogène (adhérence aux parois) Multiplying the momentum equations by the velocity itself and integrating over the domain yields

$$\rho \int_{\Omega} \frac{\partial u}{\partial t} \cdot u + \rho \int_{\Omega} ((u \cdot \nabla) u) \cdot u - \mu \int_{\Omega} \Delta u \cdot u + \int_{\Omega} \Omega u \cdot \nabla p = \int_{\Omega} f \cdot u.$$

The right-hand side is the power of external forces. The first term can be written

$$\frac{d}{dt} \int_{\Omega} \frac{\rho}{2} |u|^2,$$

it is the time derivative of the kinetic energy. For the second one, we can use the formula

$$(u \cdot \nabla) u = \frac{1}{2} \nabla |u|^2 - u \times (\nabla \times u),$$

which leads to

$$\rho \int_{\Omega} ((u \cdot \nabla) u) \cdot u = \rho \int_{\Omega} \frac{1}{2} \nabla |u|^2 \cdot u = \int_{\Gamma} \frac{\rho}{2} |u|^2 u \cdot n,$$

that vanishes since u is zero on the boundary. The viscous term is

$$-\mu \int_{\Omega} \Delta u \cdot u = \mu \int_{\Omega} |u|^2,$$

it corresponds to viscous dissipation.

We finally obtain that the time derivative of kinetic energy equals the power of external forces (rate of energy injected into the system) minus the rate of energy lost as heat by viscous effects.

Bilan énergétique d'un objet rigide volant à vitesse constante.

On s'intéresse ici au mouvement d'un objet rigide de type avion dans un fluide visqueux newtonien. On note $U e_x$ la vitesse de l'avion, avec $U > 0$, et l'on note ω_t le domaine occupé par l'avion à l'instant t . On se place en un temps t (pris égal à 0), et l'on suppose que la vitesse du fluide est nulle loin de l'objet, plus précisément qu'il existe un (grand) domaine Ω_t , contenant ω_t et se déplaçant en translation avec lui, à la frontière duquel la vitesse du fluide est nul.

7. 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 pénétration du fluide dans la paroi, mais autorisent une vitesse tangentielle non nulle.

$$\left\{ \begin{array}{ll} \rho \frac{\partial u}{\partial t} + \rho (u \cdot \nabla) u - \mu \Delta u + \nabla p & = g \quad \text{dans } \Omega_t \setminus \omega_t \\ \nabla \cdot u & = 0 \quad \text{dans } \Omega_t \setminus \omega_t \\ \nabla \cdot u & = U e_x \quad \text{on } \partial \omega(t) \\ \nabla \cdot u & = 0 \quad \text{on } \partial \Omega(t). \end{array} \right.$$

On effectue le produit scalaire de l'équation de conservation de la quantité de mouvement avec la vitesse elle même, et l'on intègre sur le domaine $\Omega_t \setminus \omega_t$. On a

$$\int_{\Omega_t \setminus \omega_t} \rho \frac{\partial u}{\partial t} = \frac{d}{dt} \int_{\Omega_t \setminus \omega_t} \frac{1}{2} \rho |u|^2 - \int_{\partial \Omega_t \cup \partial \omega_t} \frac{1}{2} \rho |u|^2 U e_x \cdot n.$$

L'intégrale de bord ci-dessus est nulle sur $\partial \Omega_t$ (la vitesse y est nulle), et

$$\int_{\partial \omega_t} \frac{1}{2} \rho |u|^2 U e_x \cdot n = \frac{1}{2} \rho |U|^2 U e_x \cdot \int_{\partial \omega_t} n = 0.$$

On a par ailleurs

$$\int_{\Omega_t \setminus \omega_t} \rho (u \cdot \nabla) u \cdot u = \int_{\Omega_t \setminus \omega_t} \frac{1}{2} \rho \nabla |u|^2 \cdot u = \int_{\Omega_t \setminus \omega_t} \nabla \cdot \left(\frac{1}{2} \rho \nabla |u|^2 u \right) = \int_{\partial \Omega_t \cup \partial \omega_t} \frac{1}{2} \rho |u|^2 u \cdot n,$$

qui est nul pour les mêmes raisons que précédemment.

Pour le traitement des termes de viscosité et de pression, on écrit maintenant

$$\Delta u = \nabla \cdot (\nabla u + {}^t \nabla u).$$

On a

$$\begin{aligned} -\mu \int_{\Omega_t \setminus \omega_t} \nabla \cdot (\nabla u + {}^t \nabla u) \cdot u + \int_{\Omega_t \setminus \omega_t} u \cdot \nabla p = \\ \frac{\mu}{2} \int_{\Omega_t \setminus \omega_t} |\nabla u + {}^t \nabla u|^2 + \int_{\partial \Omega_t \cup \partial \omega_t} (-(\nabla u + {}^t \nabla u) \cdot n + pn) \cdot u. \end{aligned}$$

L'intégrale de bord ci-dessus est nulle sur $\partial \Omega_t$ (la vitesse y est nulle), elle s'écrit donc

$$\int_{\partial \omega_t} (-\sigma \cdot n) \cdot u.$$

Le champ n ci-dessus est la normale sortante au domaine fluide, la force élémentaire exercée par le fluide sur la surface de l'objet est donc $-\sigma \cdot n$. L'intégrale ci-dessus s'écrit donc

$$\int_{\partial \omega_t} F \cdot u = U e_x \cdot \left(\int_{\partial \omega_t} F \right) = -UT,$$

où $T \geq 0$ est la *trainée*, c'est à dire la composante horizontale de la force exercée par le fluide sur l'objet, comptée positivement lorsqu'elle s'oppose à la vitesse.

On a donc finalement

$$\frac{d}{dt} \int_{\Omega_t \setminus \omega_t} \frac{1}{2} \rho |u|^2 + \frac{\mu}{2} \int_{\Omega_t \setminus \omega_t} |\nabla u + {}^t \nabla u|^2 = TU.$$

Le premier terme correspond à la variation en temps de l'énergie cinétique. Si l'on considère un objet volant de façon stationnaire, il est raisonnable de penser que cette quantité oscille autour de 0. Le bilan moyen s'écrit donc

$$\frac{\mu}{2} \int_{\Omega_t \setminus \omega_t} |\nabla u + {}^t \nabla u|^2 = TU, \tag{9.9}$$

qui exprime simplement que la puissance \mathcal{P} développée par l'avion, égale à TU , s'équilibre avec la puissance dissipée au sein du fluide visqueux (qui se retrouve dans l'air environnant sous forme de chaleur).

On peut poursuivre cette démarche dans le cas d'un avion, caractérisé par sa *finesse* f , qui est le rapport entre la portance P (composante verticale de la force) et la trainée T , pour les conditions standard d'utilisation (vitesse de croisière) :

$$f = \frac{P}{T}.$$

La portance équilibrant le poids Mg de l'avion, on a donc

$$\mathcal{P} = TU = U \frac{P}{f} = \frac{1}{f} UMg.$$

Pour un trajet de longueur L parcouru à vitesse U , le temps est $\tau = L/U$, d'où une énergie totale dépensé égale à

$$W = \tau \mathcal{P} = \frac{1}{f} LMg.$$

9.5 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*⁸ rigide et fixe (typiquement un amas tridimensionnel 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 9.12). On a alors une relation de proportionnalité entre flux de fluide et gradient de pression. Plus précisément, Darcy a mis en évidence (voir figure 9.2) 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 9.15. (Loi de Darcy en milieu isotrope)

On considère l'écoulement d'un fluide visqueux dans un milieu poreux saturé⁹.

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

$$u = -k \nabla p, \tag{9.10}$$

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

Remarque 9.16. La notion de vitesse moyenne évoquée ci-dessus correspond en fait à un flux (volumique) par unité de surface. Cette quantité, en $\text{m}^3 \text{ s}^{-1}$ 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.

8. 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.

9. On dit que le milieu 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://pablonet.princeton.edu/pgd/papers/prl/prl84p02064.pdf>

9. 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.

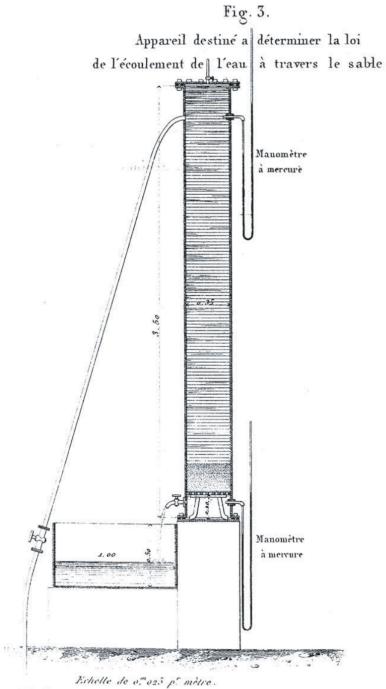


FIGURE 9.2 – Description de l’expérience de Darcy (1856)

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 9.17. (Écoulement en milieu poreux)

L’écoulement en milieu poreux saturé d’un fluide visqueux incompressible est régi par

$$\begin{cases} u + k\nabla p = U \\ \nabla \cdot u = 0 \end{cases} \quad (9.11)$$

où p est la pression au sein du fluide, u la vitesse de Darcy (voir remarque 9.16), $k = K/\mu$ 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).

9.6 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

$$\begin{cases} u + \nabla p = U & \text{dans } \Omega, \\ \nabla \cdot u = 0 & \text{dans } \Omega, \\ p = 0 & \text{sur } \Gamma, \end{cases} \quad (9.12)$$

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

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

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

$$\begin{cases} u \in K = \left\{ v \in L^2(\Omega)^2, \int_{\Omega} v \cdot \nabla q = 0 \quad \forall q \in H_0^1(\Omega) \right\}, \\ J(u) = \inf_{v \in K} J(v), \quad \text{avec } J(v) = \frac{1}{2} \int_{\Omega} |v|^2 - \int_{\Omega} v \cdot f. \end{cases} \quad (9.13)$$

Proposition 9.18. Soit Ω un domaine borné de frontière Lipschitz, et $U \in L^2(\Omega)^d$. Le problème de minimisation (9.13) ci-dessus admet une solution unique $u \in K$, et il existe un unique $p \in V = H_0^1(\Omega)$ tel que

$$u + \nabla p = U \quad p.p.$$

Démonstration. Le problème (9.13) 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. Notons en premier lieu que, du fait que B a été défini à valeur dans l'espace dual d'un espace de Hilbert, sans que l'on fasse l'identification entre les deux espaces, B^* est naturellement défini de Λ dans V' . On peut vérifier que l'application B est surjective, car son adjoint

$$B^* : q \in H_0^1(\Omega) \longmapsto \nabla q \in L^2(\Omega)^2$$

est tel que

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

d'après l'inégalité de Poincaré, ce qui assure bien la surjectivité de B selon la proposition ??, page ??.
D'après la proposition 13.54, page 278, on a donc existence d'un multiplicateur de Lagrange p tel que $u + \nabla p = U$, qui est unique du fait du caractère injectif du gradient sur $H_0^1(\Omega)$. \square

9.7 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

$$\begin{cases} -\Delta u + \nabla p &= f, \\ \nabla \cdot u &= 0, \end{cases} \quad (9.14)$$

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

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

On considère le problème de minimisation sous contrainte

$$\begin{cases} u \in K, \\ J(u) = \inf_{v \in K} J(v), \end{cases} \quad (9.15)$$

où J est la fonctionnelle

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

Proposition 9.19. La fonctionnelle J admet un unique minimiseur sur K , caractérisé par

$$\int_{\Omega} \nabla u : \nabla v = \int_{\Omega} f \cdot v \quad \forall v \in K.$$

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, caractérisé par la formulation variationnelle annoncée. \square

Ce premier résultat assure le caractère bien posé du problème dans un certain sens, mais il est manifestement incomplet puisque, du fait que le problème de minimisation a été posé dans l'espace constraint, la pression (multiplicateur de Lagrange de la contrainte) a disparu. Or cette pression est plus qu'un auxiliaire abstrait, elle a un sens physique, et il est important de lui donner un statut mathématique, et d'aboutir à un résultat d'existence et d'unicité qui porte véritablement sur la forme complète du problème de Stoke (9.15).

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

$$\Lambda = L_0^2(\Omega) = \left\{ p \in L^2(\Omega), \int_{\Omega} p = 0 \right\},$$

et l'opérateur

$$\mathcal{B} : v \in V \longmapsto \mathcal{B}v = -\nabla \cdot v.$$

L'espace K peut s'écrire

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

ce qui conduit au Lagrangien

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

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

Proposition 9.20. 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 $(u, p) \in V \times \Lambda$, où u est la solution du problème de minimisation sous contrainte (9.15). De façon équivalente, il existe un unique couple $(u, p) \in H_0^1(\Omega)^N \times L_0^2(\Omega)$ tel que

$$\int_{\Omega} \nabla u : \nabla v - \int_{\Omega} p \nabla \cdot v = \int_{\Omega} f \cdot v \quad \forall v \in H_0^1(\Omega)^N \tag{9.16}$$

$$\int_{\Omega} q \nabla \cdot u = 0 \quad \forall q \in L_0^2(\Omega). \tag{9.17}$$

Démonstration. Malgré l'analogie formelle avec le problème de Darcy (l'opérateur \mathcal{B} est l'opérateur de divergence dans les deux cas), la démonstration est plus délicate (voir par exemple [?]). L'existence et l'unicité d'un point-selle est une conséquence de la surjectivité de l'opérateur de divergence \mathcal{B} , qui est assurée par le lemme 9.21 ci-après. \square

Lemme 9.21. Soit Ω un domaine connexe, borné, de frontière Γ Lipschitzienne, et soit q dans $L_0^2(\Omega)$. Il existe $v \in H_0^1(\Omega)$ tel que $\nabla \cdot v = q$.

Démonstration. On se reportera à [?, lemme 3.2] pour la démonstration, assez délicate, 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 $\|v\|_{H^1} \leq C \|q\|_{L^2}$. \square

Remarque 9.22. 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 \mathcal{B}^* et le caractère fermé de son image. L'opérateur \mathcal{B}^* va de $H_0^1(\Omega)$ dans $L^2(\Omega)^2$, et l'inégalité est conséquence directe de l'inégalité de Poincaré

$$\|q\|_{L^2(\Omega)} \leq C \|\nabla q\|_{L^2(\Omega)^N} \quad \forall q \in H_0^1(\Omega).$$

Dans le cas de Stokes, la surjectivité de l'opérateur \mathcal{B} peut être établie comme conséquence directe d'une inégalité à première vue très similaire, l'opérateur \mathcal{B}^* étant toujours dans un certain sens l'opérateur de gradient, mais vu cette fois comme un opérateur de $L^2(\Omega)$ dans $H^{-1}(\Omega) = (H_0^1(\Omega)^N)'$. Cette inégalité peut s'écrire

$$\|q\|_{L^2(\Omega)} \leq C \|\nabla q\|_{H^{-1}(\Omega)} \quad \forall q \in L_0^2(\Omega),$$

où ∇q représente la forme linéaire sur $H_0^1(\Omega)^N$ définie par

$$v \longmapsto \int_{\Omega} q \nabla \cdot v, \quad \|\nabla q\|_{H^{-1}(\Omega)} = \sup_{v \in H_0^1(\Omega)} \frac{\int_{\Omega} q \nabla \cdot v}{\|v\|_{H_0^1(\Omega)^N}}.$$

9.8 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.

$$\begin{cases} -\mu \Delta u + \nabla p = 0 \\ \nabla \cdot u = 0, \end{cases}$$

Le domaine est défini par

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

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

$$u(x, y, z) = U \left(1 - \frac{r^2}{a^2}\right) \vec{e}_z, \quad p(x, y, z) = -4 \frac{\mu U}{a^2} (z - z_0), \quad (9.18)$$

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 :

$$Q = U \pi \frac{a^2}{2} = \frac{\pi}{8} \frac{a^4}{\mu L} (P_{in} - P_{out}). \quad (9.19)$$

Cette relation s'appelle la *Loi de Poiseuille*, et s'écrit en général¹⁰

$$P_{in} - P_{out} = RQ, \quad (9.20)$$

avec

$$R = \frac{8\mu}{\pi} \frac{L}{a^4}. \quad (9.21)$$

10. 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.

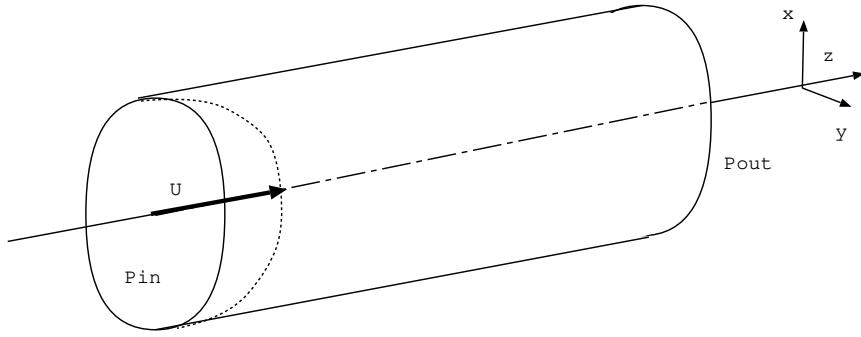


FIGURE 9.3 – Écoulement de Poiseuille

La résistance visqueuse s'exprime en Pa s m^{-3} , Les forces de viscosité dissipent l'énergie au taux¹¹

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

Un calcul direct permet d'établir que $\mathcal{P} = RQ^2$ (on reconnaîtrait 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

$$\left\{ \begin{array}{ll} -\mu \Delta u + \nabla p &= 0 & \text{in } \Omega, \\ \nabla \cdot u &= 0 & \text{in } \Omega, \\ u &= 0 & \text{on } \Gamma_w, \\ \mu \nabla u \cdot n - p n &= -P_{in} n & \text{on } \Gamma_{in}, \\ \mu \nabla u \cdot n - p n &= -P_{out} n & \text{on } \Gamma_{out}. \end{array} \right. \quad (9.22)$$

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 9.23. (Résistance d'un domaine (Stokes))

Soit u le champ de vitesse solution de (9.22). Le flux Q est défini comme

$$Q = - \int_{\Gamma_{in}} u \cdot n = \int_{\Gamma_{out}} u \cdot n. \quad (9.23)$$

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

$$P_{in} - P_{out} = RQ. \quad (9.24)$$

11. L'expression devrait être

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

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.

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

Proposition 9.24. 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 9.23) s'exprime alors

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

9.9 Ecoulement autour d'une sphère

On peut décrire explicitement le champ de vitesse correspondant à l'écoulement d'un fluide visqueux en milieu infini autour d'une sphère fixe. On considère une sphère de rayon a centrée à l'origine d'un repère $(O, \vec{e}_x, \vec{e}_y, \vec{e}_z)$, et l'on se place dans le système de coordonnées sphériques (O, r, θ, ϕ) : pour tout point de \mathbb{R}^3 représenté par son rayon vecteur $\mathbf{r} = (x, y, z)$, r est le module de \mathbf{r} , θ est l'angle que fait $(x, y, 0)$ avec \vec{e}_x (longitude, comprise entre 0 et 2π), et Φ est l'angle que fait \mathbf{r} avec l'axe des z (latitude, comprise entre 0 et π). On suppose que la vitesse à l'infini est égale à $U\vec{e}_z$. Les vecteurs unitaires associés à ce système de coordonnées qui vont nous servir à exprimer le champ des vitesses sont

$$\vec{e}_r = \frac{\mathbf{r}}{r}, \quad \vec{e}_{\Phi} = \frac{1}{r} \frac{\partial \mathbf{r}}{\partial \Phi}.$$

On peut vérifier que tout couple (u, p) défini par

$$u = u_r \vec{e}_r + u_{\Phi} \vec{e}_{\Phi}, \quad u_r = U \cos \Phi \left(1 - \frac{3a}{2r} + \frac{a^3}{2r^3} \right), \quad u_{\Phi} = -U \sin \Phi \left(1 - \frac{3a}{4r} - \frac{a^3}{4r^3} \right),$$

$$p - p_0 = -\frac{3}{2} \frac{\mu U a}{r^2} \cos \Phi,$$

où p_0 est une constante arbitraire, est solution des équations de Stokes dans le domaine $\mathbb{R}^3 \setminus B(0, a)$, avec des conditions d'adhérence ($u = 0$) sur la sphère $\{r = a\}$, et des conditions à l'infini

$$\lim_{r \rightarrow +\infty} u(r, \theta, \Phi) = U\vec{e}_z.$$

On en déduit l'expression du module de la force exercée par le fluide sur la sphère, appelée *loi de Fáxen*

$$F = 6\pi\mu a U. \tag{9.25}$$

De façon plus générale, si une particule sphérique est en mouvement à la vitesse v dans un fluide dont la vitesse locale (voir remarque 9.25 ci-dessous concernant cette question de *localité*) est égale à U , la force exercé par le fluide sur la particule s'écrira $6\pi\mu a(U - v)$.

Remarque 9.25. La situation que nous considérons ici comporte 3 échelles en espace :

1. microscopique : diamètre de la particule sphérique
2. mesoscopique : zone du fluide au voisinage de la particule, petite par rapport à l'échelle macroscopique
3. macroscopique, taille caractéristique de l'écoulement. On pourra considérer ici qu'il s'agit d'une distance correspondant à une variation significative du champ de vitesse, de l'ordre de U/G , où U est l'ordre de grandeur du module de la vitesse, et G l'ordre de grandeur de la norme du gradient.

Nous supposons que ces échelles sont bien séparées $1 \ll 2 \ll 3$. Cette hiérarchie des échelles exprime que la particule doit voir la zone mésoscopique comme infinie ($1 \ll 2$), de façon à ce que le développement ci-dessus puisse être considéré comme valide. Par ailleurs, toujours pour assurer la pertinence de la solution analytique, le fluide doit être considéré comme étant en translation uniforme sur la zone mésoscopique, ce qui nécessite $2 \ll 3$.

Deuxième partie

Fondamentaux

Chapitre 10

Graphes

Sommaire

10.1 Définitions	223
10.2 Laplacien(s)	229
10.3 Exercices	231

10.1 Définitions

Definition 10.1. (Graphe orienté)

On appelle graphe orienté un couple (V, E) où V est un ensemble de sommets, et $E \subset V \times V$ l'ensemble des arêtes (on parle aussi d'*arcs*) du graphe.

Sauf mention contraire, les graphes considérés dans la suite de ce cours sont *finis*, i.e. V est un ensemble fini.

Noter qu'un graphe au sens défini ci-dessus n'est autre qu'une *relation*¹ définie sur un ensemble (appelé ici ensemble de sommets).

Les graphes orientés permettent de décrire des liens *non symétriques* entre des entités (les sommets du graphes). Il peut s'agir par exemple de graphes d'influence : les sommets du graphes sont des personnes (souvent appelées *agents*) dans ce contexte, et l'on écrit $(x, y) \in E$, ou $x \rightarrow y$, si x est influencé² par y . La non symétrie peut aussi exprimer que le transport d'une certaine substance

1. Une relation sur un ensemble X est la donnée d'une partie R de $X \times X$, dénotée par le symbole \mathcal{R} selon la convention

$$(x, x') \in R \iff x \mathcal{R} x'.$$

2. Nous ferons ce choix d'orientation, qui peut sembler surprenant puisqu'il conduit à considérer que l'information remonte le sens des flèches, comme un saumon remonte le courant de sa rivière natale. Diverses raisons à ce choix :

1. Pour les réseaux de personnes, la flèche représente le regard ou l'attention portée à un autre agent : x pointe son attention visuelle (ou autre) vers y . L'agent x peut décider à tout moment de ne plus suivre y , auquel cas la flèche disparaît par décision de sa base (son point d'ancrage).
2. Dans ce même contexte, il est possible que x suive y mais que y n'exerce en fait aucune influence, ou une influence négative sur tel ou tel paramètre. Le choix étant fait de matérialiser l'attention portée et pas la transmission effective d'information ou d'opinion, ces situations ne posent pas de problème.
3. Pour les processus de diffusion d'opinion sur un réseau social, on verra qu'il existe une marche aléatoire associée canoniquement au modèle de propagation, avec des mouvements qui se font précisément dans le sens des flèches (et donc dans le sens opposé à la propagation de l'opinion elle-même) tel que nous l'avons choisi.
4. Comme on le verra plus loin (voir autour de la définition 10.23 ci-après), si chaque individu n'est influencé que par un seul autre, on peut associer à E une application F de V dans V . Si l'on part d'une opinion représentée par un champ scalaire $u \in \mathbb{R}^V$, l'opinion transférée par notre réseau d'influence s'écrira, selon la convention

entre x et y ne peut se faire que dans un sens, par exemple si V est l'ensemble des points de jonctions entre les rues d'une ville, les arc sont ces rues, dont certaines peuvent être à sens unique. Dans le cas d'une route à double sens entre x et y , on aura 2 arcs (x, y) et (y, x) .

Definition 10.2. (Graphe non orienté)

On appelle graphe non orienté un couple (V, E) où V est un ensemble de sommets, et

$$E \subset \{\{x, y\}, x, y \in V\},$$

l'ensemble des arêtes du graphe. On peut considérer E comme une partie de $V \times V$ quotientée par la relation d'équivalence qui consiste à identifier (x, y) et (y, x) , pour tous $x, y \in V$. Nous utiliserons l'abus d'écriture consistant à désigner par $(x, y) \in E$ une arête non orientée, sans préciser qu'il s'agit d'un élément de $V \times V$ quotienté par les transpositions. Comme nous le préciserons, dans le cas où l'on somme des quantités sur l'ensemble des arêtes, on ne comptera qu'une fois chaque arête. Deux sommets reliés par une arête sont dits *adjacents*.

Les graphes non orientés sont adaptés pour décrire des liens *symétriques* entre des entités (les sommets du graphes). Dans le contexte d'une population d'agents, il peut simplement s'agir d'une relation de proximité : $(x, y) \in E$ si x et y se connaissent, se sont déjà croisés, ou font partie de la même famille.... Dans un contexte de transport de substance, une arête non orientée correspond à un canal de transmission qui laisse passer de la matière dans un sens ou dans l'autre, comme un fil électrique qui peut être traversé par un courant dans un sens ou dans l'autre, ou un tuyau au travers duquel un fluide peut s'écouler dans les deux sens.

Nous avons pris le parti d'exclure le cas où il peut exister plusieurs arêtes reliant 2 points. Plus précisément, dans le cas d'un graphe orienté, il y a au plus un arc allant de x à y pour tous sommets x et y distincts. Pour un graphe non orienté, pour tous $x \neq y$, il y a au plus une arête reliant x et y (même si nous considérons (x, y) et (y, x) sont des arêtes, elles sont identifiées). Nous ne considérons donc, sauf avis contraire, aucun *multigraphe*³. Si par ailleurs le graphe considéré ne contient aucune boucle ($(x, x) \notin E$ pour tout $x \in V$), on parlera de graphe *simple*. Tous les graphes non orientés que nous considérerons seront simples. Dans le cas de graphes orientés, il peut être en revanche pertinent de considérer des boucles $x \rightarrow x$.

Definition 10.3. Un graphe non orienté est dit *complet* si tous les sommets sont reliés entre eux deux à deux. Un graphe orienté est complet si $E = V \times V$. On qualifiera de *vide* un graphe sans arête.

Definition 10.4. (Graphe biparti)

Un graphe non orienté $G = (V, E)$ est dit *biparti* si V admet une partition $V = V_1 \cup V_2$ tel que chaque arête du graphe contienne une extrémité dans V_1 et une extrémité dans V_2 .

La figure 10.1 (droite) donne un exemple de graphe biparti. Ces objets sont souvent utilisés pour représenter des relations entre des entités de natures différentes, par exemple V_1 est un ensemble de personnes, V_2 un ensemble de films, et $x_1 \sim x_2$ si x_1 a vu x_2 .

Definition 10.5. (Graphe pondéré)

On appelle graphe pondéré un triplet (V, E, w) où (V, E) est un graphe, orienté ou pas (au sens des définitions précédentes), et $w \in \mathbb{R}_+^E$ un ensemble de poids afférents aux arêtes. De façon plus générale et formelle, on sera amené à munir certains graphes d'une application de E dans un ensemble X , qui peut être \mathbb{R}_+ dans le cas de longueurs, de résistances, conductances, etc ... L'ensemble X peut aussi être discret, par exemple $\{+, -\}$ pour représenter des influences dans un système différentiel (voir exemple ?? ci-après).

choisie, $u' = u \circ F$ (pullback de u par F).

3. On parle de multigraphe lorsqu'il peut exister plusieurs arcs (ou arêtes dans le cas non orienté) entre deux points x et y . On peut définir un multigraphe G comme un triplet (V, E, Φ) , où E peut peut-être vu comme l'ensemble des labels des arêtes, et Φ est une application (pas forcément injective) de E dans $V \times V$. Dans ce contexte on note en général $e \in E \longmapsto (e, e_+) \in V \times V$. On parle d'arête multiple lorsque le même couple de sommets est associé à plusieurs arêtes. Nous utiliserons peu ce formalisme ici, mais il peut être utile par exemple pour exprimer les dépendances dans un modèle multicompartment (voir figure ??).

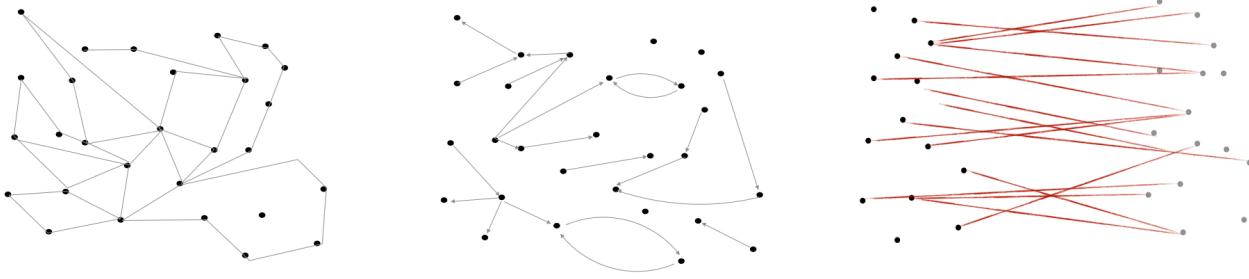


FIGURE 10.1 – Graphes non orienté, orienté, biparti

La notion enrichie de graphe pondéré permet d'affecter à chaque arête une valeur, qui peut prendre des significations très différentes selon le contexte. Dans le cas réseaux électriques par exemple (voir chapitre 2), on considère un ensemble d'arêtes non orientées auxquelles on attribue des *résistances* (ou leurs inverses, les conductances). Dans le cas d'un réseau social (voir chapitre ??), le graphe (orienté) encode les relations d'influence entre les personnes, auxquelles on pourra affecter des poids permettant de quantifier la force de cette influence.

Remarque 10.6. Le terme de “poids” est ambigu, et de fait prend des sens très différents, voire contraires, selon les contextes. Dans un contexte d’algorithme par exemple (comme la recherche de plus courts chemins), les poids correspondent en général aux longueurs entre les arêtes, donc un poids important correspond à des points éloignés, il est alors naturel de considérer qu’un poids infini équivaut à la disparition de l’arête. Mais dans d’autres contextes, comme les réseaux électriques, les poids correspondent parfois aux conductances, une conductance forte correspondant d’une certaine manière à des points proches. Un poids - conductance infini correspondrait alors non pas à la disparition de l’arête, mais à l’identification des deux points la constituant. De même dans le contexte de réseaux de contacts, si le poids quantifie l’intensité des contacts entre deux sommets-individus, un poids fort correspond à une grande proximité.

Definition 10.7. (Ordre)

L’ordre d’un graphe $G = (V, E)$ est le cardinal de V .

Definition 10.8. (Degré, dégré sortant, dégré rentrant)

Soit $G = (V, E)$ un graphe. On appelle degré de $x \in V$, que l’on note d_x , le nombre d’arêtes dont x est une extrémité. Si le graphe est orienté, on appelle d_x^- le degré entrant, qui est le nombre d’arcs dont x est le point d’arrivée, et d_x^+ le degré sortant, nombre d’arcs dont x est le point de départ.

Definition 10.9. (Graphe régulier)

Un graphe non orienté $G = (V, E)$ est dit régulier s’il est simple et si tous les sommets ont le même degré. On parlera d’un graphe p -régulier si ce degré commun est $p \in \mathbb{N}$.

Definition 10.10. (Graphe dual)

Soit $G = (V, E)$ un graphe non orienté. On définit son graphe dual $G' = (V', E')$ par $V' = E$, avec la règle que les arêtes (vues ici comme sommets du graphe dual) e_1 et e_2 sont en contact si elles partagent un même sommet de V :

$$(e_1, e_2) = ((x_1, y_1), (x_2, y_2)) \in E' \iff \{x_1, y_1\} \cap \{x_2, y_2\} \neq \emptyset.$$

Definition 10.11. (Isomorphisme / automorphisme de graphe)

On appelle isomorphisme entre deux graphes $G = (V, E)$ et $G' = (V', E')$ (orientés ou non) une bijection Φ de V dans V' telle que

$$(\Phi(x), \Phi(y)) \in E' \iff (x, y) \in E.$$

Si $G' = G$, on parlera d’automorphisme. L’ensemble de ces automorphismes est de façon évidente un sous-groupe du groupe symétrique associé à V , on parlera du groupe des automorphismes du graphe. Si le groupe des automorphismes est réduit à l’identité, on dit que le graphe est *asymétrique*.

On notera que le transport par un isomorphisme préserve les degrés.

Exercice 10.1. Combien (à isomorphisme près) y a-t-il de graphes orientés à N sommets ? de graphes non orientés à N sommets ?

Exercice 10.2. (Exemples)

- Donner un exemple de graphe non orienté à N sommets isomorphe à son graphe dual.
- Préciser le graphe dual du graphe complet non orienté à N sommets.
- Que peut-on dire d'un graphe dont le dual est vide ?
- Donner un exemple de graphe tel que la suite des ordres des graphes duals soit croissante, et un exemple de graphe tel que cette suite soit décroissante, et un exemple pour lequel la suite des ordres est constante.

Exercice 10.3. On considère un ensemble V fini, muni d'une loi de composition interne $(x, z) \mapsto x \star y = z$. Comment peut-on représenter cette loi sous la forme d'un graphe ? (On distinguera le cas d'une loi commutative d'un loi non commutative, et l'on précisera la nature des graphes correspondants.)

Remarque 10.12. On peut montrer⁴ que tout groupe fini est isomorphe au groupe des automorphismes d'un graphe non orienté fini. On peut même choisir ce graphe 3-régulier.

Definition 10.13. (Chemin / circuit, chaîne / cycle)

Dans un graphe orienté (V, E) , un *chemin* est une suite finie de sommets connectés :

$$x_0, x_1, \dots, x_n \text{ avec } (x_i, x_{i+1}) \in E \quad \forall i = 0, \dots, n-1, \quad n \geq 1.$$

Quand il existe un chemin de x à y , on écrira $x - \cdot \rightarrow y$.

Un chemin est dit *simple* s'il ne passe pas deux fois par la même arête.

Un chemin est dit *élémentaire* s'il ne passe pas deux fois par le même sommet.

Un *circuit* est un chemin qui termine où il commence.

Une *boucle* est un circuit constitué d'un arc unique (x, x) .

Dans le cas d'un graphe non orienté, on parlera de *chaîne* reliant deux points (on notera $x \sim \cdot \sim y$ si une telle chaîne existe), et de *cycle* si les deux extrémités sont confondues. Les attributs simple et élémentaire se définissent comme précédemment.

Important : Dans la pratique, et en particulier dans ce polycopié, on utilise souvent le terme de *cycle* même quand le graphe est orienté, et le terme de chemin même s'il est non orienté, notamment pour parler de *plus court chemin* (voir ci-dessous).

La *longueur* d'un chemin est le nombre des arêtes qui le constituent ou, si le graphe est pondéré, la somme des poids des arêtes⁵.

Definition 10.14. (Connexité, forte connexité)

On dit qu'un graphe non orienté est *connexe* s'il existe un chemin reliant tout x à tout y . On dira de même d'un graphe orienté si le graphe non orienté associé (on oublie l'orientation des arêtes) est connexe. On dira qu'un graphe orienté est *fortement connexe* s'il existe un chemin de tout x vers tout y , i.e. $x - \cdot \rightarrow y$ pour tous $x, y \in V$.

4. http://www.numdam.org/item/CM_1939_6_239_0.pdf (en allemand).

5. Le terme de *poids* est ambigu, et prend des significations différentes selon les modèles considérés. On prendra en particulier garde au fait que, dans certaines situations comme des réseaux de contacts où le poids correspond au nombre de contacts entre individus, un poids fort correspond à une grande proximité, auquel cas il est naturel de définir la longueur d'une arête comme une fonction *décroissante* du poids. Dans un contexte électrique, si les poids sont les conductances, il sera naturel (voir chapitre 2) de prendre pour longueur l'inverse du poids (qui est alors la résistance), mais dans d'autres contextes, d'autres fonctions peuvent s'avérer plus pertinentes. Il est néanmoins courant, en particulier dans la littérature informatique recherche opérationnelle, d'identifier les poids à des longueurs comme dans la définition proposée.

Definition 10.15. (Composantes connexes)

Soit (V, E) un graphe non orienté. On définit la relation d'équivalence \mathcal{R} comme suit : $x \mathcal{R} y$ si $x = y$ ou s'il existe un chemin entre x et y . Les classes d'équivalence sont appelées *composantes connexes* du graphe.

Definition 10.16. (Graphe acyclique)

On dit qu'un graphe orienté G est acyclique s'il ne contient aucun cycle (ou circuit, à strictement parler). On appelle un tel graphe une *hiérarchie* (voir proposition ci-dessous).

On dit qu'un graphe non orienté G est acyclique s'il ne contient aucun cycle *simple*⁶.

Proposition 10.17. (Ordre partiel sur un graphe acyclique)

Soit $G = (V, E)$ un graphe orienté acyclique. On pose $x \leq y$ si $x = y$, ou si $x - \cdot \rightarrow y$, i.e. s'il existe un chemin de x à y . Cette relation définit un ordre partiel sur V .

Démonstration. On a $x \leq x$. Si $x \leq y$, $y \leq x$, et $x \neq y$, alors la concaténation des deux chemins $x \rightarrow y$ et $y \rightarrow x$ forme un cycle, ce qui est exclu, donc nécessairement $x = y$. Enfin, si $x \leq y$, et $y \leq z$, la concaténation des deux chemins est un chemin de x à z , donc $x \leq z$. \square

Proposition 10.18. (Tri topologique)

Soit $G = (V, E)$ un graphe orienté acyclique, et ' \leq' l'ordre partiel associé. Il existe une numérotation de points de V compatible avec la structure de graphe : i.e. il existe une bijection Φ de $I_N = \{1, \dots, N\}$ vers V , avec $N = \sharp(V)$, telle que $x \leq y$ implique $\Phi^{-1}(x) \leq \Phi^{-1}(y)$.

Démonstration. Montrons tout d'abord qu'il existe un élément maximal pour l'ordre partiel induit (voir proposition 10.17). On part d'un sommet $x_0 \in V$ arbitraire, et l'on considère un chemin

$$x_0 \rightarrow x_1 \rightarrow \dots$$

Comme V est fini, on finit par aboutir à un point x_k qui ne peut être connecté qu'à un point déjà visité, ce qui est exclu par acyclicité, le point x_k n'est donc en fait connecté à aucun point, il est donc maximal pour \leq . L'indice $N = \sharp(V)$ est affecté à cet élément maximal x . On considère maintenant le graphe $V \setminus \{x\}$, avec les arêtes qui ne contiennent pas x , et l'on applique la même démarche à ce graphe, pour construire l'élément d'indice $N - 1$. Par construction cette numérotation est compatible avec l'ordre partiel initial. \square

Definition 10.19. (Arbre)

Un graphe non orienté acyclique et connexe est appelé un *arbre*. On parlera de forêt si l'on n'a pas l'hypothèse de connexité, chaque composante connexe étant alors un arbre. Une forêt peut n'être constituée que d'un arbre. Les sommets de degré 1 d'un arbre sont dits *pendants*. Une arête est dite *terminale* si l'une de ses extrémités est pendante. On dira qu'un arbre est enraciné si l'on a affecté à un sommet particulier le statut de *racine*, sans condition particulière sur ce sommet singularisé (il peut en particulier être pendante ou pas).

Definition 10.20. (Arbre)

Un arbre couvrant pour le graphe non orienté G est un arbre qui contient tous les sommets de G .

Definition 10.21. (Métrique sur un graphe non orienté)

Soit $G = (V, E)$ un graphe non orienté connexe. La structure de graphe induit sur V une métrique canonique d , telle que $d(x, y)$ est la longueur du plus court chemin reliant x et y . Si le graphe est pondéré, et que l'on assimile les poids à des longueurs des arêtes, on remplace simplement la longueur des chemins par la somme des longueurs d'arêtes. Cette métrique permet de définir le diamètre d'un graphe

$$\text{diam}(G) = \max_{x, y \in V} d(x, y),$$

ainsi que toutes les notions usuelles de boule, sphère, distance d'un point à un ensemble de points,

6. Si l'on ne précisait pas 'simple', le seul graphe rentrant dans la définition serait le graphe vide. En effet, si (x, y) est une arête d'un graphe non orienté, alors $x \longleftrightarrow y \longleftrightarrow x$ est un cycle.

Remarque 10.22. (Small world)

On qualifie de *small world* un graphe tel que, même si son nombre de sommets est grand, deux points du réseau sont toujours à petite distance l'un de l'autre. Pour le graphe constitué des habitants de notre planète, si l'on considère que deux points sont connectés quand les personnes correspondantes ont eu un contact physique, il a été estimé⁷ que le diamètre est de l'ordre de 6.

Definition 10.23. (Matrice d'adjacence, polynôme caractéristique, spectre)

On considère un graphe $G = (V, E)$, et l'on se donne une indexation x_1, \dots, x_N , de V . On peut associer à ce graphe indexé une matrice A dite d'adjacence définie par

$$A = (a_{ij}), \quad a_{ij} = 1 \text{ si } (x_i, x_j) \in E, \quad a_{ij} = 0 \text{ si } (x_i, x_j) \notin E,$$

avec la convention que, si le graphe est non orienté, on considère que pour toute arête (x, y) et (y, x) sont dans V . Selon cette définition, la matrice dépend de l'indexation choisie. On pourra être amené à utiliser la notation $A = (a_{xy})_{x,y \in V}$, qui désigne d'une certaine manière la matrice *indépendamment de l'indexation choisie*⁸. Si le graphe est non orienté, on considère qu'à la fois (x, y) et (y, x) sont dans E , de telle sorte que la matrice d'adjacence est symétrique.

Dans le cas d'un graphe biparti (non orienté par défaut) sur $V_1 \cup V_2$, on assimilera la matrice d'adjacence au bloc rectangulaire en haut à droite de la matrice définie ci-dessus, qui a $n_1 = |V_1|$ lignes et $n_2 = |V_2|$ colonnes.

Le *polynôme caractéristique* d'un graphe est défini comme le polynôme caractéristique de sa matrice d'adjacence. Le *spectre* d'un graphe est l'ensemble des racines (avec multiplicités) dans \mathbb{C} de ce polynôme caractéristique⁹.

Exercice 10.4. Préciser le spectre du graphe (non orienté) complet d'ordre N (avec et sans boucles), des graphes cycliques orientés et non orientés d'ordre N .

Remarque 10.24. (Forte connexité et irréductibilité)

Un graphe orienté est fortement connexe (i.e. il existe un chemin de tout x vers tout y) si et seulement si sa matrice d'adjacence est *irréductible*, c'est à dire si elle n'est pas semblable par permutation à une matrice diagonale par blocs avec au moins deux blocs diagonaux

Proposition 10.25. Soit $G = (V, E)$ un graphe (orienté ou non), et A la matrice d'adjacence associée. Pour $p \geq 1$, on note a_{xy}^p (où p est ici un indice et pas une puissance) les éléments de la matrice A^p (il s'agit bien ici d'une puissance). Le nombre a_{xy}^p est le nombre de chemins de longueur p de x vers y .

Démonstration. Pour $p = 2$, on a

$$a_{xy}^2 = \sum_{z \in V} a_{xz} a_{zy}.$$

Il s'agit donc d'une somme de termes nuls, ou égaux ou à 1 si $x \rightarrow z \rightarrow y$, c'est à dire s'il existe un chemin de longueur 2 passant par z . La propriété générale se montre par récurrence sur p . \square

Corollaire 10.26. Soit $G = (V, E)$ un graphe (orienté ou non), et A la matrice d'adjacence associée. Pour $p \geq 1$, on note comme précédemment a_{xy}^p les éléments de la matrice A^p (p est un indice dans le premier cas, une puissance dans le second). Le graphe est connexe (ou fortement connexe s'il s'agit d'un graphe orienté, voir définition 10.14) si et seulement si

$$\forall x, y, \exists p \geq 1, a_{xy}^p > 0.$$

7. Travers, J. & Milgram, S. An Experimental Study of the Small World Problem. *Sociometry* 32, 425 (1969). <https://www.jstor.org/stable/2786545>

8. Il s'agit plus précisément d'une classe d'équivalence de l'espace des matrices pour la relation d'équivalence

$$ARA' \iff \exists \sigma \in S^N, A' = U_\sigma^{-1} A U_\sigma,$$

où $\sigma \in S^N$ est le groupe symétrique (bijections de l'ensemble à N éléments), et U_σ la matrice de permutation associée.

9. On appelle parfois spectre d'un graphe l'ensemble des racine du polynôme caractéristique du laplacien combinatoire (voir définition 10.28, page 229).

Remarque 10.27. On peut associer certains graphes orientés à des applications de V dans V . Plus précisément, si $G = (V, E)$ est un graphe orienté, avec $d^+ \equiv 1$ (chaque sommet n'est le point de départ que d'une flèche et une seule), on peut associer à E une application F de V dans V qui est telle que $(x, F(x)) \in E$ pour tout x . Si de plus tous les degrés entrants sont eux-mêmes égaux à 1, cette application est une bijection. De façon plus générale, sans aucune hypothèse sur les degrés, on peut associer de façon univoque à tout graphe orienté G une application F dite *multivaluée*, qui à tout point de V associe une partie de V , définie par

$$\begin{aligned} F : V &\longrightarrow 2^V \\ x &\longmapsto F(x) = \{y \in V, (x, y) \in E\}. \end{aligned}$$

Dans le cas d'un graphe biparti (définition 10.4), on peut voir l'ensemble des arêtes comme représentant une multi-application F_{12} de V_1 vers V_2 , ou une multi-application F_{21} de V_2 vers V_1 . Noter que certains points de V_1 peuvent être isolés, leur image par F_{12} est donc \emptyset .

10.2 Laplacien(s)

Il existe plusieurs manière de définir, sur un graphe, un opérateur de type *Laplacien*¹⁰. Ces opérateurs jouent un rôle central dans les développements théoriques autour des graphes, ainsi que dans les aspects de modélisation. Nous présentons ici quelques éléments, qui sont développés dans les autres chapitres.

Definition 10.28. (Laplacien associé à un graphe non orienté)

Soit $G = (V, E, K)$ un graphe non orienté pondéré¹¹, sans points isolés.

On définit le laplacien comme l'endomorphisme linéaire

$$L : u \in \mathbb{R}^V \longmapsto Lu \in \mathbb{R}^V, \quad Lu(x) = \sum_{y \sim x} K_{xy} (u_x - u_y). \quad (10.1)$$

Si le graphe n'est pas pondéré, où si l'on souhaite définir un opérateur qui ne prenne pas en compte les poids, on utilise la même définition en considérant que toutes les arêtes ont un poids unitaire :

$$L : u \in \mathbb{R}^V \longmapsto Lu \in \mathbb{R}^V, \quad Lu(x) = \sum_{y \sim x} (u_x - u_y).$$

On parle alors de *Laplacien combinatoire*. Noter que, si l'on note D l'opérateur diagonal dont les coefficients sont les degrés des sommets, et A la matrice d'adjacence (définition 10.23), on a $L = D - A$.

Definition 10.29. (Champ harmonique)

Soit $G = (V, E)$ un graphe non orienté sans point isolé, et L le Laplacien associé (définition 10.28). Les champs $u \in \mathbb{R}^V$ qui annulent L sont dits *harmoniques* sur G . Pour un tel champ, la valeur en un point de degré supérieur ou égal à 1 est la moyenne pondérée des valeurs aux points voisins :

$$u(x) = \frac{1}{\sum_y K_{xy}} \sum_{y \sim x} K_{xy} u_y. \quad (10.2)$$

ou la valeur moyenne simple dans le cas du laplacien combinatoire

$$u_x = \frac{1}{d_x} \sum_{y \sim x} u_y. \quad (10.3)$$

10. Conformément à l'usage, nous appelons Laplacien des opérateurs qui sont en fait des pendants discrets de $-\Delta$, i.e. l'*opposé* Laplacien continu.

11. On prendra garde au fait que cette définition conduit à un objet pertinent en termes de modélisation si les poids sont considérés comme quantifiant une proximité entre les sommets, i.e. un coefficient K_{xy} grand correspond à une "distance" petite (voir remarque 10.6). Dans le cas d'un réseau électrique par exemple, le Laplacien aura un sens clair si les poids sont les *conductances*, une conductance petite entre 2 points indiquant que les valeurs des potentiels aux sommets sont peu corrélées, jusqu'à la situation limite d'une conductance nulle qui équivaut à la disparition de l'arête.

Proposition 10.30. La matrice associée au Laplacien défini ci-dessus est symétrique, semie-définie positive, i.e.

$$Lu \cdot u \geq 0 \quad \forall u \in \mathbb{R}^V.$$

Toutes ses valeurs propres sont donc réelles positives ou nulles. La dimension du noyau de L est égale au nombre de composantes connexes du graphe.

Démonstration. La matrice L est symétrique, ses valeurs propres sont donc réelles. Elle est par ailleurs à diagonale dominante, ces valeurs propres sont donc positives ou nulles (voir proposition 19.27, page 395). Soit maintenant $x \in V$ qui réalise le maximum de u sur la composante connexe \bar{x} à laquelle appartient x . d'après la relation (10.3), on a $u(y) = u(x)$ pour tout y connecté à x . On ainsi de proche en proche que u est constante sur la composante connexe, d'où l'on déduit que u est constant sur \bar{x} . \square

Remarque 10.31. Dans le cas d'un réseau cartésien, on retrouve le Laplacien obtenu par discrétilisation de l'opérateur continu par différences finies. Ainsi pour un graphe obtenu par subdivision d'un intervalle réel, on trouve (à facteur multiplicatif près) la matrice du Laplacien discret, avec des 2 sur la diagonale et des -1 sur les premières extra-diagonales (voir équation 19.16, 395), sauf en première et dernière positions de la diagonale, pour lesquelles l'élément vaut 1.

Proposition 10.32. Soit L le Laplacien associé à un graphe non orienté pondéré $G = (V, E, K)$ (expression (10.1)). Pour tout $u \in \mathbb{R}^V$, Lu est le gradient en u de la fonctionnelle d'énergie

$$J(v) = \frac{1}{2} \sum_{(x,y) \in E} K_{xy} (v_x - v_y)^2,$$

où, comme il a été précisé en préambule, on ne compte qu'une fois dans la somme chaque arête non orientée.

Laplacien non symétrique

Il n'y a pas unanimité dans la littérature sur ce qu'il est permis d'appeler laplacien. Nous ferons le choix de donner un sens très général à ce terme, y compris dans le cas de graphes pondérés orientés.

Definition 10.33. (Laplacien non symétrique)

Soit (V, E, K) un graphe orienté pondéré. Plus précisément, on se donne une collection de poids K_{xy} positifs, sans propriété de symétrie particulière, avec

$$E = \text{supp}(K) = \{(x, y) \mid K_{xy} > 0\}.$$

On introduit la matrice diagonale $D = \text{diag}(K_x \bar{y})$, avec $K_{x\bar{y}} = \sum_y K_{xy}$. On appelle laplacien non symétrique associé au graphe l'opérateur

$$L = D - K, \quad (Lu)_x = \sum_{x \rightarrow y} K_{xy} (u_x - u_y).$$

10.3 Exercices

Exercice 10.5. (Graphes acycliques)

1) Soit $G = (V, E)$ un graphe (fini) orienté acyclique connexe (au sens où le graphe non orienté associé est connexe), tel que $\sharp(V) \geq 2$. Montrer qu'il existe au moins un sommet x de degré sortant nul, et au moins un élément y de degré rentrant nul. Comment peut-on qualifier ces points vis-à-vis de l'ordre partiel canonique associé au graphe ?

2) Soit $G = (V, E)$ un graphe (fini) non orienté acyclique connexe, tel que $\sharp(V) \geq 2$. Montrer qu'il existe au moins deux sommets de degré 1.

3) Les propriétés précédentes restent-elles vraies dans le cas de graphes infinis ?

4) Soit G un graphe orienté, et A la matrice d'adjacence associée. Quelle propriété de la suite (A^p) est équivalente à l'acyclicité du graphe ? Si le graphe est bien acyclique, que peut-on dire sur les lignes ou colonnes associées aux points particuliers introduits dans la question 1 ?

Exercice 10.6. a) Montrer que l'ordre d'un graphe 3-régulier est nécessairement pair.

b) Pour tout $p \geq 2$, montrer qu'il existe un graphe 3-régulier d'ordre $2p$.

Exercice 10.7. Soit G un graphe non orienté simple d'ordre $2p$, tel que le degré de chaque sommet est supérieur ou égal à p . Montrer que G est connexe.

Exercice 10.8. (Tournoi)

On dit qu'un graphe orienté simple G est un *tournoi* si entre deux sommets il y a exactement un arc : pour tout x, y distincts, $(x, y) \in E \iff (y, x) \notin E$. Quand $x \rightarrow y$ (i.e. $(x, y) \in E$), on dira que y domine x . On appelle *roi* un sommet x tel que pour tout $y \neq x$, x domine y , ou alors il existe z tel que x domine z qui domine y .

a) Montrer que dans tout tournoi il existe un roi.

b) Ce roi peut-il être unique ? Est-il nécessairement unique ?

c) Exprimer la propriété en langage commun.

Exercice 10.9. Soit X un ensemble de cardinal fini $N \geq 2$, et $\mathcal{P} = 2^X$ l'ensemble de ses parties (y compris la partie vide). On définit un graphe orienté (V, E) par $V = \mathcal{P}$ et, pour tous A, B dans \mathcal{P} , $A \rightarrow B$ si A est strictement inclus dans B et $\sharp(B) = \sharp(A) + 1$. On note (V, \bar{E}) le graphe non orienté associé.

1) a) Quel est l'ordre partiel (défini par la proposition 10.17) associé canoniquement à ce graphe ?

b) Montrer que ce graphe est acyclique, connexe (i.e. (V, \bar{E}) est connexe), mais pas fortement connexe.

c) Représenter graphiquement le graphe pour $N = 2$, et $N = 3$.

2) a) Montrer qu'il suffit de rajouter une arête (un arc) à (V, E) pour qu'il devienne fortement connexe.

b) Pour ce nouveau graphe maintenant fortement connexe, si l'on se donne deux parties A et B de V , préciser la longueur du plus court chemin de A vers B .

3) Préciser l'ensemble des éléments maximaux, ainsi que l'ensemble des éléments minimaux, pour l'ordre partiel associé à (V, E) ? Répondre aux mêmes questions si l'on retire du graphe ces éléments maximaux (resp. minimaux).

4) Quelle est la distribution des degrés entrants et sortants de (V, E) ? Des degrés de (V, \bar{E}) ?

5) Décrire une procédure pour numérotter les sommets de (V, E) (i.e. les parties de X) de façon compatible avec l'ordre partiel sous-jacent (défini par la proposition 10.18).

Exercice 10.10. (Support de la matrice des itérés)

Soit $G = (V, E)$ un graphe (orienté ou non orienté), et A la matrice d'adjacence associée. On note $A^p = (a_{xy}^p)$ la puissance p -ième de A .

1) Montrer que, si le graphe est “bouclé” ($(x, x) \in E$ pour tout x), alors

$$\text{supp}(A^p) = \{(x, y) \mid a_{x,y}^p \neq 0\},$$

est croissante pour l’inclusion, mais que cette propriété n’est pas vraie en général.

2) On suppose le graphe bouclé. Montrer que la suite $\text{supp}(A^p)$ est strictement croissante pour l’inclusion jusqu’à un certain indice P , puis qu’elle est stationnaire. Comment peut-on interpréter l’entier P ?

Exercice 10.11. (Groupe des automorphismes)

Décrire le groupe des automorphismes des différents graphes non orientés à N sommets, pour N “petit” ($N = 2, 3, \dots$).

Exercice 10.12. (Laplacien et minimisation d’énergie)

On se place sur un graphe non orienté (V, E) , supposé connexe, on se donne $K = (K_{xy}) \in]0, +\infty[^E$, et l’on définit

$$v \in \mathbb{R}^V \longmapsto J(v) = \sum_E K_{xy} |v_y - v_x|^2 \in \mathbb{R}$$

a) Montrer que la fonctionnelle J est convexe. Est-elle strictement convexe ?

b) Écrire la forme bilinéaire $a(\cdot, \cdot)$ associée à cette forme quadratique, ainsi que la matrice (dans la base canonique) correspondante.

c) La forme bilinéaire $a(\cdot, \cdot)$ est-elle un produit scalaire ?

d) On introduit

$$H = \left\{ v \in \mathbb{R}^V \mid \sum v_x = 0 \right\}.$$

Montrer que $a(\cdot, \cdot)$ est un produit scalaire sur $H \times H$.

e) Préciser la différentielle de J en u , ainsi que le gradient pour le produit scalaire canonique. Quel est le gradient de J vue comme fonctionnelle sur H , pour le produit scalaire $a(\cdot, \cdot)$?

Exercice 10.13. (Matrices à diagonale dominante, cercles de Gershgorin)

1) Soit $A = (a_{ij}) \in \mathcal{M}_N(\mathbb{C})$ supposée à diagonale strictement dominante, i.e.

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}| \quad \forall i = 1, \dots, N.$$

Montrer que A est inversible.

2) Soit $A = (a_{ij}) \in \mathcal{M}_N(\mathbb{C})$. On note $\text{Sp}(A)$ l’ensemble des valeurs propres de A , et $D(z, r) \subset \mathbb{C}$ le disque fermé de centre $z \in \mathbb{C}$ et de rayon $r \geq 0$. Montrer que

$$\text{Sp}(A) \subset \bigcup D \left(a_{ii}, \sum_{j \neq i} |a_{ij}| \right).$$

3) On rappelle qu’une matrice $A = (a_{ij}) \in \mathcal{M}_N(\mathbb{R})$ est dite stochastique si

$$a_{ij} \in [0, 1] \quad \forall i, j, \quad \sum_j a_{ij} = 1 \quad \forall i.$$

Montrer que les valeurs propres d’une matrice stochastique sont dans le disque unité de \mathbb{C} .

4) Montrer que les valeurs propres du Laplacien combinatoire associé à un graphe orienté appartiennent à \mathbb{R}^+ .

5) On note L le Laplacien associé à un graphe pondéré ($(K_{xy} \in]0, +\infty[^E)$ non orienté). Montrer que le spectre est inclus dans un segment de \mathbb{R} que l’on précisera.

Exercice 10.14. (Champs harmoniques)

On considère $V = \{0, 1, \dots, N\}$ et le graphe linéaire non orienté associé :

$$0 \longleftrightarrow 1 \longleftrightarrow \dots \longleftrightarrow N.$$

- 1) Écrire la matrice L du Laplacien combinatoire associé à ce graphe. Montrer que, pour tout $k = 0, \dots, N$, le vecteur

$$v^k = (v_j^k)_{0 \leq j \leq N}, \quad v_j^k = \cos\left(k(j + 1/2)\frac{\pi}{N+1}\right)$$

est vecteur propre de L , et préciser son spectre.

- 2) Décrire l'ensemble Λ_\emptyset des champs de \mathbb{R}^V harmoniques sur V , l'ensemble Λ_0 des champs de \mathbb{R}^V harmoniques sur $V \setminus \{0\}$, et l'ensemble $\Lambda_{0,N}$ des champs de \mathbb{R}^V harmoniques sur $V \setminus \{0, N\}$.

- 3) De façon plus générale, soit $\Gamma \subset V$ une partie de V . Décrire l'ensemble Λ_Γ des champs de \mathbb{R}^V harmoniques sur $V \setminus \Gamma$.

Exercice 10.15. On considère un graphe orienté (V, E) , tel que, pour tout x , le degré sortant de x , défini par

$$d_x^+ = \text{Card}\{y, (x, y) \in E\},$$

est non nul. On accepte que le graphe puisse contenir des boucles, i.e. que (x, x) puisse être dans E pour certains x .

On note $\mathcal{P}(V)$ l'ensemble des mesures de probabilité sur V , c'est à dire l'ensemble des champs $(\mu_x)_{x \in V}$ tels que $\mu_x \geq 0$ pour tout x , et

$$\sum_x \mu_x = 1.$$

On définit le *poussé en avant* de μ comme $\nu = P(\mu) \in \mathbb{R}^V$ défini par

$$\nu_y = \sum_{x, x \rightarrow y} \frac{1}{d_x^+} \mu_x \quad \forall y \in V,$$

ou simplement $\nu_y = 0$ si le degré rentrant d_y^- de y est nul.

- 1) Montrer que, pour tout $\mu \in \mathcal{P}$, le poussé en avant de μ est dans $\mathcal{P}(V)$.

- 2) Donner un exemple de graphe (V, E) et de mesure $\mu \in \mathcal{P}(V)$ tels que $P(\mu) = \mu$. Donner un exemple de graphe pour lequel l'opération P modifie la mesure uniforme.

- 3) Montrer que P ne vérifie pas le principe du maximum, en donnant un exemple de graphe (V, E) et de mesure $\mu \in \mathcal{P}(V)$ tels que l'ensemble des valeurs de $P(\mu)$ ne soit pas dans $[\min(\mu), \max(\mu)]$.

Pour $\mu^0 \in \mathcal{P}(V)$ donné, on définit la suite (μ^k) définie par

$$\mu^{k+1} = P(\mu^k).$$

- 4) Soit $p \in \mathbb{N}$ donné. Donner un exemple de graphe (V, E) et de mesure $\mu^0 \in \mathcal{P}(V)$ tels que la suite (μ^k) est p -périodique.

- 5) Que peut on dire de la suite (μ^k) dans le cas où le graphe (V, E) ne contient aucun cycle autre que des boucles (x, x) ?

Pour tout $f \in \mathbb{R}^V$ on définit le *tiré en arrière* de f comme le champ $g = T(f) \in \mathbb{R}^V$ défini par

$$g_x = \frac{1}{d_x^+} \sum_{y, x \rightarrow y} f_y.$$

6) Montrer que T vérifie le principe du maximum, au sens où, si $g = T(f)$, alors g_x est dans $[\min(f), \max(f)]$.

7) Montrer que T n'est pas conservatif, en donnant un exemple de graphe et de champ f tels que, si $g = T(f)$, on ait

$$\sum_x g_x \neq \sum_x f_x.$$

Pour $f^0 \in \mathbb{R}^V$, on définit (comme précédemment pour P) la suite des itérés de T :

$$f^{k+1} = T(f^k).$$

8) Décrire les comportements possibles de la suite f^{k+1} (on pourra illustrer la réponse sur des exemples).

Chapitre 11

Équations différentielles ordinaires

Sommaire

11.1	Théorème de Cauchy–Lipschitz	235
11.2	Points d'équilibre, stabilité	239
11.3	Compléments	241
11.3.1	Équations linéaires	241
11.3.2	Lemme(s) de Gronwall	242
11.4	Exercices	244

11.1 Théorème de Cauchy–Lipschitz

Étant donnés un ouvert U de \mathbb{R}^d , $y_0 \in U$, un intervalle $I = [t_0, T[$ de \mathbb{R} , une fonction f de $U \times I$ dans \mathbb{R}^d , on appelle problème de Cauchy la recherche de $t \in I \mapsto x(t) \in U$ vérifiant

$$\begin{cases} \dot{y}(t) &= f(y, t), \\ y(0) &= y_0. \end{cases} \quad (11.1)$$

On dit que l'équation est *autonome* si f ne dépend pas du temps, i.e. si $f = f(y)$.

Definition 11.1. (Solution maximale)

On appelle solution maximale du problème de Cauchy (11.1) une fonction $t \mapsto y(t) \in \mathbb{R}^d$ définie sur un intervalle $[t_0, T^*[\subset [t_0, T[$, solution de (11.1), 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 $t \mapsto z(t) \in U$ est solution de (11.1) sur $J[t_0, T^{**}[$, et s'identifie à y sur $[0, \min(T^*, T^{**})[$, alors nécessairement $T^{**} \leq T^*$.

Definition 11.2. (Solution globale)

On appelle solution *globale* du problème de Cauchy (11.1) une solution $t \mapsto y(t) \in \mathbb{R}^d$ définie sur l'intervalle $[t_0, T[$ tout entier. Si la solution n'est définie que sur un sous-intervalle, on parlera de solution *locale*.

Un premier résultat général assure l'existence de solutions maximales

Théorème 11.3. (Cauchy-Peano-Arzelà)

On suppose $f(\cdot, \cdot)$ continue sur $U \times I$. Alors le problème (11.1) admet une solution maximale.

Cette solution n'est pas unique en général. On pourra considérer par exemple le cas $U = \mathbb{R}$, $I = [0, +\infty[, f(y) = \sqrt{y}$ si $y \geq 0$, f identiquement nulle sur \mathbb{R}_- , et la condition initiale $y(t_0) = y(0) = 0$. Les fonctions

$$y_1(t) \equiv 0, \quad y_2(t) = \frac{1}{4}t^2,$$

sont solution globales du même problème de Cauchy.

Remarque 11.4. On prendra garde au fait que le problème ci-dessus est peut sembler d'une certaine manière bien posé, au sens où la connaissance du présent permettrait de prédire l'évolution instantanée et donc de déduire l'avenir. De fait, une discréétisation explicite de ce problème (on parle de schéma d'Euler explicite) est bien un algorithme, i.e. une suite d'opérations élémentaires permettant de calculer tous les termes d'une suite sans ambiguïté :

$$h > 0, \quad y^0 = y_0, \quad y^1 = y^0 + hf(y^1, t^1), \quad \dots \quad y^{n+1} = y^n + hf(y^{n+1}, t^n), \quad \dots$$

Dans le cas du contre-exemple ci-dessus, on retrouve la solution identiquement nulle. Noter que, si l'on considère le schéma dit d'Euler implicite

$$y^{n+1} = y^n + hf(y^{n+1}),$$

le caractère univoque de la relation n'est plus assuré. Pour le contre-exemple précédent, le premier pas s'écrit

$$y^1 = h\sqrt{y^1},$$

qui admet deux solutions : $y^1 = 0$ et $y^1 = h^2$.

Exercice 11.1. Décrire l'ensemble des suites potentiellement produites par le schéma d'Euler implicite de la remarque ci-dessus.

Definition 11.5. (Cylindre de sécurité)

On appelle cylindre de sécurité pour (\bar{y}, \bar{t}) un ensemble $\bar{B}(\bar{y}, \bar{r}) \times [\bar{t}, \bar{t} + \eta]$ tel que toute solution $y(t)$ du problème de Cauchy sur $[\bar{t}, \bar{t} + \eta]$ soit contenue dans $\bar{B}(\bar{y}, r)$, et tel que $|f|$ est borné par une constante M sur le cylindre, avec $r \leq \eta M$.

Proposition 11.6. On suppose que f est continue sur $U \times I$. Alors f admet un cylindre de sécurité en tout point $(\bar{y}, \bar{t}) \subset U \times I$.

Démonstration. La fonction f est définie et continue sur un ensemble du type $\bar{B}(\bar{y}, r) \times [\bar{t}, \bar{t} + \tau]$, qui est compact. Elle est donc notamment bornée par $M > 0$. On choisit $\eta = \min(\tau, r/M)$. Toute solution du problème de Cauchy $y(\bar{t}) = \bar{y}$ est telle que

$$|y(t) - \bar{y}| = \left| \int_0^t f(y(s), s) ds \right| \leq Mt \leq M\eta \leq r,$$

ce qui assure que $\bar{B}(x_0, r) \times [\bar{t}, \bar{t} + \eta]$ est un cylindre de sécurité. \square

Remarque 11.7. Dans le cas où \mathbb{R}^d est remplacé par un espace de Banach de dimension infinie, la démonstration ci-dessus n'est plus valable, car $\bar{B}(\bar{y}, r)$ n'est pas compact. On peut néanmoins montrer une propriété analogue en supposant f localement Lipschitzienne par rapport à la première variable.

Definition 11.8. (Caractère Lipschitz local)

On dit que $f : U \times I \mapsto \mathbb{R}^d$ 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

$$|f(y_2, s) - f(y_1, s)| \leq k |y_2 - y_1| \quad \forall y_1, y_2 \in \bar{B}(y, r), \quad s \in [\max(t - \eta, 0), t + \eta].$$

Proposition 11.9. Soit $f : U \times I \mapsto \mathbb{R}^d$ continue, et telle que les dérivées partielles $\partial f_i / \partial y_j$ existent et sont continues par rapport au couple (y, t) sur $U \times I$. Alors f est localement Lipschitzienne par rapport à la première variable.

Démonstration. Soit $(y, t) \in U \times I$. Il existe r et η tels que $\bar{B}(y, r) \times [\max(t - \eta, 0), t + \eta]$ soit dans $U \times I$. Les dérivées partielles étant continues sur ce compact, elle sont bornées, donc la différentielle de f est elle-même bornée par une constante k . On en déduit le caractère Lipschitz par le théorème des accroissements finis \square

Théorème 11.10. (Cauchy-Lipschitz)

On considère une donnée de Cauchy $(y_0, t_0) \in U \times [t_0, T[$ (avec U ouvert de \mathbb{R}^d), et on suppose que la fonction f , définie de $U \times I$ dans \mathbb{R}^d , est continue sur $U \times I$ et localement Lipschitzienne par rapport à la première variable. Alors le problème de Cauchy (11.1) admet une unique solution maximale définie sur $[t_0, T^*[\subset [t_0, T[$.

Démonstration. D'après la proposition 11.6, le point (y_0, t_0) admet un cylindre de sécurité $\overline{B}(y_0, \bar{r}) \times [t_0, t_0 + \eta]$, sur lequel $|f| \leq M$, avec $\eta M \leq r$. On introduit l'espace X des applications continues sur $[t_0, t_0 + \eta]$ à valeurs dans $\overline{B}(y_0, r)$, muni de la norme de la convergence uniforme, et pour tout $y(\cdot) \in X$, on définit Φy par

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

On a

$$|\Phi y(t) - y_0| \leq \int_{t_0}^t |f(y(s), s)| ds \leq M\eta \leq r,$$

et ainsi Φ est une application de X dans lui-même, et une fonction de $[t_0 - \eta, t_0 + \eta]$ est solution du problème de Cauchy si elle seulement si c'est un point fixe de Φ .

Montrons qu'il existe $n \in \mathbb{N}$ tel que Φ^n soit strictement contractante. Soient $y, z \in X$. On note $y_n = \Phi^n y$ (de même pour z). On a

$$|z_1(t) - y_1(t)| = \left| \int_{t_0}^t (f(z(s), s) - f(y(s), s)) ds \right| \leq kt \|z - y\|_\infty.$$

De même

$$|z_2(t) - y_2(t)| = \left| \int_{t_0}^t (f(z_1(s), s) - f(y_1(s), s)) ds \right| \leq k^2 \left| \int_{t_0}^t s ds \right| \|z - y\|_\infty = \frac{k^2 t^2}{2} \|z - y\|_\infty.$$

On montre ainsi par récurrence que

$$|z_n(t) - y_n(t)| \leq \frac{k^n t^n}{n!} \|z - y\|_\infty \text{ d'où } \|z_n - y_n\|_\infty \leq \frac{k^n \eta^n}{n!} \|z - y\|_\infty$$

de telle sorte que Φ^n est contractante pour n suffisamment grand. D'après le théorème de point fixe de Picard (voir Th. 19.7, page 375), l'application Φ admet un unique point fixe, et l'on en déduit l'existence d'une solution au problème de Cauchy définie sur $[t_0, t_0 + \eta]$, et unique solution sur cet intervalle.

On considère deux solutions y et z du problème de Cauchy, définies sur des intervalles I_y et I_z , de bornes supérieures respectives T_y et T_z . On introduit

$$T^* = \sup \{ t \geq t_0, t \in I_y \cap I_z, y(s) = z(s) \quad \forall s \in [t_0, t] \}.$$

Supposons que $T^* < \min(T_y, T_z)$. On peut reprendre la construction précédente au point $(y^*(T^*), T^*) = (z^*(T^*), T^*)$, et construire ainsi une solution au problème de Cauchy sur $[T^*, T^* + \varepsilon[\subset I_y \cap I_z[, qui s'identifie à y et z par unicité, ce qui est absurde d'où $T = \min(T_y, T_z)$. Soit maintenant J la réunion des intervalles contenant t_0 sur lesquels le problème de Cauchy associé à (y_0, t_0) admet une solution. Pour tout $t \in J$, toutes les solutions définies sur un intervalle contenant t s'identifient d'après ce qui précède. Si J n'est pas égal à I , alors il est du type $[0, T^*[, avec $T^* < T$. En effet, si l'intervalle était fermé en T^* , on pourrait étendre la solution au delà de T^* .$$

On note $y(t)$ cette valeur commune, construisant ainsi une solution maximale unique. \square

Proposition 11.11. On se place dans les hypothèses du théorème précédent, et l'on note y la solution maximale définie sur $[t_0, T^*[$. Pour tout $\tau < T^*$, la solution du problème rétrograde

$$\dot{z} = -f(z, \tau - t)$$

avec condition initiale (que l'on pourrait appeler ici *finale*) $z(0) = y(\tau)$, s'identifie à y parcourue dans le sens rétrograde.

Démonstration. Le problème rétrograde rentre dans le cadre du théorème de Cauchy-Lipschitz (avec $T = \tau - t_0$), il admet donc une solution maximale unique. Or $t \mapsto z(t) = y(\tau - t)$ vérifie

$$\frac{d}{dt}z(t) = -\dot{y}(\tau - t) = -f(y(\tau - t), \tau - t) = -f(z, \tau - t),$$

et la condition initiale $z(0) = y(\tau)$, il s'agit donc bien de cette solution unique. \square

Exercice 11.2. On considère la fonction f définie par $f(y) = y^p$ pour y , f nulle pour y négatif, avec $p \geq 0$. Faire l'analyse du problème de Cauchy sur $[0, +\infty[$ associé à cette fonction (existence et unicité de solutions, caractère global ou pas de ces solutions). (Correction page ??)

Comportement des solutions

Proposition 11.12. (Sortie des compacts)

On se place dans le cadre du théorème 11.10, et l'on note $y(\cdot)$ la solution maximale, définie sur $J = [t_0, T^*[$. Si $T^* < T$, alors y sort de tout compact de U lorsque t tend vers T^* , i.e.

$$\forall K \text{ compact } \subset U, \exists \eta > 0, y(t) \notin K \quad \forall t > T^* - \eta.$$

Démonstration. On suppose $T^* < T$. Si la propriété de sortie des compacts n'est pas vérifiée, il existe un compact $K \subset U$ et une suite (t^n) croissante tendant vers T^* tels que $y(t^n) \in K$ pour tout n . On peut extraire une sous-suite (que l'on note toujours (t^n)) qui converge vers un élément y^∞ de $K \subset U$. On peut placer un cylindre de sécurité passé - avenir¹ $\overline{B}(y_\infty, r) \times [T^* - \eta, T^* + \eta]$ sur lequel f est majoré par M , avec $r \leq \eta M$. Pour n assez grand, $y(t^n)$ est tel que $|y(t^n) - y_\infty| < r/2$ et $T^* - t^n < \eta/2$. Le fermé $\overline{B}(y_\infty, r/2) \times [t^n, t^n + \eta/2]$ est alors un cylindre de sécurité inclus dans le premier. Le procédé de construction d'une solution dans la preuve du théorème 11.10 peut donc être mis en œuvre pour construire une solution au problème de Cauchy de donnée initiale $(y(t^n), t^n)$. Cette solution est définie sur $[t^n, t^n + \eta/2]$ avec $t^n + \eta/2 > T^*$. Cette solution s'identifie à y sur $[t_n, T^*[$ par unicité, et la prolonge strictement au-delà de T^* , ce qui est absurde. \square

Proposition 11.13. On se place dans le cadre des hypothèses du théorème de Cauchy-Lipschitz 11.10, et l'on se donne une fonction V de U dans \mathbb{R} telle que

$$\forall M, \exists K \text{ compact tel que } V(x) \geq M \quad \forall x \in U \setminus K.$$

Alors une solution $t \mapsto y(t)$ du problème de Cauchy sur $[t_0, T^*[$, telle que $V(y(t))$ reste borné sur cet intervalle, ne saurait être maximale.

Démonstration. On considère y une telle solution, avec $V(y(t))$ borné par M sur $[0, T^*[$. D'après l'hypothèse sur V , il existe un compact K tel que $V(y) \geq M + 1$ pour tout y dans $U \setminus K$. On a donc $y(t) \in K$ pour tout $t < T^*$, ce qui exclut le caractère maximal d'après la proposition 11.12. \square

Exercice 11.3. (Existence de solution globale)

Soit Φ une fonction C^2 de \mathbb{R} dans \mathbb{R} , coercive (i.e. $\Phi(y)$ tend vers $+\infty$ quand $|y|$ tend vers l'infini). On considère l'EDO scalaire posée sur $\mathbb{R} \times [0, +\infty[$

$$\dot{y} = -\Phi'(y) + f(t),$$

où f est une fonction continue bornée sur \mathbb{R}_+ , avec donnée initiale $y(0) = y_0$. Montrer qu'il existe une unique solution maximale, et que cette solution est *globale*, i.e. définie sur $[0, +\infty[$.

1. On peut étendre immédiatement la construction de la proposition 11.6 en construire un tel cylindre symétrique par rapport à T^* .

Dépendance par rapport aux conditions initiales

Proposition 11.14. Soit U un ouvert de l'espace de \mathbb{R}^d , $I = [t_0, T[$ un intervalle de \mathbb{R} , et f une fonction continue de $U \times I$ dans \mathbb{R} , Lipschitzienne par rapport à la première variable. Pour y_0, z_0 donnés dans U , on note y et z les solutions au problèmes de Cauchy associées à ces conditions initiales. Alors, sur leur intervalle commun de définition, on a

$$|y(t) - z(t)| \leq e^{k(t-t_0)} |y_0 - z_0|.$$

Démonstration: On a

$$|y(t) - z(t)| = \left| y_0 - z_0 + \int_{t_0}^t (f(y(s), s) - f(z(s), s)) \right| \leq |y_0 - z_0| + k \int_{t_0}^t |y(s) - z(s)|$$

Le lemme de Gronwall 11.24 assure l'inégalité annoncée. \square

Proposition 11.15. Soit $f : \mathbb{R}^d \times I \rightarrow \mathbb{R}^d$ vérifiant les hypothèses du théorème de Cauchy-Lipschitz. On suppose qu'il existe deux constantes A et B telles que

$$|f(y, t)| \leq A|y| + B \quad \text{sur } \mathbb{R}^d \times I.$$

Alors toute solution au problème de Cauchy est définie sur I tout entier.

Démonstration. D'après la proposition 11.12, les solutions maximales ne sont définies sur un sous-intervalle strict que si $|y|$ tend vers $+\infty$. Or on a ici

$$|y(t)| \leq |y_0| + B(t - t_0) + A \int_{t_0}^t |y(s)|.$$

D'après le lemme de Gronwall 11.24 appliqué à $\varphi(t) = |y(t_0 + t)|$, on ne peut donc avoir divergence de $|y|$ vers $+\infty$ en temps fini. \square

11.2 Points d'équilibre, stabilité

Definition 11.16. On se place dans le cadre du problème de Cauchy 11.1, dans le cas autonome, i.e. f ne dépend que de y . On dit que y_{eq} est un point *d'équilibre*, ou point *stationnaire*, si $f(y_{eq}) = 0$, de telle sorte que la fonction constante $t \in I \mapsto y_{eq}$ est solution du problème de Cauchy.

Definition 11.17. (Stabilité, stabilité asymptotique)

Soit $t \mapsto y(t)$ une solution du problème de Cauchy (11.1) associé à (y_0, t_0) , que l'on suppose définie sur $[t_0, +\infty[$. On dit que la solution y est

- (i) *stable* si pour tout $\varepsilon > 0$, il existe $\eta > 0$ tel que, pour tout z_0 tel que $|z_0 - y_0| < \eta$, la trajectoire $t \mapsto z(t)$ associée à la condition initiale z_0 reste à distance de $y(t)$ inférieure à ε ;
- (ii) *asymptotiquement stable* si (i) est vérifié, et que de plus $|z(t) - y(t)|$ tend vers 0 quand t tend vers $+\infty$.

Dans le cas d'un système autonome et d'une solution constante égale à y_0 , on parlera de point d'équilibre stable et asymptotiquement stable.

Le théorème suivant donne une condition suffisante de stabilité asymptotique, ainsi qu'une condition suffisante de non stabilité, pour un point d'équilibre dans \mathbb{R}^d . Il formalise les considérations informelles suivantes. Considérons un point d'équilibre y_{eq} d'une équation autonome. La différence $y - y_{eq}$ vérifie

$$\frac{d}{dt}(y - y_{eq}) = f(y) - f(y_{eq}) \approx \nabla F \cdot (y - y_{eq}).$$

On peut de fait montrer que les solutions du problème linéarisé ci-dessus se comportent au voisinage du point d'équilibre comme les solution du problème de départ. On a donc en particulier les propriétés exprimées par la proposition 11.23, page 242.

Théorème 11.18. Soit y_{eq} un point d'équilibre de l'équation $\dot{y} = f(y)$ dans $U \subset \mathbb{R}^d$. On suppose f continûment différentiable dans un voisinage de y_{eq} , et l'on introduit le gradient

$$\nabla f = \left(\frac{\partial f_i}{\partial y_j} \right)_{1 \leq i, j \leq N}$$

1. Si toutes les valeurs propres de ∇f sont de parties réelles strictement négatives, alors le point y_{eq} est asymptotiquement stable.
2. Si l'une (au moins) des valeurs propres a une partie réelle strictement positive, alors y_{eq} n'est pas stable.

Exemple 11.2.1. 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

$$f(y) = \begin{pmatrix} -y_2 + \alpha |y|^2 y_1 \\ y_1 + \alpha |y|^2 y_2 \end{pmatrix}$$

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 $y_0 = 0$ est stable (mais non asymptotiquement stable). Pour $\alpha > 0$, le point est instable, et pour $\alpha < 0$, le point est asymptotiquement stable.

Stabilité non locale

Definition 11.19. (Fonction de Lyapunov)

On considère un point d'équilibre de l'équation autonome $\dot{y} = f(y)$ dans \mathbb{R}^N , c'est-à-dire un point y_{eq} tel que $f(y_{eq}) = 0$. On appelle fonction de Lyapunov pour y_{eq} une fonction φ continue sur un voisinage V de y_{eq} , continûment différentiable sur $V \setminus \{y_{eq}\}$, et telle que

1. y_{eq} est un minimum strict de φ sur V ,
2. $\nabla \varphi(y) \cdot f(y) \leq 0$ pour tout $y \in V \setminus \{y_{eq}\}$,

Proposition 11.20. Si le point d'équilibre y_{eq} admet une fonctionnelle de Lyapunov, alors il est stable. Si la fonctionnelle peut être choisie de telle sorte que l'inégalité du point 2 est stricte (pour $y \neq y_{eq}$), alors y_{eq} est asymptotiquement stable.

Démonstration. Soit $\varepsilon > 0$, suffisamment petit pour que $\overline{B}(y_{eq}, \varepsilon)$ soit dans V . Le minimum de φ sur la sphère est atteint, il est donc strictement plus grand que la valeur en y_{eq} . On choisit β compris strictement entre ces deux valeurs, et l'on introduit

$$W = \varphi^{-1}(-\infty, \beta] \cap B(y_{eq}, \varepsilon).$$

C'est un ouvert qui contient y_{eq} , il contient donc une boule $B(y_{eq}, \eta)$. Pour toute condition initiale dans cette boule, la trajectoire reste dans $B(y_{eq}, \varepsilon)$, car $\varphi(y(t))$ est décroissant, donc reste inférieur à β , donc ne peut s'approcher de la frontière de $B(y_{eq}, \varepsilon)$.

On suppose maintenant l'inégalité est stricte. On considère une trajectoire $t \mapsto y(t)$ issue de $y(0) \in B(y_{eq}, \eta)$. 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 y de la trajectoire vérifie $\varphi(y) = \ell$, d'où $y = y_{eq}$, et on a convergence de la trajectoire (qui est incluse dans le compact $\overline{B}(y_{eq}, \varepsilon)$) vers y_{eq} . Si la limite est strictement supérieure à ce minimum, on considère l'ensemble

$$A = \varphi^{-1}([\ell, +\infty[) \cap \overline{B}(y_{eq}, \varepsilon).$$

Cet ensemble est compact car fermé borné. La fonction

$$y \mapsto \nabla \varphi(y) \cdot f(y)$$

y atteint donc son maximum, qui est strictement négatif d'après l'hypothèse :

$$\nabla \varphi(y) \cdot f(y) \leq \alpha < 0 \quad \forall y \in A.$$

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. \square

Les résultats précédents portent sur des propriétés de stabilité locale. La notion de fonctionnelle de Lyapunov permet dans certains cas d'assurer le caractère attractif d'un point d'équilibre de façon globale, ou au moins sur une certaine zone de l'espace.

Proposition 11.21. On considère un point d'équilibre y_{eq} de l'équation autonome $\dot{y} = f(y)$ dans un ouvert U de \mathbb{R}^N , sur lequel f est localement Lipschitzienne. Soit $V \subset U$ un ouvert contenant y_{eq} , tel que toute trajectoire issue d'un point de V reste dans un compact inclus dans V . On suppose qu'il existe sur V une fonctionnelle de Lyapunov Ψ stricte au sens suivant :

1. Ψ est continue sur V ,
2. y_{eq} est un minimum strict de Ψ sur V ,
3. La fonction Ψ est strictement décroissante le long de toute trajectoire dans V , sauf pour la trajectoire triviale $y(t) \equiv y_{eq}$.

Alors y_{eq} est attractif sur V , i.e. toute trajectoire issue d'un point de V converge vers y_{eq} .

Démonstration. Soit $y_0 \in V$. On note $y(t)$ la trajectoire issue de y_0 . La quantité $\Psi(y(t))$ est décroissante, elle converge donc vers une limite $\ell \geq \Psi(y_{eq})$ quand t tend vers $+\infty$. Si cette limite est $\Psi(y_{eq})$, alors toute suite extraite convergente de la trajectoire converge vers une limite z dans un compact inclus dans V , donc dans V , et ce z vérifie $\Psi(z) = \Psi(y_{eq})$, on a donc nécessairement $z = y_{eq}$.

On suppose maintenant que $\ell > \Psi(y_{eq})$ (en vue de montrer que c'est impossible). Par hypothèse on peut extraire de la trajectoire une sous-suite qui converge vers z , avec $z \in V$, et l'on a $\Phi(z) = \ell > \Psi(y_{eq})$. On considère la trajectoire z_t issue de $z_0 = z$. Comme $z_0 \neq y_{eq}$, $\Psi(z_t)$ est strictement décroissante, on a $z_1 = \ell - \varepsilon$, avec $\varepsilon > 0$. Par continuité de la solution par rapport aux conditions initiales, il existe $\eta > 0$ tel que, pour $|y_0 - z_0| < \eta$, on a $\varphi(y_1) < \ell - \varepsilon/2$. On peut donc trouver un point y_0 de la suite extraite précédente tel que y_1 , qui fait partie de la trajectoire issue de y_0 , donne à φ une valeur strictement inférieure à ℓ , ce qui est absurde. \square

11.3 Compléments

11.3.1 Équations linéaires

Proposition 11.22. On considère le problème de Cauchy

$$\dot{y} = Ay \quad A \in \mathcal{M}_d(\mathbb{R}), \quad y(0) = y_0.$$

Ce problème admet une solution unique donnée par

$$y(t) = e^{tA}y_0 = \left(\sum_{k=0}^{+\infty} \frac{t^k}{k!} A^k \right) y_0.$$

Si la matrice A est diagonalisable, elle s'écrit PDP^{-1} , on en déduit

$$y(t) = P \left(\sum_{k=0}^{+\infty} \frac{t^k}{k!} D^k \right) P^{-1} y_0 = Pe^{tD}P^{-1}y_0.$$

Si l'on écrit la solution dans une base de vecteurs propres, $x = (x_1, \dots, x_d)$, on a

$$x_i = e^{\lambda_i t} x_i(0).$$

Considérons maintenant une matrice sous forme de bloc de Jordan :

$$A = \begin{pmatrix} \lambda & 1 & 0 & \cdots & 0 \\ 0 & \lambda & 1 & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & 0 \\ \cdot & \cdot & \cdot & \lambda & 1 \\ 0 & \cdot & \cdot & 0 & \lambda \end{pmatrix} = \lambda \text{Id} + N,$$

où N est une matrice nilpotente. On a

$$e^{tA} = e^{\lambda t} \left(\text{Id} + tN + \frac{t^2}{2!} N^2 + \frac{t^3}{3!} N^3 + \cdots + \frac{t^{d-1}}{(d-1)!} N^{d-1} \right).$$

On en déduit la forme générale de l'exponentielle pour une matrice non diagonalisable, en utilisant sa forme réduite de Jordan constituée de blocs diagonaux du type du précédents, avec commutation des matrices correspondantes. On en déduit la propriété suivante :

Proposition 11.23. Si toutes les valeurs propres de A ont des parties réelles strictement négatives, alors e^{tA} tend vers 0 quand t tend vers $+\infty$. Si au moins une valeur propre a une partie réelle strictement positive, alors $\|e^{tA}\|$ tend vers $+\infty$.

11.3.2 Lemme(s) de Gronwall

Proposition 11.24. 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

$$\varphi(t) \leq C + \int_0^t g(s)\varphi(s) \, ds \quad \forall t \in [0, T].$$

On a alors

$$\varphi(t) \leq C \exp \left(\int_0^t g(s) \, ds \right) \quad \forall t \in [0, T].$$

Démonstration: On suppose tout d'abord $C > 0$. La fonction $z(t) = C + \int_0^t g(s)\varphi(s) \, ds$ est dérivable et de dérivée $z' = g\varphi \leq gz$. On a donc (on sait que z par définition ne s'annule pas)

$$\frac{z'}{z} \leq g \implies \varphi \leq z(t) \leq z(0) \exp \left(\int_0^t g(s) \, ds \right) = C \exp \left(\int_0^t g(s) \, ds \right).$$

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 $g \equiv M = \text{constante}$, on a $\varphi(t) \leq C \exp(Mt)$.

La proposition suivante permet d'obtenir, pour certains systèmes dynamiques (en particulier pour les systèmes exprimant le principe fondamental de la dynamique) des estimations plus fines.

Proposition 11.25. 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

$$\varphi(t) \leq C + 2 \int_0^t g(s) \sqrt{\varphi(s)} \, ds \quad \forall t \in [0, T].$$

On a alors

$$\varphi(t) \leq \left(\sqrt{C} + \int_0^t g(s) \, ds \right)^2 \quad \forall t \in [0, T].$$

Démonstration. La démonstration est analogue à la précédente, en considérant maintenant la fonction

$$z(t) = C + 2 \int_0^t g(s) \sqrt{\varphi(s)}.$$

□

Definition 11.26. (Flot d'une équation différentielle)

On considère l'équation différentielle (11.1), sous les hypothèses du théorème (11.10). On appelle flot de l'équation différentielle l'application Φ qui au triplet $(y_0, t_0; t)$ associe la solution au temps t du problème de Cauchy pour la donnée (y_0, t_0) . Cette application vérifie donc

$$\begin{cases} \frac{\partial \Phi}{\partial t}(y_0, t_0; t) &= f(\Phi(y_0, t_0; t), t), \\ \Phi(y_0, t_0; t_0) &= y_0. \end{cases} \quad (11.2)$$

Cette application est définie sur

$$\bigcup_{(y_0, t_0) \in U \times I} \{(y_0, t_0)\} \times I_{(y_0, t_0)}$$

où $I_{(y_0, t_0)}$ est l'intervalle de définition de la solution maximale associée à la donnée de Cauchy (y_0, t_0) .

Proposition 11.27. On se place dans le cadre de la définition précédente, en supposant de plus que la fonction f est globalement Lipschitzienne par rapport à la première variable sur $U \times I$, de constante de Lipschitz k . Alors

$$|\Phi(y_0, t_0; t) - \Phi(y_0, t_0; t)| \leq e^{k(t-t_0)} |y_0 - y_0|.$$

Démonstration. C'est une application directe de la proposition (11.14).

□

11.4 Exercices

Exercice 11.4. On considère un modèle (appelé modèle de Verhulst) de croissance de population avec taux de reproduction de base $\beta > 0$, et une capacité maximal du milieu y_{max} :

$$\dot{y} = \beta y (1 - y/y_{max}).$$

Montrer que, pour toute donnée initiale $y_0 \in [0, y_{max}]$, le problème de Cauchy admet une solution globale à valeurs dans $[0, y_{max}]$, et préciser le comportement des solutions. (Correction page ??)

Exercice 11.5. Dans la continuité de l'exercice 11.4, on considère un modèle de population appelé *bistable*

$$\dot{y} = y(1-y)(y-\theta),$$

où θ est un paramètre entre 0 et 1. Faire l'analyse du problème de Cauchy à donnée initiale $y_0 \in [0, 1]$ (existence et unicité des solutions, points d'équilibre, stabilité, comportement des solutions).

Exercice 11.6. On considère un système de N particules de même masse en interaction gravitationnelle : la force exercée par la masse j sur la masse i est

$$f_{ji} = \frac{e_{ij}}{|x_j - x_i|^2}.$$

Écrire le système d'ordre 1 sur les positions et les vitesses qui exprime le principe fondamental de la dynamique, et faire l'analyse du problème (existence et unicité des solutions, caractère global ou non de ces solutions).

Exercice 11.7. On considère un système de N particules de même masse et de même charge en interaction coulombienne : la force exercée par la masse j sur la masse i est

$$f_{ji} = -\frac{e_{ij}}{|x_j - x_i|^2}.$$

Écrire le système d'ordre 1 sur les positions et les vitesses qui exprime le principe fondamental de la dynamique, écrire la conservation de l'énergie totale pour ces solutions, et faire l'analyse du problème (existence et unicité des solutions, caractère global ou non de ces solutions).

De façon plus générale, on considère un système de N particules dans \mathbb{R}^d en interaction selon un potentiel $V = V(x_1, \dots, x_N)$ deux fois continûment différentiable sur \mathbb{R}^{dN} , c'est à dire l'on a

$$f_{ji} = -\nabla_{x_i} V.$$

Montrer que, si V est minoré, les solutions sont définies globalement.

Si l'on suppose seulement le potentiel défini et régulier en dehors de l'ensemble des configurations telles que deux particules au moins sont confondues, quelle condition supplémentaire sur V assure-t-elle le caractère global des solutions ?

Exercice 11.8. Soit f une fonction continûment différentiable de \mathbb{R}^d dans \mathbb{R}^d , et qui vérifie

$$\langle f(x) | x \rangle \leq |x|^2 \quad \forall x.$$

a) Montrer que le problème de Cauchy sur \mathbb{R}_+ admet une solution globale pour toute condition initiale.

b) donner des exemples d'applications f pour lesquelles les solutions sont définies globalement, mais tel que le modèle rétrograde (on remplace f par $-f$) admettent des solutions qui explosent en temps fini.

Exercice 11.9. (Ventilation humaine)

On considère le modèle suivant de ventilation humaine :

$$R \frac{dV}{dt} + \varphi(V) = -P(t),$$

où $V = V(t)$ et le volume d'air contenu à l'instant t dans les poumons, $R > 0$ est la résistance de l'arbre bronchique, $P(t)$ la pression exercée à l'extérieur des alvéoles (pression négative lors de la contraction du diaphragme, qui tend à augmenter le volume), et $\varphi(V)$ une fonction encodant la propriétés mécaniques du poumon (tendance à retourner vers sa forme d'équilibre).

a) On suppose dans un premier temps que le modèle de déformation est linéaire, et s'écrit

$$\varphi(V) = E(V - V_0),$$

où V_0 est le volume à l'équilibre, et $E > 0$ un paramètre appelé *elastance*².

Donner l'expression explicite de la solution du problème de Cauchy avec condition initiale $V(0) = V_{init}$, et donner l'allure des solutions dans le cas où P est périodique et constante par morceaux (pression négative pour l'inspiration, pression nulle pour l'expiration).

b) Écrire le bilan énergétique du système et commenter sa conformité avec le Premier Principe de la thermodynamique . Proposer un modèle qui prenne en compte l'inertie du poumon, et écrire le nouveau bilan énergétique. On pourra introduire le paramètre I , appelé *inertance* de l'appareil respiratoire, tel que $I\dot{V}^2/2$ correspond à une énergie cinétique.

c) Si l'on suppose le terme de forçage $P(\cdot)$ périodique, montrer que la solution se comporte asymptotiquement de façon périodique (dans un sens que l'on précisera).

d) Montrer que, selon ce modèle, le volume peut prendre des valeurs arbitrairement grandes et arbitrairement petites, en particulier négatives.

e) Pour pallier les problèmes évoqués à la question précédente, on considère un modèle élastique non linéaire. On considère que φ est définie sur $]V_{min}, V_{max}[$, avec $V_{min} > 0$, telle que $\varphi(V) = \Psi'(V)$, où Ψ est une fonction C^2 , qui tend vers $+\infty$ quand V tend vers V_{min}^+ ou V_{max}^- . On suppose P continue. Montrer que le problème admet une solution unique définie sur $[0, +\infty[$ tout entier.

Exercice 11.10. (Modèle SIR)

On considère le modèle dit SIR de propagation d'une épidémie, basé sur les fractions S (susceptible), I (infected), et R , recovered, de la population globale.

$$\begin{cases} \frac{dS}{dt} = -\beta SI \\ \frac{dI}{dt} = +\beta SI - \gamma I \\ \frac{dR}{dt} = +\gamma I \end{cases} \quad (11.3)$$

où $\beta > 0$ est le taux d'incidence (produit $p \times K$ du nombre de contacts moyen par unité de temps, et p la probabilité d'infection lors d'un contact), et $\gamma = 1/D$ le taux de guérison, inverse de la durée moyenne de la maladie.

a) Vérifier la conservativité du système, en montrant que, pour toute solution de ce système, la somme des fractions $S + I + R$ est conservée.

On ne considère maintenant que les 2 premières équations.

b) Justifier que l'on puisse se contenter d'étudier ce sous-système.

On se donne des conditions initiales $S_0 \geq 0$, $I_0 \geq 0$, telles que $S_0 + I_0 \leq 1$.

c) Montrer l'existence et l'unicité d'une solution $y(t) = (S(t), I(t))$ maximale, définie globalement, à valeurs dans $[0, 1] \times [0, 1]$.

d) Préciser les points fixes de ce système, et étudier leur stabilité. Préciser la signification en termes de modélisation de ces notions.

2. Ce paramètre joue ici, pour ce modèle dont la variable est un *volume*, le rôle d'une constante de raideur pour les modèles basés sur des *déplacements*.

N.B. : On pourra introduire le paramètre $R_0 = \beta/\gamma$, appelé *taux de reproduction de base*.

e) Montrer que la quantité

$$\Phi(S, I) = \beta(S + I) - \gamma \log(S)$$

se conserve le long des solutions. Que peut on en déduire sur la forme des trajectoires dans l'espace des phases ?

Exercice 11.11. On cherche à modéliser un matériaux composite constitué d'une phase rigide et de petites inclusions remplies de gaz parfait compressible. On considère dans ce but un système constitué de cloisons coulissantes, situées en $x_0 = 0 < x_1 < \dots < x_{N-1} < x_N = L$, de même masse. On suppose que le volume de l'air coincé entre 2 cloisons est $x_{j+1} - x_j$, et que la pression est inverse à ce volume. On suppose que les cloisons coulissent avec un frottement caractérisé par un paramètre $\nu \geq 0$, de telle sorte que le système s'écrit

$$\ddot{x}_j = \frac{1}{\tau} \left(\frac{1}{x_j - x_{j-1}} - \frac{1}{x_{j+1} - x_j} \right) - \nu \dot{x}_j \quad \forall j = 1, \dots, N-1$$

On introduit

$$\Lambda = \{x = (x_1, \dots, x_{N-1}) \in \mathbb{R}^{N-1}, , 0 < x_1 < \dots < x_{N-1} < L\}.$$

- 1) Montrer que, pour toute donnée initiale en position-vitesse $(x, u) \in \Lambda \times \mathbb{R}^{N-1}$, le problème admet une solution unique définie globalement.
- 2) Montrer que le système admet un unique point d'équilibre où tous les compartiment ont une même longueur $a = L/N$, et les vitesses sont nulles.
- 3) Étudier la stabilité du point d'équilibre ci-dessus, selon que ν soit nul ou pas.

Chapitre 12

Espaces de Sobolev

Sommaire

12.1	Définitions, propriétés générales	247
12.2	Traces	249
12.3	Injections	253
12.4	Inégalités de Poincaré	253
12.5	Problèmes aux limites elliptiques	255
12.6	Espaces de Sobolev et transformation de Fourier	258

Ce chapitre présente un certain nombre de définitions et propriétés autour de la notion d'espace de Sobolev, sans démonstration.

12.1 Définitions, propriétés générales

Definition 12.1. (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

$$\nabla \varphi = \begin{pmatrix} \frac{\partial \varphi}{\partial x_1} \\ \vdots \\ \frac{\partial \varphi}{\partial x_N} \end{pmatrix}.$$

Definition 12.2. (Espace de Sobolev)

On définit l'espace de Sobolev $H^1(\Omega)$ comme l'ensemble des fonctions u dans $L^2(\Omega)$ telles qu'il existe $v = (v_1, \dots, v_N) \in (L^2(\Omega))^N$ vérifiant

$$\int_{\Omega} u \frac{\partial \varphi}{\partial x_i} = - \int_{\Omega} \varphi v_i \quad \forall \varphi \in \mathcal{D}(\Omega), \quad \forall i = 1, \dots, N.$$

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 12.3. 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 séparable¹.

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'ambigüïté), 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 12.4. 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 12.1) s'identifie au gradient au sens de Sobolev (définition 12.2).

Proposition 12.5. 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 .

Proposition 12.6. L'espace $\mathcal{D}(\mathbb{R}^N)$ est dense dans $H^1(\mathbb{R}^N)$.

On dit que ω est fortement inclus dans Ω si $\bar{\omega}$ est compact et inclus dans Ω . On note $\omega \subset\subset \Omega$.

Proposition 12.7. Pour tout $\omega \subset\subset \Omega$, tout $u \in H^1(\Omega)$, il existe une suite (u_n) dans $\mathcal{D}(\Omega)$ telle que

$$u_n \rightarrow u \text{ dans } L^2(\Omega), \quad \nabla u_n \rightarrow \nabla u \text{ dans } L^2(\omega)^N.$$

Corollaire 12.8. Soit (ω_n) une suite de domaines fortement inclus dans Ω , et $u \in H^1(\Omega)$. Il existe une suite (u_n) dans $\mathcal{D}(\Omega)$ telle que

$$\|u_n - u\|_{L^2(\Omega)} \rightarrow 0, \quad \|\nabla u_n - \nabla u\|_{L^2(\omega_n)^N} \rightarrow 0.$$

Definition 12.9. On définit $H_0^1(\Omega)$ comme l'adhérence de $\mathcal{D}(\Omega)$ dans $H^1(\Omega)$.

Noter que, d'après la proposition 12.7, 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 12.2, la définition de l'opérateur trace sur le bord γ_0). On peut néanmoins 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 12.21), nous noterons $\tilde{H}^{1/2}$ l'espace abstrait correspondant.

Definition 12.10. 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

$$\|u\|_{H^2(\Omega)}^2 = |u|_0^2 + \sum_i \left| \frac{\partial u}{\partial x_i} \right|_0^2 + \sum_{i,j} \left| \frac{\partial^2 u}{\partial x_i \partial x_j} \right|_0^2 = |u|_{0,\Omega}^2 + |u|_{1,\Omega}^2 + |u|_{2,\Omega}^2.$$

1. Il contient un sous-ensemble dénombrable et dense

On peut définir de façon analogue les espaces $H^m(\Omega)$ pour $m = 3, 4, \dots$, mais nous n'utiliserons ici que $m \leq 2$.

Definition 12.11. (Espace H_{loc}^m)

Soit m un entier positif (on utilisera le cas $m = 2$ dans la suite). On définit l'espace $H_{\text{loc}}^m(\Omega)$ 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_{loc}^m 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} .

Proposition 12.12. 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

$$\begin{aligned} \int_{\mathbb{R}^N} \tilde{u} \nabla \varphi &= \int_{\Omega} u \nabla \varphi = \lim_{n \rightarrow +\infty} \int_{\Omega} u_n \nabla \varphi \\ &= - \lim_{n \rightarrow +\infty} \int_{\Omega} \varphi \nabla u_n = - \int_{\Omega} \varphi \nabla u = - \int_{\mathbb{R}^N} \varphi v. \end{aligned}$$

où v est le champ de vecteurs qui vaut ∇u dans Ω , et 0 à l'extérieur de Ω . □

12.2 Traces

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 a priori 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, même si cette régularité ne va pas jusqu'à la possibilité de définir des valeurs ponctuelles. Cette section décrit succinctement les étapes essentielles à la construction de ces valeurs au bord (traces).

Definition 12.13. (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 12.14. On a $H_0^1(\mathbb{R}^N) = H^1(\mathbb{R}^N)$ (d'après la proposition 12.6), 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 12.1 ci-après sur la notion de *capacité*).

Exercice 12.1. (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

$$C_\omega = \inf \left\{ \int_{\Omega} |\nabla u|^2 , v|_\omega \equiv 1 \text{ sur } \omega , v \in D(\Omega) \right\}.$$

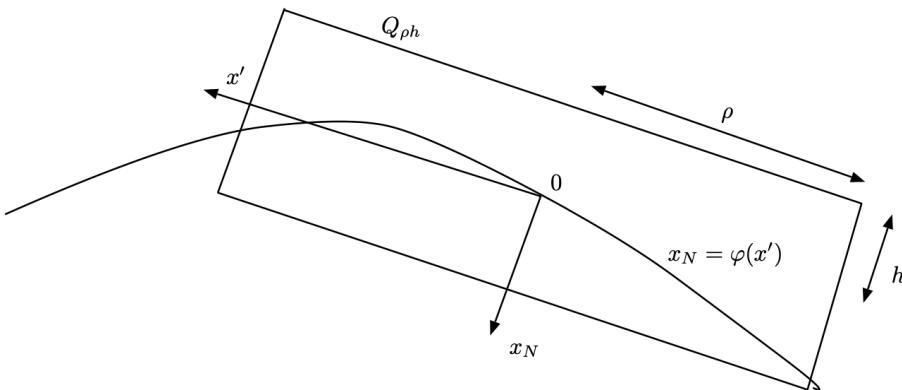


FIGURE 12.1 – Régularité de la frontière

- 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 $H^1(\Omega)$.

On définit le cylindre $Q_{\rho h}$ de \mathbb{R}^N par

$$Q_{\rho h} = \{x \in \mathbb{R}^N, x = (x', x_N) = (x_1, \dots, x_N), |x'| < \rho, -h < x_N < h\}.$$

Dans la définition qui suit, "X" représente une régularité fonctionnelle du type C^o , Lipschitz, C^k , etc...

Definition 12.15. 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 $\rho, h > 0$, tels qu'il existe une application

$$\varphi : \{x' \in \mathbb{R}^{N-1}, |x'| < \rho\} \longrightarrow \mathbb{R}$$

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 12.16. (vecteur normal)

Soit Ω un ouvert de classe 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 $\mathcal{D}(\bar{\Omega})$ l'ensemble des restrictions des fonctions de $\mathcal{D}(\mathbb{R}^N)$ à $\bar{\Omega}$.

Proposition 12.17. 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 Brezis² 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 $x \mapsto \eta_\varepsilon(x) = \eta(x/\varepsilon)$, 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'autre 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 12.18. Soit Ω un ouvert de frontière Γ Lipschitzienne. Alors $\mathcal{D}(\overline{\Omega})$ est dense dans $H^1(\Omega)$.

Proposition 12.19. 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 l'ouvrage précité pour une démonstration plus complète. Soit φ une fonction régulière à support compact ($\varphi \in \mathcal{D}(\overline{\Omega})$). Cette fonction est nulle pour $x_N \leq M$. On a donc

$$\begin{aligned} \varphi(x', 0)^2 &= \int_M^0 \partial_N(\varphi(x', s)^2) ds = 2 \int_M^0 \varphi(x', s) \partial_N \varphi(x', s) ds \\ &\leq 2 \left(\int_0^M \varphi(x', s)^2 \right)^{1/2} \left(\int_0^M (\partial_N \varphi(x', s))^2 \right)^{1/2}, \\ &\leq \int_0^M \varphi(x', s)^2 ds + \int_0^M (\partial_N \varphi(x', s))^2 ds, \end{aligned}$$

2. H. Brezis, Analyse fonctionnelle : Théorie et applications, Ed. Dunod

d'où

$$\int_{\mathbb{R}^{N-1}} \varphi(x', 0)^2 \leq \int_{\mathbb{R}^N} \varphi^2 + \int_{\mathbb{R}^N} \varphi^2 \int_{\mathbb{R}^N} (\partial_N \varphi)^2 \leq \int_{\mathbb{R}^N} \varphi^2 + \int_{\mathbb{R}^N} |\nabla u|^2,$$

d'où la continuité $(H^1(\Omega), L^2(\Gamma))$ de l'application restriction. On conclut par densité de $\mathcal{D}(\bar{\Omega})$ dans $H^1(\Omega)$. \square

Remarque 12.20. 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 12.21. (Espace $H^{1/2}(\Gamma)$)

On note $H^{1/2}(\Gamma) \subset L^2(\Gamma)$ l'image de l'application $\gamma_0 : H^1(\Omega) \mapsto L^2(\Gamma)$ 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 12.22. L'espace $H^{1/2}$ peut se définir sur l'espace entier par la transformée de Fourier (voir section 12.6), 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 12.22, page 252).

Proposition 12.23. L'espace $H_0^1(\Omega)$ est constitué des fonctions de $H^1(\Omega)$ dont la trace sur $\partial\Omega$ est nulle.

Definition 12.24. (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 $\varphi \in \mathcal{D}(\bar{\Omega})$, on appelle dérivée normale de φ en un point de Γ la quantité

$$\frac{\partial \varphi}{\partial n} = \nabla \varphi \cdot n.$$

Definition 12.25. 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 $\mathcal{D}(\bar{\Omega})$ dans $H^2(\Omega)$ (qui, de fait, n'est pas exigée).

Proposition 12.26. (Première formule de Green)

Soit Ω un ouvert borné de frontière Γ Lipschitzienne. Pour tous u et v dans $H^1(\Omega)$, on a

$$\int_{\Omega} v \nabla u = - \int_{\Omega} u \nabla v + \int_{\Gamma} u v n.$$

Proposition 12.27. (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

$$-\int_{\Omega} v \Delta u = \int_{\Omega} \nabla u \cdot \nabla v - \int_{\Gamma} \frac{\partial u}{\partial n} v.$$

Proposition 12.28. Soit Ω un ouvert borné de frontière Γ lipschitzienne. On suppose que Ω se décompose de la façon suivante

$$\bar{\Omega} = \bigcup_{i=1, \dots, p} \bar{\Omega}_i,$$

où les Ω_i sont des ouverts de frontière lipschitzienne, inclus dans Ω , deux à deux disjoints. On note $\Gamma_{ij} = \bar{\Omega}_i \cap \bar{\Omega}_j$. 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 12.26 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 $\nabla u = v$. \square

Remarque 12.29. 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 12.30. 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 $H^2(\Omega)$.

12.3 Injections

Théorème 12.31. 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 12.32. (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.

12.4 Inégalités de Poincaré

Proposition 12.33. (Inégalité de Poincaré)

Soit Ω un domaine de \mathbb{R}^N borné dans une direction, c'est-à-dire tel que

$$\Omega \subset \{x \in \mathbb{R}^N, \xi \cdot x \in]a, b[\}$$

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

$$\{x = (x_1, \dots, x_N) = (x', x_N) \in \mathbb{R}^N, x_N \in]0, L[\}.$$

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'après 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. \square

Remarque 12.34. On appelle constante de Poincaré du domaine Ω le plus petit réel C_{Ω} tel que l'inégalité ci-dessus est vérifiée. On a

$$\frac{1}{C_{\Omega}^2} = \inf_{u \neq 0} \frac{\int_{\Omega} |\nabla u|^2}{\int_{\Omega} |u|^2}.$$

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 vérifiant³

$$-\Delta u = \lambda u.$$

La proposition précédente assure $\lambda_1 \geq 1/L^2$, pour tout domaine Ω inclus dans une bande d'épaisseur L .

Corollaire 12.35. Soit Ω un domaine de \mathbb{R}^N borné dans une direction. Alors la forme bilinéaire

$$(u, v) \mapsto \int_{\Omega} \nabla u \cdot \nabla v$$

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 12.36. (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

$$|u|_0 \leq C(|Tu|_M + |\nabla u|_0) \quad \forall u \in H^1(\Omega).$$

3. L'opérateur de Laplace $-\Delta$, 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 $-\Delta u = \lambda u$ presque partout.

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

$$\|u_n\|_{L^2} > n (|Tu_n|_M + |\nabla u_n|_0) \quad \forall u \in H^1(\Omega).$$

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 $(\nabla u_n) \rightarrow 0$, on a $\nabla u = 0$, et ainsi u est constante sur Ω (voir proposition 12.5, page 248). Comme $Tu = 0$, cette constante est nulle, ce qui est absurde car $\|u\| = \lim \|u_n\| = 1$ \square

La démonstration ci-dessus permet d'établir directement la propriété suivante :

Corollaire 12.37. 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 12.38. 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 équivalente à la norme H^1 .

12.5 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

$$\begin{cases} -\Delta u &= f & \text{dans } \Omega \\ u &= 0 & \text{sur } \partial\Omega, \end{cases} \quad (12.1)$$

où f est une fonction de $L^2(\Omega)$ donnée. On parlera du problème de Poisson dans le domaine Ω .

Definition 12.39. (Solution faible)

On appellera solution faible de (12.1) une fonction de $H_0^1(\Omega)$ telle que

$$\int_{\Omega} \nabla u \cdot \nabla v = \int_{\Omega} fv \quad \forall v \in H_0^1(\Omega). \quad (12.2)$$

Proposition 12.40. (Principe de Dirichlet)

On suppose Ω borné dans une direction. Soit $f \in L^2(\Omega)$. Alors le problème 12.1 admet une unique solution faible : il existe un unique $u \in H_0^1(\Omega)$ solution de la formulation variationnelle (12.2). C'est l'unique élément de $H_0^1(\Omega)$ qui minimise la fonctionnelle

$$v \mapsto \frac{1}{2} \int_{\Omega} |\nabla v|^2 - \int_{\Omega} fv.$$

Démonstration: C'est une application directe du théorème de Lax-Milgram, avec

$$a(u, v) = \int_{\Omega} \nabla u \cdot \nabla v, \quad \langle \varphi, v \rangle = \int_{\Omega} fv.$$

Noter que la forme bilinéaire $a(\cdot, \cdot)$ est bien coercive grâce à l'inégalité de Poincaré (proposition 12.33, page 253). \square

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

$$\begin{cases} u - \Delta u = f & \text{dans } \Omega \\ \frac{\partial u}{\partial n} = 0 & \text{sur } \partial\Omega, \end{cases} \quad (12.3)$$

où f est donnée.

Definition 12.41. On appellera solution classique (dans le cas où f est au moins continue) une fonction de $C^2(\bar{\Omega})$ qui vérifie le système ci-dessus, et solution faible une fonction de $H^1(\Omega)$ telle que

$$\int_{\Omega} uv + \int_{\Omega} \nabla u \cdot \nabla v = \int_{\Omega} fv \quad \forall v \in H^1(\Omega). \quad (12.4)$$

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 (12.3), car la dérivée normale n'est en général pas définie sur le bord.

Proposition 12.42. Soit $f \in L^2(\Omega)$. Alors le problème 12.3 admet une unique solution faible. Cette solution faible est l'élément de $H_0^1(\Omega)$ qui minimise la fonctionnelle

$$v \mapsto \frac{1}{2} \int_{\Omega} |v|^2 + \frac{1}{2} \int_{\Omega} |\nabla v|^2 - \int_{\Omega} fv.$$

Démonstration: C'est de nouveau une application directe du théorème de Lax-Milgram dans $H = H^1(\Omega)$. \square

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

$$-\Delta u = f, \quad u - \Delta u = f \text{ ou } -\nabla k \cdot \nabla u = f,$$

où Δ est le Laplacien $\Delta = \sum \partial^2 / \partial x_i^2$, k est un champ scalaire régulier tel que $0 < m \leq k(x) \leq M < +\infty$.

Proposition 12.43. 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 $H_{loc}^2(\Omega)$ et vérifie

$$-\Delta u = f \quad \text{p.p.}$$

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

$$D_h u = \frac{1}{|h|} (\tau_h u - u),$$

et l'on écrit la formulation variationnelle avec $v = D_{-h} D_h u$. Il vient

$$\int_{\mathbb{R}^N} \nabla u \cdot \nabla v = \frac{1}{|h|^2} \int_{\mathbb{R}^N} \nabla u \cdot (\tau_h \nabla u - 2\nabla u + \tau_{-h} \nabla u).$$

On peut écrire

$$\int_{\mathbb{R}^N} \nabla u \cdot (-\nabla u + \tau_{-h} \nabla u) = \int_{\mathbb{R}^N} \tau_h \nabla u \cdot (-\tau_h \nabla u + \nabla u),$$

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 ?? ((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 ??, $\partial_i u \in H^1(\mathbb{R}^N)$ pour tout $i = 1, \dots, N$.

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

$$\int_{\mathbb{R}^N} \nabla(\theta u) \cdot \nabla \varphi = \int_{\mathbb{R}^N} \theta f \varphi - 2 \int_{\mathbb{R}^N} \varphi \nabla u \cdot \nabla \theta - \int_{\mathbb{R}^N} \varphi u \Delta \theta = \int_{\mathbb{R}^N} g \varphi \quad \forall \varphi \in \mathcal{D}(\Omega).$$

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 $u \in H_{\text{loc}}^2(\Omega)$. \square

Proposition 12.44. 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 (12.2) avec conditions de Dirichlet homogènes est dans $H_{\text{loc}}^2(\Omega)$ et vérifie

$$-\Delta u = f \quad \text{p.p.}$$

Démonstration: C'est une application directe de la proposition 12.43. \square

Le passage de la régularité H_{loc}^2 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 12.45. 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 $-\Delta u = f$ 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 $H_{\text{loc}}^2(\Omega)$ est assurée par la proposition 12.43. On se reportera à Brezis [?, 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 12.43, les translations étant effectuées parallèlement au bord considéré. \square

Proposition 12.46. Les conclusions du théorème ci-dessus sont valides si l'on suppose le domaine polyédrique et convexe.

Proposition 12.47. 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 $\bar{\Omega}$, minorée par une constante

Remarque 12.48. 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 Costabel⁴ une analyse détaillée de la régularité dans ce type de situation, en fonction de l'angle du raccord entre les composantes.

Remarque 12.49. Si l'on considère le problème

$$u - \Delta u = f,$$

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 12.50. Soit Ω un domaine de frontière C^2 et bornée, et f un élément de $L^2(\Omega)$. La solution de (12.4) appartient à H^2 , et sa dérivée normale est nulle sur $\Gamma = \partial\Omega$.

12.6 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 néanmoins 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 12.51. 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 12.52. L'application $u \mapsto \tilde{u}$ 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 théorème si l'on prend la définition 12.2, page 247 comme référence.

Théorème 12.53. L'espace $H^1(\mathbb{R}^N)$ est l'ensemble des fonctions u de $L^2(\mathbb{R}^N)$ telles que

$$(1 + |\xi|^2)^{1/2} \tilde{u} \in L^2(\mathbb{R}^N).$$

Nous démontrons à présent le théorème de Rellich 12.32 déjà énoncé à la page 253.

Théorème 12.54. Soit Ω un domaine borné de frontière lipschitzienne. L'injection de $H^1(\Omega)$ dans $L^2(\Omega)$ est compacte.

4. 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>

Démonstration: On considère une suite (u_n) bornée dans $H^1(\Omega)$. On note P l'opérateur de prolongement de la proposition 12.17, page 251. 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 18.32, page 365, 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 $M \geq 0$

$$\|u_n\|_{L^2}^2 = \|\tilde{u}_n\|_{L^2}^2 = \int_{|\xi| < M} |\tilde{u}_n|^2 + \int_{|\xi| > M} |\tilde{u}_n|^2 \leq \int_{|\xi| < M} |\tilde{u}_n|^2 + \frac{1}{1+M^2} \int_{|\xi| > M} (1+|\xi|^2) |\tilde{u}_n|^2.$$

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 $\|u_n\|_{L^2}$. \square

Chapitre 13

Éléments d'optimisation

Sommaire

13.1	Éléments d'analyse convexe	260
13.2	Existence d'un minimiseur, conditions d'optimalité	261
13.3	Contraintes unilatérales	264
13.4	Point-selle, théorème de Kuhn et Tucker	269
13.5	Formalisme de l'analyse non lisse, sous-différentiels	272
13.6	Dérivation du minimum par rapport aux contraintes	277
13.7	Cas de la dimension infinie	278
13.8	Contraintes non linéaires d'égalité	279
13.9	Illustrations	281
13.10	Exercices	283

13.1 Éléments d'analyse convexe

Definition 13.1. (Partie convexe)

Soit E un espace vectoriel (ou affine). On dit que $X \subset E$ est convexe si $(1 - \theta)x + \theta y \in X$, i.e.

$$(1 - \theta)x + \theta y \in X \quad \forall x, y \in X, \theta \in [0, 1].$$

Exercice 13.1. Montrer que $X \subset E$ est convexe si et seulement si

$$\sum_{i=1}^n \theta_i x_i \in X \quad \forall x_1, \dots, x_n \in X, \theta_1, \dots, \theta_n \geq 0, \sum_{i=1}^n \theta_i = 1.$$

Definition 13.2. (Enveloppe convexe)

Soit E un espace vectoriel (ou affine), et $X \subset E$. On appelle enveloppe convexe de X , et l'on note $co(X)$ le plus petit convexe contenant X , i.e.

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

Definition 13.3. (Coercivité)

Soit E un espace vectoriel normé, et J une fonctionnelle définie d'une partie X de E dans \mathbb{R} . On dit que J est coercive sur X si

$$\lim_{x \in X, \|x\| \rightarrow +\infty} J(x) = \infty.$$

Definition 13.4. (Fonctionnelle convexe, strictement convexe)

Soit E un espace affine, et J une fonctionnelle définie d'une partie convexe $X \subset E$ dans \mathbb{R} . On dit que J est convexe sur X si

$$J((1 - \theta)x + \theta y) \leq (1 - \theta)J(x) + \theta J(y) \quad \forall x, y \in X, \theta \in]0, 1[.$$

On dit que J est *strictement convexe* si l'inégalité ci-dessus est stricte dès que $x \neq y$.

Definition 13.5. (Fonctionnelle α -convexe, fortement convexe)

Soit E un espace vectoriel normé, et J une fonctionnelle définie d'une partie convexe $X \subset E$ dans \mathbb{R} . On dit que J est α -convexe, pour $\alpha \in \mathbb{R}$, si

$$J((1-\theta)x + \theta y) \leq (1-\theta)J(x) + \theta J(y) - \frac{\alpha}{2}\theta(1-\theta)\|x-y\|^2 \quad \forall x, y \in X, \theta \in [0, 1].$$

Une fonctionnelle α -convexe avec $\alpha > 0$ est dite *fortement convexe*.

N.B. : la définition de l' α -convexité est telle que, dans un espace de Hilbert, la fonctionnelle canonique $|x|^2/2$ est exactement 1-convexe. Pour cette fonctionnelle particulière, l'inégalité ci-dessus est une égalité.

La notion d' α -convexité porte sur des propriétés à la fois locales et globales : elle induit une minoration de la dérivée seconde (voir exercice 13.6), et si $\alpha > 0$ elle implique la coercivité (voir exercice 13.7, page 283), qui est une propriété globale.

Une fonctionnelle fortement convexe est de façon évidente strictement convexe. Une fonctionnelle strictement convexe peut en revanche ne pas être fortement convexe (par exemple $x \mapsto x^4$, qui viole la condition en 0, ou $x \mapsto |x|^{3/2}$, qui viole la condition en $\pm\infty$).

Pour $\alpha < 0$, une fonctionnelle α -convexe peut ne pas être convexe, il s'agit d'une notion *affaiblie* de convexité. Une telle fonction peut en revanche être rendue convexe par l'ajout d'un terme quadratique. Certaines fonctions ne sont α -convexe pour aucun α , considérer par exemple $x \mapsto -|x|$ (la concavité singulière en 0 n'est pas rattrapable).

Definition 13.6. (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 . On dira que J est continûment différentiable sur U si elle est différentiable en tout point de U , et si la correspondance $x \mapsto DJ(x)$ est continue.

Definition 13.7. (Gradient d'une fonctionnelle)

Dans le cadre de la définition précédente, si l'on suppose de plus que E est un espace de Hilbert, alors la différentielle $DF(u)$ de F en u s'identifie, par le théorème de Riez-Fréchet, à un vecteur de E . On appelle ce vecteur le gradient de F en u , et on le note $\nabla F(u)$.

Exercice 13.2. On se place sur \mathbb{R}^d , et l'on considère une fonctionnelle J différentiable en x . On associe à une matrice symétrique définie positive M le produit scalaire

$$\langle x | y \rangle_M = \langle Mx | y \rangle.$$

a) Exprimer $\nabla^M J(x)$, gradient de J en x relativement à ce produit scalaire, en fonction du gradient relativement au produit scalaire canonique.

b) Quel est l'ensemble décrit par $\nabla^M J(x)$ lorsque M décrit le cône des matrices symétriques définies positives ?

13.2 Existence d'un minimiseur, conditions d'optimalité

Proposition 13.8. Soit J une fonctionnelle continue d'un fermé non vide $F \subset \mathbb{R}^d$ dans \mathbb{R} . On suppose que l'une des conditions suivantes est vérifiée¹ :

1. A strictement parler la deuxième recouvre les deux cas, d'après la convention précisée dans la définition, mais nous les séparons pour mieux distinguer les cas.

- (i) F est borné,
- (ii) la fonctionnelle J est coercive sur F .

Le minimum de J sur F est alors atteint, i.e.

$$\exists u \in F, J(u) = \min_{v \in F} J.$$

Si J est strictement convexe et F est convexe, ce minimiseur est unique.

Démonstration. Si le fermé F est borné, alors il est compact non vide, le minimum est donc atteint sur F .

Si F n'est pas borné et J est coercive sur F , on considère $x_0 \in F$. Par coercivité il existe M tel que

$$\forall x \in F, \|x\| > M, J(x) > J(x_0).$$

L'ensemble $F_M = \{x \in F, \|x\| \leq M\} = F \cap \overline{B}(0, M)$ est compact, le minimum de J sur F_M est donc atteint en un point $x \in F_M$, qui est aussi un minimiseur sur F par construction. \square

Dans le cas d'une fonctionnelle non convexe, l'unicité d'un minimiseur n'est pas assurée, cependant l'occurrence de minimiseurs multiples n'est pas générique². Plus que la non-unicité du minimiseur, la conséquence essentielle de la non-convexité d'une fonctionnelle est le fait qu'il puisse exister plusieurs minima locaux. En conséquence, les conditions nécessaires d'optimalité abordées dans la suite *ne sont pas suffisantes*. Dans le cas où plusieurs minima locaux co-existent, il faut comparer les valeurs respectives de tous ces minimiseurs pour déterminer le minimiseur global, s'il existe³.

Proposition 13.9. (Conditions nécessaires d'optimalité)

Soit U un ouvert d'un espace vectoriel normé E , et J une fonctionnelle différentiable sur U . Si u est un minimum local de J sur U , alors $DJ(u) = 0$.

Démonstration. Pour tout $h \in H$, $u + \varepsilon h$ est dans U pour ε suffisamment petit, on a donc

$$J(u + \varepsilon h) = J(u) + \varepsilon \langle DJ(u), h \rangle + o(\varepsilon) \geq J(u),$$

d'où $\langle DJ(u), h \rangle \geq 0$ pour tout h . Comme on peut prendre h et $-h$ dans l'inégalité, cela implique $\langle DJ(u), h \rangle = 0$. \square

Remarque 13.10. Lorsque la fonctionnelle J est l'énergie mécanique d'un système qui dépend d'un certain nombre de paramètres, la condition $DJ(u) = 0$ exprime en général l'équilibre des forces. Considérer par exemple un point situé sur l'axe réel, relié en 0 et 1 à des ressorts de longueur au repos nulles et de raideurs k_1 et k_2 . L'énergie élastique s'écrit

$$J(x) = \frac{1}{2}k_1x^2 + \frac{1}{2}k_2(1-x)^2.$$

La condition nécessaire d'optimalité s'écrit $-\nabla J(x) = 0$, soit $-k_1x - k_2(1-x) = 0$, qui exprime la somme des forces exercées par 0 et 1, respectivement, au travers des ressorts.

La condition d'annulation de la différentielle assure le caractère minimisant sous certaines hypothèses de convexité :

Proposition 13.11. (Condition suffisante d'optimalité)

Soit U un ouvert convexe d'un espace vectoriel normé E , et J une fonctionnelle différentiable sur U . On suppose que J est convexe sur U . Si $DJ(u) = 0$, alors u est un minimiseur global de J sur U .

2. Considérer par exemple l'espace des polynômes de degré d pair plus grand que 4, à coefficient directeur égal à 1, identifié à une partie de \mathbb{R}^d . L'ensemble des coefficients pour lesquels le minimiseur est unique est un ouvert dense de \mathbb{R}^d .

3. Il peut y avoir des minimiseurs locaux sans qu'aucun d'entre eux ne soit global, considérer par exemple la fonction $1/x + \sin x$ sur $]0, +\infty[$.

Démonstration. Soit $v \in U$. On a, pour tout $\theta \in]0, 1]$,

$$J((1 - \theta)u + \theta v) \leq (1 - \theta)J(u) + \theta J(v),$$

d'où

$$J(v) - J(u) \geq \frac{1}{\theta} (J(u + \theta(v - u)) - J(u))$$

qui tend vers $\langle DJ(u) | v - u \rangle = 0$ quand ε tend vers 0. \square

On se reportera l'exercice 13.13, page 284, pour une condition nécessaire et suffisante d'optimalité sous contrainte convexe.

L'essentiel de ce qui suit est consacré à la notion de multiplicateur de Lagrange, variable auxiliaire 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 précédemment, toutes les directions étaient permises, ce qui a permis de conclure à l'annulation de la différentielle. Dans le cas constraint, seules les variations qui ne font pas sortir de l'ensemble sont autorisées. La proposition ci-dessous est un cas particulier d'une propriété plus générale démontrée plus loin (théorème 13.24, ou plus précisément corollaire 13.25, page 268), mais permet de traiter un grand nombre de situations.

Proposition 13.12. (Conditions nécessaires d'optimalité, cas avec contraintes)

Soit J une fonctionnelle C^1 sur un ouvert U de $V = \mathbb{R}^n$. On suppose que J admet un minimum local sur $U \cap K$ en u , avec

$$K = u_0 + \ker B, \quad B \in \mathcal{M}_{mn}(\mathbb{R}).$$

Il existe alors $\lambda \in \mathbb{R}^m$ tel que

$$\begin{cases} \nabla J(u) + B^* \lambda &= 0 \\ Bu &= Bu_0 \end{cases}$$

Démonstration. Soit $v \in \ker B$. Pour 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 donc⁴ $\nabla J(u) \in (\ker B)^\perp = \text{Im } B^*$, d'où le résultat. \square

Remarque 13.13. Tant que le nombre de contraintes reste fini, la proposition précédente s'applique immédiatement au cas où V est un espace de Hilbert, qui peut lui être de dimension infinie, il suffit de remplacer la matrice B exprimant les contraintes par une application qui envoie V dans \mathbb{R}^m :

$$B : v \mapsto (\langle \varphi_i, v \rangle)_i,$$

où les φ_i sont éléments de V' . L'image de B étant fermée (c'est un sous-espace vectoriel de \mathbb{R}^m , on a $(\ker B)^\perp = \text{Im } B^*$, d'où l'existence du vecteur λ de multiplicateurs de Lagrange. Le cas où à la fois la dimension de V et le nombre de contraintes sont infinis est plus délicat, nous renvoyons à la section 13.7 pour plus de détails.

Le cas de contraintes d'égalité dans le cas non linéaire est beaucoup plus délicat. La difficulté vient du fait que si l'on considère une variation $u + h$ d'un point u admissible, il est délicat de caractériser l'admissibilité de $u+h$. On peut néanmoins énoncer une propriété permettant de définir des multiplicateurs de Lagrange dans ce contexte, en dimension finie, pour un nombre fini de contraintes (avec condition d'indépendance des gradients au point considéré), voir proposition 13.60, page 279.

4. On se reportera à l'exercice 13.10, page 283, qui propose une démonstration de $(\ker B)^\perp = \text{Im } B^*$ dans l'esprit des développements qui vont suivre, sans utiliser le théorème du rang, et fait le lien entre cette propriété matricielle et un résultat général d'analyse fonctionnelle sur des formes linéaires d'un espace vectoriel normé.

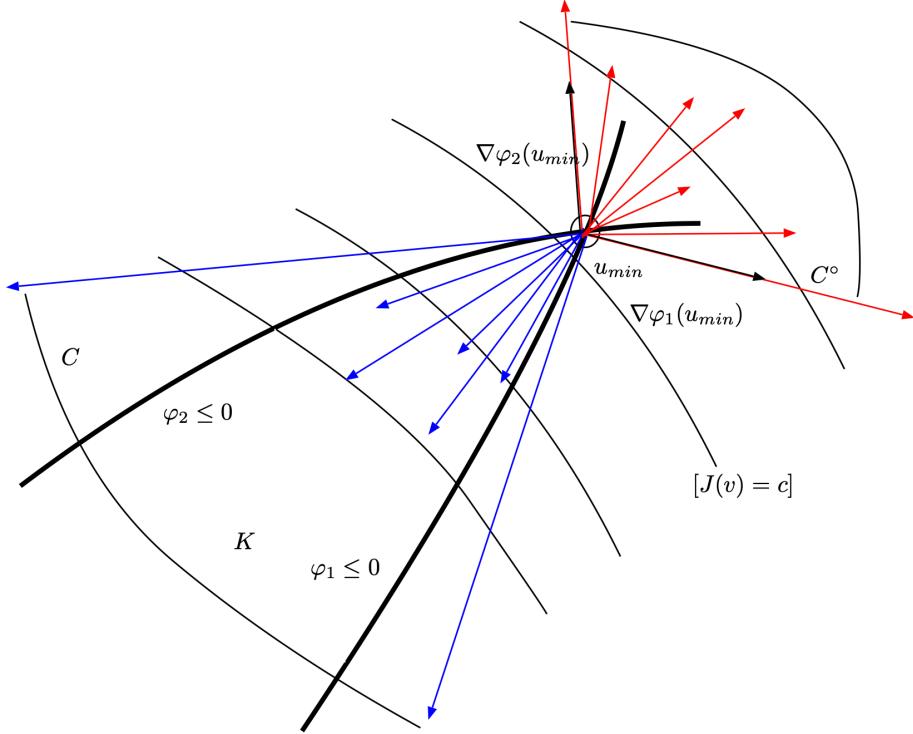


FIGURE 13.1 – Cône des directions admissibles et son polaire

13.3 Contraintes unilatérales

Cette section établit, dans l'esprit de la proposition 13.12, des conditions nécessaires d'optimalité pour des contraintes d'inégalité. La figure 13.1 illustre la démarche d'ensemble, que nous décrivons ici informellement. Le point u_{min} est supposé réaliser le minimum d'une fonctionnelle dont sont tracées quelques courbes isovaleurs (la fonction est supposée ici décroître quand on va vers la droite), sur un ensemble admissible K , défini ici comme l'intersection entre deux ensembles définis implicitement : $\{v, \varphi_1(v) \leq 0\}$ et $\{v, \varphi_2(v) \leq 0\}$. Comme dans la preuve de la proposition 13.12, on écrit le caractère minimisant de u_{min} en effectuant des variations admissibles. Les directions correspondantes appartiennent au cône C dit des directions admissibles (en bleu). La caractérence minimisant de u_{min} implique que le produit scalaire du gradient de J avec toutes ces directions soit positif, c'est-à-dire que $-\nabla J$ soit dans ce que nous allons définir comme le cône polaire C° (en rouge) de C . Nous établirons une version unilatérale de la propriété matricielle $(\ker B)^\perp = \text{Im } B^*$, qui assurera que ce cône polaire est le cône convexe engendré par $\nabla \varphi_1(u_{min})$ et $\nabla \varphi_2(u_{min})$, c'est à dire l'ensemble des combinaisons positives de ces deux vecteurs, ce qui permettra d'écrire, dans un cas comme celui-là,

$$\nabla J(u_{min}) + \lambda_1 \nabla \varphi_1(u_{min}) + \lambda_2 \nabla \varphi_2(u_{min}) = 0.$$

H désigne dans la suite un espace de Hilbert.

Definition 13.14. (Cône)

On appelle cône de sommet 0 une partie C de H telle que

$$u \in C \implies \lambda u \in C \quad \forall \lambda > 0,$$

ce qui peut aussi s'exprimer $\mathbb{R}_+^* C \subset C$. On appellera cône de sommet $s \in H$ un ensemble C tel que $C - s$ est un cône de sommet 0.

Sauf indication contraire, les cônes qui nous considérerons dans la suite seront de sommet l'origine 0.

Definition 13.15. (Polaire d'un ensemble)

Soit C une partie de H , on définit le polaire de C comme

$$C^\circ = \{v \in H, \langle v | u \rangle \leq 0 \quad \forall u \in C\}.$$

Noter que dans le cas où C est un sous-espace vectoriel de H , l'ensemble C° est simplement l'orthogonal de C .

On notera également que dans la définition ci-dessus, C intervient comme un ensemble de contraintes, de telle sorte que l'application $C \mapsto C^\circ$, de 2^H dans lui-même, est *décroissante* pour l'inclusion.

Exercice 13.3. On se place sur $H = \ell^2$. Dans les deux cas suivants, montrer que C est un cône convexe fermé, et identifier son polaire.

(i) C est l'ensemble des suites de ℓ^2 à termes positifs ou nuls.

(ii) ($\bullet\bullet$) C est l'ensemble des suites de ℓ^2 décroissantes.

Proposition 13.16. Pour tout $C \subset H$, C° est un cône convexe fermé.

Les caractères cônique, convexe, et fermé, étant stables par intersection, on peut définir la notion d'*enveloppe*.

Definition 13.17. (Enveloppe convexe conique, enveloppe convexe conique fermée)

Soit $C \subset H$. On appelle enveloppe convexe conique de C le plus petit cône convexe qui contient C , i.e. l'intersection des cônes convexes qui contiennent C . On la note $co(C)$. On appelle enveloppe conique fermée le plus petit cône convexe fermé qui contient C . Il s'agit de l'adhérence de $co(C)$, que l'on notera en conséquence $\overline{co}(C)$.

Proposition 13.18. Soit $C \subset H$ une partie de H , C° son polaire, et $C^{\circ\circ} = (C^\circ)^\circ$ son bipolaire. Alors $C^{\circ\circ}$ est l'enveloppe convexe fermée conique de C . En particulier, si C est un cône convexe fermé (de sommet 0), alors $C^{\circ\circ} = C$.

Démonstration. L'inclusion $C \subset C^{\circ\circ}$ est immédiate : tout v dans C a un produit scalaire négatif contre tout élément de C° , il est donc dans $C^{\circ\circ}$. Comme $C^{\circ\circ}$ est un cône convexe fermé, l'inclusion demeure par passage à l'enveloppe convexe fermé conique.

Si l'inclusion $\overline{co}(C) \subset C^{\circ\circ}$ est stricte, il existe $z \in C^{\circ\circ}$ qui n'appartient pas à $\overline{co}(C)$. On peut alors, d'après⁵ le théorème de Hahn-Banach 18.14, page 361, séparer le convexe fermé $\overline{co}(C)$ de $\{z\}$: il existe h tel que

$$\langle h | v \rangle \leq \alpha < \langle h | z \rangle \quad \forall v \in \overline{co}(C).$$

Comme v décrit un cône de sommet 0, $\langle h | v \rangle$ 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 $h \in \overline{co}(C)^\circ = C^\circ$. Par ailleurs le maximum de $\langle h | v \rangle$ est 0, et donc $\alpha \geq 0$, d'où $\langle h | z \rangle > 0$ ce qui est absurde car $h \in C^\circ$ et $z \in C^{\circ\circ}$. \square

Avant d'écrire la propriété principale de cette section, qui donnera des conditions nécessaires d'optimalité sous contraintes, nous établissons le caractère fermé d'un cône engendré par un nombre fini de vecteur, i.e. un ensemble du type

$$C = \left\{ \sum_{i=1}^m \lambda_i g_i, \lambda_i \geq 0 \quad \forall i = 1, \dots, m \right\}, \tag{13.1}$$

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

⁵ Il s'agit ici du "petit" théorème de Hahn-Banach, c'est à dire dans un cadre Hilbertien, qui ne nécessite pas l'axiome du choix, et peut se démontrer en quelques lignes à l'aide de la projection sur un convexe fermé.

vecteurs est fermée, il est un peu plus délicat⁶ de démontrer une telle propriété de fermeture pour le cône (convexe) engendré par une telle famille.

C'est l'objet du lemme suivant :

Lemme 13.19. Le cône convexe C défini par (13.1) est fermé.

Démonstration. On raisonne par récurrence sur le nombre de vecteurs g_i . Supposons que tout cône convexe engendré par m vecteurs est fermé, et considérons une famille de $m+1$ vecteurs g_i .

Si les g_i forment une famille libre, on se place dans l'espace vectoriel W engendré par les g_i , et l'on introduit

$$G : \lambda \in \mathbb{R}^{m+1} \mapsto \sum_{i=1}^{m+1} \lambda_i g_i \in W.$$

Cette application est inversible par hypothèse, d'inverse G^{-1} linéaire continu (la dimension est finie). Considérons maintenant une suite $v^k = \sum \lambda_i^k g_i$ qui converge vers $v \in W$. Alors $G^{-1}v^k$ converge vers $G^{-1}v$, i.e. le vecteur λ^k converge vers un vecteur λ de \mathbb{R}^{m+1} , dont toutes les composantes sont positives ou nulles par continuité, on a donc bien $v \in C$.

Si maintenant la famille est liée, il existe μ_1, \dots, μ_{m+1} , non tous nuls, tels que

$$\sum_{i=1}^{m+1} \mu_i g_i = 0. \quad (13.2)$$

On considère une suite dans C qui converge vers $v \in H$:

$$\sum_{i=1}^{m+1} \lambda_i^k g_i \longrightarrow v.$$

Il s'agit de montrer que v est combinaison positive des g_i . On suppose (quitte à prendre la combinaison opposée) que l'un des coefficients de la combinaison non triviale (13.2) est strictement négatif. On considère alors, pour tout k , le plus grand $\beta^k \geq 0$ tel que $\lambda_i^k + \beta^k \mu_i \geq 0$ pour tout $1 \leq i \leq m+1$. 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 m vecteurs $(g_i)_{i \neq i_0}$ (d'après l'hypothèse de récurrence), donc dans C . \square

Lemme 13.20. (Lemme de Farkas)

Soient $(g_i)_I$ une famille finie de vecteurs d'un espace de Hilbert H . On introduit

$$C = \{h \in H, \langle g_i | h \rangle \leq 0 \quad \forall i \in I\}.$$

Le polaire de C est le cône engendré par les g_i :

$$C^\circ = \left\{ \sum_{i \in I} \lambda_i g_i, \lambda_i \geq 0 \quad \forall i \right\}.$$

Démonstration. L'ensemble C est de façon évidente le cône polaire de

$$F = \left\{ \sum_{i \in I} \lambda_i g_i, \lambda_i \geq 0 \quad \forall i \right\},$$

6. On peut d'ailleurs vérifier que, dans le cas d'une famille infinie (même dans un espace de dimension finie), le résultat est faux en général (voir exercice 13.15, page 284).

L'ensemble F est un cône convexe, fermé d'après le lemme 13.19. D'après la proposition 13.18), il s'identifie donc à son bipolaire : on a donc $C^\circ = F^{\circ\circ} = F$. \square

Remarque 13.21. On peut voir ce lemme de Farkas comme une version unilatérale du lemme dit *des noyaux*⁷ qui est elle-même une généralisation de la propriété $(\ker B)^\perp = \text{Im } B^*$ pour les matrices. Si l'on interprète cette propriété en considérant la matrice B dont les lignes sont les g_i écrits dans la base canonique de \mathbb{R}^n , elle assure que si un vecteur g est orthogonal à tout vecteur h lui-même orthogonal à des vecteurs g_1, \dots, g_m de \mathbb{R}^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 $\text{vect}(g_i)$.

Contraintes d'inégalité

On s'intéresse ici à la minimisation de fonctionnelles sur des ensembles du type

$$K = \{ v \in H, \varphi_i(v) \leq 0, i = 1, \dots, m \}, \quad (13.3)$$

où les φ sont des fonctionnelles de H dans \mathbb{R} , continûment différentiables.

Definition 13.22. (Contraintes actives)

On dit que la contrainte i est active en $u \in H$ dès que $\varphi_i(u) = 0$. On note I_u l'ensemble des i tels que la contrainte i est active en u .

Definition 13.23. (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

$$\langle \nabla \varphi_i(u) | h \rangle < 0$$

ou simplement $\langle \nabla \varphi_i(u) | h \rangle \leq 0$ si φ_i est affine, pour tout $i \in I_u$.

On notera que, si toutes les contraintes sont affines, on peut prendre $h = 0$ dans la définition ci-dessus, ce qui assure trivialement les inégalités larges. Dans le cas de contraintes affines, ces contraintes sont *toujours* qualifiées.

Théorème 13.24. (Conditions nécessaires d'optimalité sous contrainte)

Soit J une fonctionnelle C^1 définie sur un ouvert U d'un espace de Hilbert H , et u un minimiseur local de J sur $U \cap K$ défini par (13.3), où les φ_i sont continûment différentiables sur U . On suppose que les contraintes sont qualifiées en u . Il existe alors $\lambda_1, \lambda_2, \dots, \lambda_m \geq 0$ tels que

$$\nabla J(u) + \sum_{i=1}^m \lambda_i \nabla \varphi_i(u) = 0,$$

avec⁸ $\sum \varphi_i(u) \lambda_i = 0$.

Démonstration. Notons en premier lieu que, si toutes les contraintes sont affines, il est possible que le h de la propriété de qualification soit nul. Si seul ce vecteur nul réalise les inégalités, cela signifie que le cône polaire de la famille des $\nabla \varphi_i$ est réduit au vecteur nul, donc que son double polaire est l'espace entier, et donc que

$$\left\{ \sum \lambda_i \nabla \varphi_i(u), \lambda_1, \dots, \lambda_m \geq 0 \right\} = H.$$

7. Soit X un espace vectoriel, et $\varphi, \varphi_1, \dots, \varphi_n$ des formes linéaires sur X , telles que

$$\cap \ker \varphi_i \subset \ker \varphi.$$

Alors φ est combinaison linéaire des φ_i (voir exercice 13.10, page 283).

8. On notera que, du fait que les composantes de $\varphi(u)$ (resp. λ) sont négatives (resp. positives), cette identité implique que $\lambda_i = 0$ dès que la contrainte i n'est pas saturée.

Alors $-\nabla J(u)$, comme tout autre vecteur, peut s'écrire comme combinaison linéaire positive de ces vecteurs. Même si elle ne présente *aucun intérêt*, puisqu'elle ne donne aucune information sur $\nabla J(u)$, la conclusion est donc alors bien vérifiée. Cela correspond au cas où le minimiseur est un point isolé de K , au voisinage duquel on ne peut faire aucune variation admissible.

Nous nous proposons maintenant de montrer que $\nabla J(u)$ a un produit scalaire positif ou nul contre tout élément du cône

$$C = \{h \in H, \langle \nabla \varphi_i | h \rangle \leq 0 \quad \forall i \in I_u\}.$$

Le cas de figure $h = 0$ ayant été traité précédemment, on considère maintenant le cas où il existe un $h \neq 0$ vérifiant $\langle \nabla \varphi_i(u) | h \rangle < 0$ pour toute contrainte i active en u (avec éventuellement égalité pour une contrainte affine). Pour $t > 0$ suffisamment petit, on a $u + th \in K \cap U$, et donc

$$J(u + th) \geq J(u) \quad \forall t \in [0, t^*[,$$

d'où

$$J(u) + t \langle \nabla J(u) | h \rangle + o(t) \geq \nabla J(u),$$

et donc nécessairement

$$\langle \nabla J(u) | h \rangle \geq 0.$$

Soit maintenant h tel que l'on ait simplement l'inégalité au sens large $\langle \nabla \varphi_i(u) | h \rangle \leq 0$ pour l'ensemble des contraintes. Montrons que la propriété reste vérifiée. En effet, considérons un h^* pour lequel on a les inégalités strictes (ou larges pour les contraintes affines), on préserve les inégalités strictes pour $(1 - \varepsilon)h + \varepsilon h^*$, d'où

$$\langle \nabla J(u) | ((1 - \varepsilon)h + \varepsilon h^*) \rangle \geq 0,$$

et donc $\langle \nabla J(u) | h \rangle \geq 0$ par passage à la limite $\varepsilon \rightarrow 0$.

Le vecteur $-\nabla J$ est donc dans C° , polaire de

$$C = \{h \in H, \langle \nabla \varphi_i | h \rangle \leq 0 \quad \forall i \in I_u\}.$$

Le vecteur $-\nabla J(u)$ est donc dans le cône convexe fermé

$$\left\{ \sum_{i \in I_u} \lambda_i \nabla \varphi_i(u), \lambda_i \geq 0 \right\}.$$

d'après le lemme de Farkas 13.20.

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. \square

Dans le cas de contrainte d'égalité affines, on retrouve la propriété déjà démontrée, conséquence de $(\ker B)^\perp = \text{Im } B^*$ (voir proposition 13.12). On peut voir de fait ce qui précède comme une généralisation unilatérale de cette propriété matricielle (voir remarque 13.21, page 267).

Corollaire 13.25. (Contraintes d'égalité affines)

Soit J une fonctionnelle C^1 définie sur un ouvert U de H , et u un minimiseur local de J sur $U \cap K$, avec

$$K = \{v \in H, \varphi_i(v) = 0, i = 1, \dots, m\},$$

où les φ_i sont des fonctions *affines*. Il existe alors $\lambda_1, \lambda_2, \dots, \lambda_m$ tels que

$$\nabla J(u) + \sum_{i=1}^m \lambda_i \nabla \varphi_i = 0.$$

Démonstration. On écrit simplement chaque contrainte d'égalité comme deux contraintes d'inégalité mutuellement opposées. \square

Le lien entre cette proposition et la proposition 13.12 peut être explicité en considérant l'écriture matricielle B des contraintes. Chaque ligne de B contient alors les coefficients de $\nabla \varphi_i$, et B^* est ainsi l'écriture matricielle de

$$\lambda \longmapsto \sum_{i=1}^m \lambda_i \nabla \varphi_i.$$

Remarque 13.26. Dans le cas de contraintes affines, on peut bien sûr panacher entre des contraintes d'égalité et des contraintes d'inégalité, l'écriture de la propriété correspondante est laissée en exercice.

13.4 Point-selle, théorème de Kuhn et Tucker

Lemme 13.27. Soient V et Λ deux *ensembles*, et $L(\cdot, \cdot)$ une application de $V \times \Lambda$ dans \mathbb{R} . On définit

$$G(\mu) = \inf_{v \in V} L(v, \mu) \in [-\infty, +\infty[, \quad F(v) = \sup_{\mu \in \Lambda} L(v, \mu) \in]-\infty, +\infty]. \quad (13.4)$$

On a alors

$$G(\mu) \leq F(v) \quad \forall \mu \in \Lambda, v \in V.$$

Par suite, s'il existe u et λ tels que $G(\lambda) = F(u)$, alors

$$G(\lambda) = \max G = \min F = F(u) = L(u, \lambda).$$

Démonstration. On écrit simplement, pour tout $\mu \in \Lambda$, tout $v \in V$,

$$G(\mu) \leq L(v, \mu) \leq F(v).$$

ce qui conclut la démonstration. \square

Definition 13.28. 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 13.29. (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, λ) est un point selle de L (sur $V \times \Lambda$) si

$$L(u, \mu) \leq L(u, \lambda) \leq L(v, \lambda) \quad \forall \mu \in \Lambda, v \in V.$$

Proposition 13.30. Soient V et Λ deux *ensembles*, $L(\cdot, \cdot)$ une application de $V \times \Lambda$ dans \mathbb{R} , et G et F définies par (13.4). Les assertions suivantes sont équivalentes :

- (i) $L(\cdot, \cdot)$ admet un point-selle (u, λ) (Def. 13.29)
- (ii) Il existe $u \in V$ et $\lambda \in \Lambda$ tels que $G(\lambda) = F(u) = L(u, \lambda)$.

Démonstration. (i) \implies (ii) Comme (u, λ) est point-selle, on a $L(u, \lambda) \leq L(v, \lambda)$ pour tout v , d'où $G(\lambda) = L(u, \lambda)$. On démontre de la même manière $F(u) = L(u, \lambda)$.

(ii) \implies (i) On suppose maintenant

$$G(\lambda) = F(u) = L(u, \lambda).$$

On a

$$L(u, \lambda) = F(u) = \sup_{\mu} L(u, \mu),$$

d'où $L(u, \lambda) \geq L(u, \mu)$ pour tout μ dans Λ . On a de même $L(u, \lambda) = G(\lambda) = \inf L(v, \lambda)$. \square

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 13.31. (Lagrangien)

Soit J une fonctionnelle d'un ensemble V dans \mathbb{R} , et K un ensemble défini par m contraintes d'inégalité

$$K = \{v \in V, \varphi_i(v) \leq 0, \forall i, 1 \leq i \leq m\}$$

Le Lagrangien associé au problème de minimisation de J sur K est défini par

$$(v, \mu) \in V \times \mathbb{R}_+^m \longmapsto L(v, \mu) = J(v) + \sum_{i=1}^m \mu_i \varphi_i(v). \quad (13.5)$$

Conformément à la définition 13.29, avec $\Lambda = \mathbb{R}_+^m$, on dira que $(u, \lambda) \in V \times \mathbb{R}_+^m$ est point-selle du Lagrangien défini par (13.5) si

$$L(u, \mu) \leq L(u, \lambda) \leq L(v, \lambda) \quad \forall \mu \in \mathbb{R}_+^m, v \in V.$$

Chaque contrainte d'égalité pouvant s'écrire comme deux contraintes d'inégalité, on peut toujours se ramener à un Lagrangien limité aux contraintes unilatérales (en dédoublant les multiplicateurs de Lagrange associés aux contraintes d'égalité).

Pour exprimer le lien entre point-selle et propriétés de minimisation, nous nous limiterons en revanche au cas d'inégalités, le cas des contraintes d'égalité est laissé en exercice au lecteur.

Proposition 13.32. On considère une fonctionnelle J d'un ensemble V dans \mathbb{R} , et l'on suppose que le Lagrangien associé au problème de minimisation de J sur

$$K = \{v \in V, \varphi_i(v) \leq 0, \forall i, 1 \leq i \leq m\}$$

admet un point-selle $(u, \lambda) \in V \times \mathbb{R}_+^m$, c'est à dire que

$$J(u) + \sum_{i=1}^m \mu_i \varphi_i(u) \leq J(u) + \sum_{i=1}^m \lambda_i \varphi_i(u) \leq J(v) + \sum_{i=1}^m \lambda_i \varphi_i(v) \quad \forall \mu \in \mathbb{R}_+^m, v \in V.$$

Alors u minimise J sur K , et l'on a $\lambda_i \varphi_i(u) = 0$ pour tout i .

Si V est un ouvert d'un espace de Hilbert, et que les fonctions $J, \varphi_1, \dots, \varphi_m$ sont dérivables, alors on a de plus

$$\nabla J(u) + \sum_{i=1}^m \lambda_i \nabla \varphi_i(u) = 0.$$

Démonstration. D'après la première inégalité du point-selle, la quantité $\sum \mu_i \varphi_i(u)$ est bornée sur \mathbb{R}_+^m , 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 $\mu = 0$, l'inégalité $0 \leq \sum \lambda_i \varphi_i(u)$. Comme il s'agit d'une somme de termes négatifs ou nuls, tous les termes sont nuls : $\lambda_i \varphi_i(u) = 0$, et ainsi $\lambda_i = 0$ dès que $\varphi_i(u) < 0$ (i.e. quand la contrainte n'est pas activée). On utilise maintenant la seconde inégalité :

$$J(u) = J(u) + \sum_{i=1}^m \lambda_i \varphi_i(u) \leq J(v) + \sum_{i=1}^m \lambda_i \varphi_i(v)$$

qui est en particulier inférieur à $J(v)$ pour tout $v \in K$.

Si maintenant V 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

$$v \longmapsto J(v) + \sum_{i=1}^m \lambda_i \varphi_i(v)$$

est régulière, et le fait que u la minimise implique que son gradient soit nul en u (proposition 13.9), ce qui conclut la démonstration. \square

Théorème 13.33. (Kuhn et Tucker)

On considère un ouvert convexe U de \mathbb{R}^n , J convexe différentiable sur U , et l'ensemble admissible

$$K = \{v, \varphi_i(v) \leq 0, 1 \leq i \leq m\}.$$

On suppose les φ_i différentiables et convexes sur U .

On suppose qu'il existe $(u, \lambda) \in U \times \mathbb{R}^m$ tel que

$$\left| \begin{array}{rcl} \nabla J(u) + \sum_{i=1}^m \lambda_i \nabla \varphi_i(u) & = & 0 \\ \varphi_i(v) & \leq & 0 \quad \forall i \\ \lambda & \geq & 0 \\ \sum_{i=1}^m \lambda_i \varphi_i(u) & = & 0 \end{array} \right. \quad (13.6)$$

Le couple (u, λ) est alors point-selle du Lagrangien

$$L(v, \mu) = J(v) + \sum_{i=1}^m \mu_i \varphi_i(v)$$

sur $U \times \mathbb{R}_+^m$ et u minimise donc J sur $U \cap K$ (proposition 13.32).

Démonstration. De la dernière condition de (13.7) on déduit que u minimise la fonctionnelle (convexe)

$$v \mapsto J(v) + \lambda \cdot \varphi(v),$$

sur le convexe U (voir proposition 13.11, page 262). On en déduit la seconde inégalité du point-selle. On a par ailleurs, comme les $\varphi_i(u)$ sont négatifs,

$$J(u) + \mu \cdot \varphi(u) \leq J(u)$$

pour tout $\mu \in \mathbb{R}_+^m$. Mais on a aussi $J(u) = J(u) + \lambda \cdot \varphi(u)$ par hypothèse (deuxième de (13.7)), d'où la première inégalité du point-selle. \square

Corollaire 13.34. (Contraintes affines)

Le théorème précédent s'applique au cas de contraintes d'égalité dès que les contraintes sont affines. Plus précisément, Si l'on considère un ouvert convexe U de \mathbb{R}^d , J convexe différentiable sur U , et l'ensemble admissible

$$K = \{v, \varphi_i(v) = 0, 1 \leq i \leq m\},$$

où les φ_i sont *affines*. On suppose qu'il existe $(u, \lambda) \in (U \cap K) \times \mathbb{R}^m$ tel que

$$\nabla J(u) + \sum_{i=1}^m \lambda_i \nabla \varphi_i(u) = 0. \quad (13.7)$$

Le couple (u, λ) est alors point-selle du Lagrangien $L(v, \mu) = J(v) + \mu \cdot \varphi(v)$ sur $U \times \mathbb{R}_+^m$ et u minimise ainsi J sur $U \cap K$.

Démonstration. Il suffit d'écrire chaque contrainte d'égalité comme deux contraintes d'inégalité. Plus précisément, si l'on sépare en I^+ et I^- les indices correspondant à des λ_i respectivement positifs et négatifs, on peut écrire

$$\sum_{i=1}^m \lambda_i \nabla \varphi_i(u) = \sum_{i \in I^+} \lambda_i \nabla \varphi_i(u) + \sum_{i \in I^-} (-\lambda_i) \nabla (-\varphi_i(u)),$$

on est donc ramené à la situation du théorème 13.33 avec les contraintes d'inégalité associées aux fonctions $\varphi_1, \dots, \varphi_n, -\varphi_1, \dots, -\varphi_n$. \square

Synthèse des sections précédentes

On considère une fonctionnelle différentiable sur un ouvert U d'un espace de Hilbert H , et le problème consistant à minimiser J sur $U \cap K$, où K est défini par

$$K = \{v \in U, \varphi_i(v) \leq 0 \quad \forall i, j, 1 \leq i \leq m\}$$

avec φ différentiable. On appelle parfois formulation point-selle (ou mixte) le problème consistant à trouver $(u, \lambda) \in U \cup \mathbb{R}_+^n$ tel que

$$\left| \begin{array}{rcl} \nabla J(u) + \sum_{i=1}^m \lambda_i \nabla \varphi_i(u) & = & 0 \\ \varphi_i(v) & \leq & 0 \quad \forall i \\ \lambda & \geq & 0 \\ \sum_{i=1}^m \lambda_i \varphi_i(u) & = & 0 \end{array} \right. \quad (13.8)$$

Le théorème 13.24 (conditions nécessaires d'optimalité) assure que, si u est minimiseur local de J sur $U \cap K$, et si les contraintes sont qualifiées en u , alors il existe λ tel que (u, λ) soit solution du problème (13.8).

Réiproquement, le théorème 13.33 assure que, si J et les φ_i sont convexes, et si (u, λ) est solution de 13.8, alors u est minimiseur global de J sur $U \cap K$.

Ces deux propriétés sont parfois regroupées dans un même théorème énonçant une équivalence entre le fait d'être solution de (13.8) et de minimiser J sur $U \cap K$, mais cette équivalence ne peut être établie que si l'on fait l'ensemble des hypothèses ci-dessus, en particulier la qualification des contraintes (qui n'est pertinente que pour la condition nécessaire), et la convexité (qui n'est pertinente que pour la condition suffisante).

Précisons pour terminer que, dans le cas très étudié d'une fonctionnelle quadratique

$$J(v) = \langle Av | v \rangle - \langle b | v \rangle$$

et de contraintes affines, on peut introduire une matrice B encodant ces contraintes :

$$K = \{v \in V = \mathbb{R}^n, Bv \leq z\}.$$

On peut écrire la formulation de type point-selle (ou mixte) associée au problème sous la forme

$$\left| \begin{array}{rcl} Au + B^*p & = & b \\ Bu & \leq & z \\ p & \geq & 0 \\ \langle p | Bu \rangle & = & 0. \end{array} \right. \quad (13.9)$$

13.5 Formalisme de l'analyse non lisse, sous-différentiels

Nous présentons dans cette section un cadre formel qui permet d'exprimer de façon plus concise, et potentiellement plus générale, les propriétés des sections précédentes. Cette approche permet en particulier de remplacer un problème de minimisation sous contrainte par un problème de minimisation sans contrainte, en rajoutant simplement à la fonction à minimiser une fonction qui prend la valeur $+\infty$ à l'extérieur de l'ensemble admissible. La fonction obtenue étant non lisse, la notion de gradient sera remplacé par la notion plus générale de *sous-différentiel*.

On se place dans un espace de Hilbert H , et les fonctions que nous considérons sont à valeurs dans $\mathbb{R} \cup \{+\infty\}$. La valeur $+\infty$ ne sera utilisée qu'en lien avec la loi ‘+’, au sens où l'on considérera que

$$(+\infty) + (+\infty) = +\infty, \quad c + (+\infty) = +\infty \quad \forall c \in \mathbb{R},$$

et avec la relation d'ordre usuelle sur \mathbb{R} , étendue canoniquement à $\mathbb{R} \cup \{+\infty\}$:

$$+\infty \leq +\infty, \quad c \leq +\infty \quad \forall c \in \mathbb{R}.$$

Definition 13.35. (Domaine d'une fonction, fonction propre)

Soit $\varphi : H \rightarrow \mathbb{R} \cup \{+\infty\}$, on appelle domaine de φ l'ensemble

$$D(\varphi) = \{u \in H, \varphi(u) < +\infty\}.$$

On dira qu'une fonction est *propre* si son domaine est non vide.

On vérifie immédiatement que le domaine d'une fonction convexe est nécessairement convexe :

Proposition 13.36. Si $\varphi : H \rightarrow \mathbb{R} \cup \{+\infty\}$ est convexe, son domaine est convexe.

Definition 13.37. (Épigraphe)

Soit φ une fonction sur H à valeurs dans $\mathbb{R} \cup \{+\infty\}$. On appelle *épigraphe* de φ l'ensemble

$$\text{epi}(\varphi) = \{(x, t) \in H \times \mathbb{R}, \varphi(x) \leq t\}.$$

Definition 13.38. (Semi-continuité inférieure)

Soit φ une fonction sur H . On dit que φ est semi-continue inférieurement si son épigraphe est fermé.

Proposition 13.39. Soit φ une fonction sur H . La fonction φ est s.c.i. si et seulement si

$$\forall x \in D(\varphi), \forall \varepsilon > 0, \exists \eta > 0, \forall y \in B(x, \eta), \varphi(y) \geq \varphi(x) - \varepsilon,$$

ou, de façon équivalente, si et seulement si, pour $x \in D(\varphi)$, $x_n \rightarrow x$,

$$\varphi(x) \leq \liminf \varphi(x_n).$$

Démonstration. On suppose la propriété ci-dessus vérifiée. Soit (x_n, t_n) une suite de l'épigraphe qui converge vers (x, t) . On a

$$\varphi(x) \leq \liminf \varphi(x_n) \leq \liminf t_n \leq t,$$

d'où $(x, t) \in \text{epi}(\varphi)$.

Réciproquement, supposons que la propriété ne soit pas vérifiée, i.e.

$$\exists \varepsilon > 0, \forall \eta > 0, \exists y \in B(x, \eta), \varphi(y) \leq \varphi(x) - \varepsilon.$$

On peut alors construire une suite x_n qui converge vers x telle que

$$\varphi(x_n) \leq \varphi(x) - \varepsilon.$$

La suite $(x_n, \varphi(x) - \varepsilon)$ est dans l'épigraphe de φ , et converge vers $(x, \varphi(x) - \varepsilon)$ qui n'y est pas, ce qui contredit le caractère fermé de l'épigraphe. \square

L'épigraphe permet de caractériser la convexité :

Proposition 13.40. Une fonction φ sur H à valeurs dans $\mathbb{R} \cup \{+\infty\}$ est convexe si et seulement si son épigraphe est convexe.

Definition 13.41. (Fonction indicatrice)

Soit K une partie de H , la fonction indicatrice de K est la fonction I_K qui vaut 0 sur K , et $+\infty$ ailleurs. On n'utilisera cette notion que pour des ensembles convexes fermés.

Proposition 13.42. Une partie $K \subset H$ est convexe si et seulement si I_K est une fonction convexe.

Proposition 13.43. Une partie $K \subset H$ est fermée si et seulement si I_K est une fonction s.c.i.

Definition 13.44. (Sous-différentiel)

Soit $\varphi : H \rightarrow \mathbb{R} \cup \{+\infty\}$ une fonction convexe, et $u \in D(\varphi)$. On appelle sous-différentiel de φ en u l'ensemble

$$\partial\varphi(u) = \{w \in H, \varphi(u) + \langle w | h \rangle \leq \varphi(u+h) \quad \forall h \in H\}.$$

On dira que φ est sous-differentiable en $u \in D(\varphi)$ si $\partial\varphi(u) \neq \emptyset$.

La définition ci-dessus est en général réservée aux fonctions convexes. Son application brutale à des fonctions non convexes conduit en effet à ce que le sous-différentiel soit vide en des points du domaine (si φ est strictement concave par exemple). On est parfois amené à considérer une définition plus souple de cette notion.

Definition 13.45. (Sous-différentiel au sens de Fréchet)

Soit $\varphi : H \rightarrow \mathbb{R} \cup \{+\infty\}$ une fonction, et $u \in D(\varphi)$. On appelle sous-différentiel au sens de Fréchet de φ en u l'ensemble

$$\partial\varphi(u) = \{w \in H, \varphi(u) + \langle w | h \rangle \leq \varphi(u+h) + o(h)\}.$$

Proposition 13.46. J une fonction convexe différentiable en $u \in H$. On a

$$\partial J(u) = \{\nabla J(u)\}$$

Démonstration. Soit $w \in \partial J$. On a

$$J(u+h) \geq J(u) + \langle w | h \rangle \text{ et } J(u+h) = J(u) + \langle \nabla J(u) | h \rangle + o(h),$$

d'où

$$\langle w - \nabla J(u) | h \rangle \leq o(h),$$

Pour tout vecteur g fixé, on écrit que l'inégalité précédente est vérifiée pour $h = \varepsilon g$, pour tout ε , ce qui implique $w = \nabla J(u)$. \square

On se reportera à l'exercice 13.19, page 285, pour quelques exemples de sous-différentiels de fonctions non lisses.

On peut montrer immédiatement une généralisation des propositions 13.9 et 13.11 au cas non lisse.

Proposition 13.47. Soit $\varphi : H \rightarrow \mathbb{R} \cup \{+\infty\}$ une fonction convexe propre. On a l'équivalence

$$u \in \arg \min_H \varphi \iff 0 \in \partial\varphi(u).$$

Démonstration. On suppose que $0 \in \partial\varphi(u)$, ce qui s'écrit

$$\varphi(u) + \langle 0 | h \rangle = \varphi(u) \leq \varphi(u+h),$$

qui exprime que u minimise φ . Réciproquement, si u minimise φ , l'inégalité ci-dessus implique que u est dans le sous-différentiel de φ . \square

La notion de sous-différentiel est à manier avec précaution. En particulier le sous-différentiel d'une somme de fonctions convexes n'est en général pas la somme (ensembliste) des sous-différentiels⁹. Cette propriété est néanmoins vérifiée si l'une des fonctions est lisse au point considéré.

Proposition 13.48. Soit $\varphi : H \rightarrow \mathbb{R} \cup \{+\infty\}$ une fonction convexe propre, $u \in D(\varphi)$, et J une fonction convexe différentiable en u . On a

$$\partial(J + \varphi)(u) = \nabla J(u) + \partial\varphi.$$

9. Considérer par exemple la fonction $\varphi = I_{[-1,0]} + I_{[0,1]}$ en 0.

Démonstration. On vérifie immédiatement que $\partial\varphi_1 + \partial\varphi_2 \subset \partial(\varphi_1 + \varphi_2)$ en toute généralité, on a donc $\partial(J + \varphi)(u) \supset \nabla J(u) + \partial\varphi$.

Considérons maintenant un élément w de $\partial(J + \varphi)$. Soit h tel que $u + h \in D(J + \varphi)$. On a, pour tout $t \in [0, 1]$,

$$J(u) + \varphi(u) + t \langle w | h \rangle \leq J(u + th) + \varphi(u + th).$$

Comme J est différentiable en u , on a

$$J(u + th) = J(u) + t \langle \nabla J(u) | h \rangle + o(t),$$

d'où

$$\frac{\varphi(u + th) - \varphi(u)}{t} \geq \langle w - \nabla J(u) | h \rangle + o(t).$$

La fonction φ étant convexe, on a

$$\varphi(u + h) \geq \frac{\varphi(u + th) - \varphi(u)}{t} + \varphi(u),$$

d'où, en combinant les deux dernières inégalités, et en faisant tendre t vers 0^+ ,

$$\langle w - \nabla J(u) | h \rangle \leq \varphi(u + h) - \varphi(u),$$

d'où l'on déduit $w - \nabla J(u) \in \partial\varphi$. □

On s'intéresse comme dans les sections précédentes à la minimisation d'une fonctionnelle J sur un ensemble admissible de type.

$$K = \{ v \in H, \varphi_i(v) \leq 0, i = 1, \dots, m \} \quad (13.10)$$

Proposition 13.49. Soit $K \subset H$ l'ensemble défini par (13.10) ci-dessus, où l'on suppose les fonctions φ_i convexes et différentiables sur H . Si l'on suppose les contraintes qualifiées en $u \in K$ (au sens de la définition 13.23), on a

$$\partial I_K(u) = co((\nabla \varphi_i)_{i=1, \dots, m}).$$

Démonstration. La preuve de la proposition 13.24, page 267 assure que

$$\partial I_K(u) = \{w, I_K(u) + \langle h | w \rangle \leq I_K(u + h) \quad \forall h\} = \{w, \langle h | w \rangle \leq 0 \quad \forall h, u + h \in K\}$$

est le cône polaire de

$$C = \{h \in H, \langle \nabla \varphi_i | h \rangle \leq 0 \quad \forall i \in I_u\}.$$

qui est lui-même le cône polaire de

$$co((\nabla \varphi_i)_{i=1, \dots, m}) = \left\{ \sum_{i \in I_u} \lambda_i \nabla \varphi_i(u), \lambda_i \geq 0 \right\},$$

i.e. $\partial I_K(u) = co((\nabla \varphi_i)_i)^\circ = co((\nabla \varphi_i)_i)$. □

On peut maintenant donner une formulation du théorème 13.24 dans le présent cadre, sous des hypothèses de convexité des fonctions.

Proposition 13.50. Soit J une fonctionnelle C^1 convexe définie sur espace de Hilbert H , et u un minimiseur de J sur $U \cap K$ défini par (13.3), où les φ_i sont continûment différentiables et convexes sur H . On suppose que les contraintes sont qualifiées en u . On a alors

$$\nabla J(u) \in -\partial I_K,$$

que l'on écrit parfois $\nabla J(u) + \partial I_K \ni 0$.

Démonstration. L'élément u est minimiseur global de $J + I_K$ sur H , on a donc (d'après les propositions 13.47 et 13.48)

$$0 \in \partial(J + I_K) = \nabla J(u) + \partial\varphi,$$

d'où la conclusion. \square

Remarque 13.51. La proposition ci-dessous nécessite des hypothèses de convexité qui n'étaient pas dans le théorème 13.24, de fait il s'agit de conditions nécessaires qui ne nécessitent que des informations locales. Or si K n'est pas convexe, la notion que nous avons donnée de sous-différentiel n'est pas pertinente. La notion plus générale de sous-différentiel de Fréchet (voir définition 13.45 ci-dessus) n'implique que les valeurs de la fonction au voisinage du point considéré, elle peut être définie pour des fonctions non convexes et permet d'énoncer une propriété plus générale.

Compléments

Proposition 13.52. Soit φ une fonction convexe s.c.i. d'un espace de Hilbert H dans $\mathbb{R} \cup \{+\infty\}$, alors φ est localement lipschitzienne sur l'intérieur de son domaine.

Démonstration. On montre dans un premier temps que φ est localement bornée. Si le domaine D est non vide, on considère $0 \in D(\varphi)$, et $r > 0$ tel que $B(x_0, r) \subset \overset{\circ}{D}$. On définit les ensembles F_n par

$$F_n = \{x \in \bar{B}(x_0, r), \varphi(x) \leq n\}.$$

Les F_n sont des fermés par semi-continuité inférieure, leur union est $\bar{B}(x_0, r)$, qui est d'intérieur non vide. D'après le lemme de Baire, il en existe au moins un d'intérieur non vide, donc il existe $n \in \mathbb{N}$, $y \in B(x, r)$, $\eta > 0$, tels que $\varphi(x) \leq n$ sur $\bar{B}(y, \eta)$. Si cette boule contient x_0 , c'est terminé. Sinon on considère \tilde{y} le symétrique de y par rapport à x_0 :

$$\tilde{y} = x_0 - (y - x_0) = 2x_0 - y.$$

Tout point de $\bar{B}(x_0, \eta/2)$ s'écrit comme milieu de \tilde{y} et d'un point de $\bar{B}(y, \eta)$. La convexité assure donc le caractère majoré de φ sur $\bar{B}(x_0, \eta/2)$ (le majorant est $M = (\varphi(\tilde{y}) + n)/2$).

Montrons que φ est également minorée sur cette même boule. Pour $x \in \bar{B}(x_0, \eta/2)$, x_0 s'écrit comme milieu de x et de son symétrique par rapport à x_0 , à savoir $x_0 - (x - x_0) = 2x_0 - x$. On a donc

$$\varphi(x_0) \leq \frac{1}{2}(\varphi(x) + \varphi(2x_0 - x)) \implies \varphi(x) \geq 2\varphi(x_0) - \varphi(2x_0 - x) \geq 2\varphi(x_0) - M.$$

On a donc montré que $|\varphi|$ est borné par une constante C sur une boule autour de x_0 , que l'on note $\bar{B}(x_0, 2r)$. Montrons que φ est lipschitzienne sur $\bar{B}(x_0, r)$. Soient x et y dans $\bar{B}(x_0, r)$. On note $\alpha = |y - x|$, et l'on définit z en "prolongeant" le segment $[x, y]$ d'une longueur r , de telle sorte que $z \in \bar{B}(x_0, 2r)$, et ce nouveau point s'écrit

$$z = y + \frac{r}{\alpha}(y - x).$$

On peut écrire y comme combinaison convexe de x et z :

$$y = \frac{r/\alpha}{1+r/\alpha}x + \frac{1}{1+r/\alpha}z.$$

Par convexité de φ on a

$$\varphi(y) \leq \frac{r/\alpha}{1+r/\alpha}\varphi(x) + \frac{1}{1+r/\alpha}\varphi(z) \implies \varphi(y) - \varphi(x) \leq \frac{1}{1+r/\alpha}(\varphi(z) - \varphi(x)) \leq \frac{2M}{1+r/\alpha} \leq \frac{2M}{r}\alpha.$$

On démontre la même inégalité sur $\varphi(x) - \varphi(y)$ en changeant les rôles de x et y , on a donc

$$|\varphi(y) - \varphi(x)| \leq \frac{2M}{r}\alpha = \frac{2M}{r}|y - x|,$$

d'où le caractère lipschitzien de φ sur $\bar{B}(x_0, r)$

\square

13.6 Dérivation du minimum par rapport aux contraintes

Proposition 13.53. On considère une fonctionnelle d'un ensemble V 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.

$$J(u^\alpha) + \sum \mu_i^\alpha (\varphi_i(u^\alpha) - \alpha_i) \leq J(u^\alpha) + \sum \lambda_i^\alpha (\varphi_i(u^\alpha) - \alpha_i) \leq J(\tilde{u}) + \sum \lambda_i^\alpha (\varphi_i(v) - \alpha_i) \quad \forall \mu \geq 0, v \in V.$$

On note $m(\alpha)$ la valeur du minimum correspondant aux contraintes α . On a

$$m(\alpha) \geq m(0) - \lambda^0 \cdot \alpha.$$

Si la fonction $\alpha \in \mathbb{R}^m \mapsto m(\alpha) \in \mathbb{R}$ est dérivable, alors

$$\nabla m(\alpha) = -\lambda^0, \text{ i.e. } \lambda_i^0 = -\frac{\partial m}{\partial \alpha_i}.$$

Démonstration. On a (d'après la seconde inégalité qui caractérise (u^0, p^0) comme point-selle)

$$\begin{aligned} m(0) &= J(u^0) = J(u^0) + \sum_{i=1}^n \lambda_i^0 \varphi_i(u^0) \leq J(u^\alpha) + \sum_{i=1}^n \lambda_i^0 \varphi_i(u^\alpha) \\ &= J(u^\alpha) + \sum_{i=1}^n \lambda_i^0 (\varphi_i(u^\alpha) - \alpha_i) + \sum_{i=1}^n \lambda_i^0 \alpha_i \end{aligned}$$

qui est (d'après la première inégalité qui caractérise $(u^\alpha, \lambda^\alpha)$ comme point-selle) plus petit que

$$J(u^\alpha) + \sum_{i=1}^n \lambda_i^\alpha (\varphi_i(u^\alpha) - \alpha_i) + \sum_{i=1}^n \lambda_i^0 \alpha_i = J(u^\alpha) + \sum_{i=1}^n \lambda_i^0 \alpha_i$$

On obtient donc bien $m(\alpha) = J(u^\alpha) \geq m(0) - \lambda^0 \cdot \alpha$.

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

$$\nabla m(0) \cdot \alpha + o(1) \geq -\lambda^0 \cdot \alpha,$$

pour tout α décrivant un voisinage symétrique de 0. On a donc bien $\nabla m = -\lambda^0$. \square

Exercice 13.4. Donner un exemple de problème de minimisation sous contraintes d'une fonctionnelle lisse pour lequel la correspondance $\alpha \mapsto m_\alpha$ n'est pas différentiable

Exercice 13.5. (Principe des travaux virtuels et différentiation par rapport aux contraintes)

On considère le système masses-ressort décrit dans la section 13.9, page 281 : les positions des masses sont $x_0, \dots, x_n \in \mathbb{R}$, reliées par des ressorts de même raideur $k > 0$, de telle sorte que l'énergie potentielle associée est

$$J(x) = \frac{1}{2} k \sum_{i=1}^n |x_i - x_{i-1}|^2 = \frac{1}{2} k \langle Ax | x \rangle,$$

où A est la matrice du Laplacien discret avec condition de Neuman (voir (19.16), page 395, en remplaçant les 2 en haut à gauche et en bas à droite par des 1). On s'intéresse à la minimisation de J sous les contraintes $x_0 \leq 0$ et $x_N \geq 1$. Écrire les conditions nécessaires d'optimalité, interprétez en termes de modélisation les multiplicateurs de Lagrange, ainsi que, lorsque l'on perturbe les contraintes aux voisinages de 0 et 1, respectivement, la propriété de différentiation qui fait l'objet de la proposition 13.53.

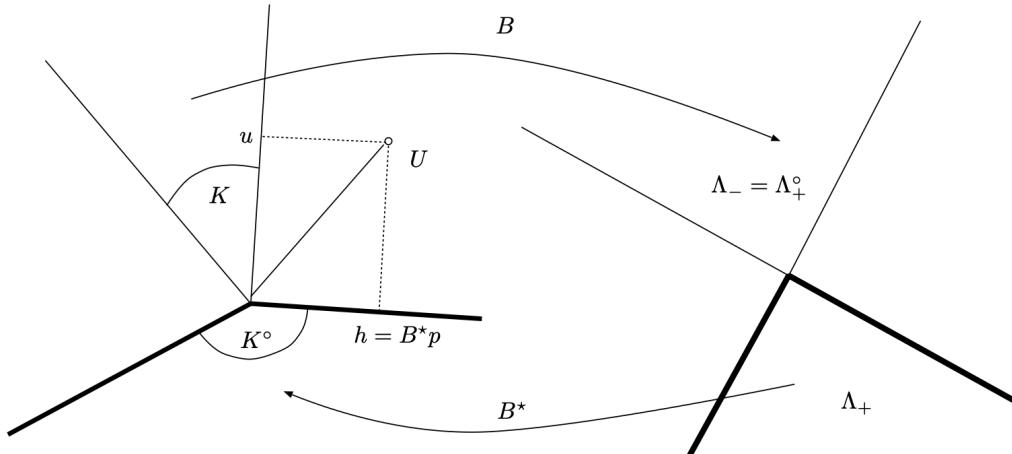


FIGURE 13.2 – Généralisation du Lemme de Farkas

13.7 Cas de la dimension infinie

Dans ce qui précède il n'est pas nécessaire de supposer que l'espace V est de dimension finie. En revanche on a utilisé de façon essentielle le fait que le nombre de contraintes est fini.

On se replace pour commencer dans le cadre de contraintes d'égalité affines traité dans la section 13.2, que l'on généralise au cas d'une infinité de contraintes. Plus précisément, on considère deux espaces de Hilbert V et Λ , $B \in \mathcal{L}(V, \Lambda)$, et un ensemble admissible affine défini comme $K = u_0 + \ker B$. La démonstration de la proposition (13.12) n'utilise la dimension finie qu'au travers de l'identité $(\ker B)^\perp = \overline{\text{Im}B^*}$. En toute généralité, on a seulement $(\ker B)^\perp = \overline{\text{Im}B^*}$. On peut ainsi démontrer :

Proposition 13.54. 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

$$\begin{aligned} \nabla J(u) + B^*\lambda &= 0 \\ Bu &= Bu_0. \end{aligned}$$

Remarque 13.55. 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

$$\nabla J(u) + B^*\lambda_\varepsilon = o(1).$$

On peut aussi généraliser l'approche développée dans la section 13.3, sur les contraintes d'inégalité. On considère comme précédemment deux espaces de Hilbert V et Λ , et $B \in \mathcal{L}(V, \Lambda)$. On considère de plus Λ_+ un cône convexe fermé¹⁰ de Λ . On définit l'ensemble admissible comme l'image réciproque par B de Λ_+° , c'est à dire

$$K = \{v \in V, \langle Bv | \mu \rangle \leq 0 \quad \forall \mu \in \Lambda_+\} = B^{-1}(\Lambda_+^\circ). \quad (13.11)$$

Lemme 13.56. Sous les hypothèses décrites ci-dessus, avec K défini par (13.11), on a

$$K^\circ = \overline{B^*\Lambda_+}.$$

10. Dans les sections précédentes traitant le cas d'un nombre fini de contraintes, $\Lambda = \mathbb{R}^m$, $\Lambda_+ = \mathbb{R}_+^m$, et B est l'application associée à la matrice dont les lignes sont les g_i .

Démonstration. Une inclusion est immédiate : pour tout $\mu \in \Lambda_+$, tout $v \in K$, on a

$$\langle v | B^* \mu \rangle = \langle Bv | \mu \rangle \leq 0,$$

d'où $B^*\Lambda_+ \subset K^\circ$ et donc $\overline{B^*\Lambda_+} \subset K^\circ$ (car K° est fermé). Supposons que l'inclusion soit stricte : il existe alors $z \in K^\circ \setminus \overline{B^*\Lambda_+}$. On peut séparer $\{z\}$ et $\overline{B^*\Lambda_+}$ d'après le théorème de Hahn-Banach, il existe donc $h \in V$ et $\alpha \in \mathbb{R}$ tel que

$$\langle h | B^* \mu \rangle \leq \alpha < \langle h | z \rangle.$$

L'ensemble décrit par le membre de gauche étant un cône de \mathbb{R} de sommet 0, on a $\langle h | B^* \mu \rangle \leq 0 \leq \alpha$ pour tout $\mu \in \Lambda_+$, et donc en particulier $h \in K$. Mais comme $\langle h | z \rangle > 0$ avec $z \in K^\circ$, c'est absurde. On a donc bien l'identité entre les deux ensembles. \square

Proposition 13.57. On considère une fonctionnelle J continûment différentiable sur l'espace de Hilbert V , et u un minimum local de J sur $V \cap K$, où K est défini par (13.3). On suppose que $B^*\Lambda_+$ est fermé. Il existe alors $\lambda \in \Lambda_+$ tel que

$$\nabla J(u) + B^* \lambda = 0. \quad (13.12)$$

Proposition 13.58. On considère une fonctionnelle J continûment différentiable sur l'espace de Hilbert V , et u un minimum local de J sur $V \cap K$, où K est défini par (13.3). On suppose que B est *surjective*, on a alors existence unicité de λ tel que (13.12) soit vérifiée.

Démonstration. Si B est surjective, alors B^* est injective et à image fermée, ce qui implique l'existence d'une constante $C > 0$ telle que

$$|B^* \mu| \geq C |\mu| \quad \forall \mu \in \Lambda.$$

Si l'on considère une suite $(B^* \mu_n)$ qui converge vers $v \in V$, avec $\mu_n \in \Lambda_+$, alors le caractère de Cauchy de $(B^* \mu_n)$ se transpose à la suite (μ_n) d'après l'inégalité ci-dessus, d'où $\mu_n \rightarrow \mu$, et $v = B^* \mu \in B^*(\Lambda_+)$. \square

Remarque 13.59. On prendra garde au fait qu'il ne suffit pas de supposer que B (ou de façon équivalente B^*) est à image fermée pour assurer le caractère fermé du cône $B^*\Lambda_+$. En effet, l'image d'un ensemble fermé par une application linéaire à image fermée n'est pas nécessairement fermée, même dans le cas d'un cône convexe, et même en dimension finie. On se reportera à l'exercice 13.11, page 283, pour un contre exemple.

13.8 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 :

$$K = \{ v \in \mathbb{R}^d, \varphi_i(v) = 0, i = 1, \dots, N \}.$$

Proposition 13.60. (Muplicateurs de Lagrange, contraintes d'égalité)

Soit $J : U \subset \mathbb{R}^d \rightarrow \mathbb{R}$ 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 en u des fonctionnelles φ_i forment une famille libre. Il existe alors $\lambda_1, \dots, \lambda_N$, 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 $\text{vect}(\nabla\varphi_i(u))_i$, ce qui termine la preuve.

Montrons maintenant que tout vecteur h orthogonal à tous les $\nabla\varphi_i(u)$, est bien une *direction admissible* en u .

On note $g_i = \nabla\varphi_i(u)$, et

$$V = \text{vect}(g_1, \dots, g_N)^\perp.$$

Comme les vecteurs g_i forment une famille libre, V est de dimension $d - N$. On considère une base (h_1, \dots, h_{d-N}) de V , on note

$$x = (x_1, \dots, x_{d-N}) \in \mathbb{R}^{d-N}, \quad y = (y_1, \dots, y_N) \in \mathbb{R}^N$$

et l'on définit γ l'application

$$\gamma : (x, y) \in \mathbb{R}^d \mapsto \gamma(x, y) = u + x_1 h_1 + \dots + x_{d-N} h_{d-N} + y_1 g_1 + \dots + y_N g_N.$$

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

$$(x_k, y_1, y_2, \dots, y_N) \mapsto \varphi \circ \gamma_k(x_k, y_1, \dots, y_N),$$

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 (théorème 19.11, page 376) pour construire une courbe $(y_1, \dots, y_N) = y = y(x_k)$ 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 -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,

$$\dot{y}(0) = -(\nabla(\varphi \circ \gamma_k))^{-1} \frac{\partial(\varphi \circ \gamma_k)}{\partial x_k}|_{(0,0)} = (G^T G)^{-1} (G^T h_k)$$

qui est nul car h_k est orthogonal à tous les g_i . On a donc

$$\frac{d}{dt} \gamma_k(t, y(t))|_{t=0} = h_k + \dot{y}_1(0) g_1 + \dots + \dot{y}_N(0) g_N = h_k,$$

ce qui termine la démonstration. □

Remarque 13.61. 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 $\varphi_2(x, y) = y - x^2$. 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.

13.9 Illustrations

Système masses - ressorts

Considérons une chaîne horizontale de $n + 1$ masses $0, 1, 2, \dots, n$, 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 $(x_0, x_1, \dots, x_n) \in \mathbb{R}^{n+1}$. L'énergie potentielle du système s'écrit

$$J(x) = \frac{1}{2}k \sum_{i=1}^n |x_i - x_{i-1}|^2 = \frac{1}{2}k \langle Ax | x \rangle,$$

où A est (à une constante multiplicative près) la matrice du Laplacien discret avec conditions de Neuman. Tout point diagonal (x, x, \dots, 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

$$E = \{x, x_0 = 0, x_n = L\} = X + \ker B, \text{ avec } B : x \in \mathbb{R}^{n+1} \mapsto (x_0, x_n) \in \mathbb{R}^2.$$

La matrice B s'écrit

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

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 extrémiales, 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

$$J_\varepsilon(x) = \frac{1}{2}k \sum_{i=1}^n |x_i - x_{i-1}|^2 + \frac{1}{2\varepsilon} (|x_0|^2 + |x_n - L|^2).$$

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 13.62. 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 transparaîsse 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 aisément de supprimer des contraintes pour parvenir à un jeu minimal équivalent qui assurera l'unicité des multiplicateurs de Lagrange. 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 13.63. 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

$$J_\varepsilon(x) = \frac{1}{2}k \sum_{i=1}^n |x_i - x_{i-1}|^2 + \frac{1}{2\varepsilon} (|x_0|^2 + |x_n - L|^2).$$

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.

Equilibre de Nash

On définit un jeu à N agents comme la donnée de N fonctions d'utilité g_1, \dots, g_N :

$$g_i : U_1 \times \cdots \times U_N \longrightarrow \mathbb{R},$$

où U_i est l'ensemble des stratégies possibles pour i . On note $u_i \in U_i$ la stratégie choisie par l'agent i . Son gain (*pay-off*) dépend donc de sa propre stratégie u_i et des stratégies des autres joueurs. On notera u_{-i} la collection des stratégies des autres joueurs, de telle sorte que g_i peut être vue comme une fonction de (u_i, u_{-i})

Definition 13.64. On appelle équilibre de Nash associé au jeu ci-dessus une collection de stratégies telle que chaque joueur maximise son utilité, au vu des stratégies des autres joueurs, i.e. $u = (u_1, \dots, u_N) \in U_1 \times \cdots \times U_N$ est un équilibre de Nash si et seulement si

$$g_i(u_i, u_{-i}) = \max_{v \in U_i} g_i(v, u_{-i}).$$

13.10 Exercices

Exercice 13.6. Soit $f \in C^2(\mathbb{R})$. On suppose que f est α -convexe. Montrer que α est un minorant de la dérivée seconde en tout point. Comment peut-on généraliser cette propriété aux dimensions supérieures ?

Exercice 13.7. a) Soit J une fonctionnelle s.c.i. propre sur un espace de Hilbert H , à valeurs dans $\mathbb{R} \cup \{+\infty\}$. Montrer que, si J est fortement convexe, alors J est coercive¹¹.
(On pourra commencer par montrer que J décroît au plus linéairement quand $|x|$ tend vers $+\infty$.)
 b) Donner un exemple de fonctionnelle définie sur un convexe K d'un espace de Hilbert H qui soit fortement convexe sans être coercive.

Exercice 13.8. Donner un exemple de fonction strictement convexe non coercive, et même non minorée. Donner un exemple de fonction coercive non convexe. Montrer qu'une fonction convexe de \mathbb{R} dans \mathbb{R} qui tend vers $-\infty$ quand x tend vers $-\infty$ tend vers $+\infty$ quand x tend vers $+\infty$.

Exercice 13.9. a) Soit J une fonctionnelle α -convexe différentiable d'un espace de Hilbert V dans \mathbb{R} . Montrer que

$$\langle \nabla J(v) - \nabla J(u) | v - u \rangle \geq \alpha |v - u|^2 \quad \forall u, v \in V.$$

b) (**) On considère maintenant une fonctionnelle J de \mathbb{R}^n dans \mathbb{R} convexe et deux fois continûment différentiable. On pour tout $t \geq 0$ on définit $\Phi_t = \text{Id} + t\nabla J$. Montrer que pour tout $t \geq 0$ Φ_t est un C^1 difféomorphisme de \mathbb{R}^n dans \mathbb{R}^n .

(Indication : on pourra montrer que Φ_t est injective, que $|\Phi_t|$ est coercive, que l'image de Φ_t est à la fois ouverte et fermée, et conclure.)

c) Montrer Φ_t est expansive pour pour $t \geq 0$, et plus précisément que, pour tout $A \subset \mathbb{R}^n$ mesurable, $t \mapsto \lambda(\Phi_t(A))$ est croissante (où λ désigne la mesure de Lebesgue).

Exercice 13.10. a) Soit $B \in \mathcal{M}_{mn}(\mathbb{R})$. Démontrer, sans utiliser le théorème du rang, l'identité

$$(\ker B)^\perp = \text{Im}B^*.$$

b) Montrer la propriété (appelée parfois lemme des noyaux) : soit E un e.v.n, $\varphi_1, \dots, \varphi_N$ des formes linéaires continues sur E , et φ une forme linéaire continue telle que

$$\ker \varphi \subset \bigcap \ker \varphi_i.$$

Montrer que φ est combinaison linéaire des φ_i .

c) Préciser le lien entre les deux propriétés ci-dessus.

Exercice 13.11. Montrer que l'image d'un cône fermé de \mathbb{R}^2 par une application linéaire est fermée, mais que ce résultat n'est plus vrai en général dans \mathbb{R}^d pour $d \geq 3$.

Exercice 13.12. Soit J une fonction C^2 de \mathbb{R}^d dans \mathbb{R} . On considère le système dynamique suivant, appelé *flot de gradient* pour J :

$$\frac{du}{dt} = -\nabla J(u) \quad (\star)$$

avec condition initiale $u(0) = u_0$.

1) Montrer que (\star) admet une solution maximale unique.

2) Montrer que, si J est coercive, alors cette solution est globale.

On considère le *schéma d'Euler implicite* pour (\star) , qui consiste à définir u^{n+1} comme solution de

$$\frac{u - u^n}{\tau} = -\nabla J(u) \quad (\star\star)$$

11. Dans le cas où la section 13.5 n'aurait pas été abordée, on pourra supposer plus simplement J continue à valeurs dans \mathbb{R} .

où $\tau > 0$ est le pas de temps. Ce schéma est pour l'instant abstrait du fait que, sans hypothèse supplémentaire, il n'est pas garanti ce cette équation admette une solution unique.

On suppose que J est α -convexe, pour $\alpha \in \mathbb{R}$.

3) Vérifier que u est solution si et seulement si $\nabla \Phi_\tau(u) = 0$, où Φ_τ est une fonctionnelle que l'on précisera.

4) Montrer qu'il existe τ^* tel que la fonctionnelle Φ_τ admet un minimiseur unique pour tout $\tau \in]0, \tau^*]$.

5) En déduire que, pour tout $\tau \in]0, \tau^*]$, le schéma d'Euler ($\star\star$) admet une solution unique.

On se place dans l'hypothèse $\tau \in]0, \tau^*]$, et l'on considère la suite des itérés construites suivant le schéma d'Euler implicite : $u^0 = u_0, u^1, u^2, \dots$

6) Montrer que la suite des $J(u^n)$ est décroissante. Montrer que, si $\nabla J(u^n) \neq 0$, on a $J(u^{n+1}) < J(u^n)$.

Exercice 13.13. Soit K une partie convexe d'un e.v.n. E , et J une fonctionnelle différentiable sur un ouvert contenant K .

1) Montrer que, si u est un minimum local de J sur K , alors

$$\langle DJ(u), v - u \rangle \geq 0 \quad \forall v \in K.$$

2) Réciproquement, montrer que, si u vérifie l'inégalité ci-dessus, et si J est convexe sur K , alors u est un minimum global de J sur K .

Exercice 13.14. On considère une institution publique proposant de distribuer à des personnes $i = 1, 2, \dots, n$ des subventions. On considère que, du fait de sa situation sociale, la personne i est légitime à toucher la somme α_i . On note $u = (u_1, \dots, u_n)$ le vecteur des sommes effectivement allouées, et l'on considère la fonctionnelle

$$u \longmapsto J(u) = \frac{1}{2} \sum_{i=1}^n |u_i - \alpha_i|^2.$$

a) Montrer que J admet un minimiseur unique, correspondant à l'attribution effective à i de la somme α_i . Montrer qu'il en est de même si l'on remplace la puissance 2 par n'importe quel réel positif (on adoptera la convention que $u^0 = 1$ pour $u > 0$, et $0^0 = 0$).

b) On suppose maintenant que l'institution dispose d'une somme $S > 0$ qu'elle prévoit d'attribuer effectivement à la population. Elle se propose de minimiser la fonctionnelle J sous la contrainte que la somme totale attribuée est inférieure ou égale à S . Faire l'analyse du problème d'optimisation sous contrainte correspondant, et préciser le minimiseur (et donc la stratégie de distribution opérée).

Vérifier que cette démarche peut aboutir à des u_i négatifs (on demande à certains agents, les moins nécessiteux, de verser de l'argent dans le pot commun qui permettra de subventionner les plus nécessiteux).

c) On considère qu'il est exclu de demander de l'argent aux agents. Proposer une nouvelle formulation du problème qui exclut cette situation, et faire son analyse. On pourra supposer les α_i distincts deux à deux pour simplifier l'analyse.

Commenter les résultats obtenus dans le cas d'exploitations agricoles de tailles différentes, où α_i est proportionnel à la taille de l'exploitation i .

d) Reprendre la question précédente (analyse et interprétation) dans le cas où chaque terme de la somme définissant i est affectée d'un poids $\beta_i > 0$.

e) Que se passe-t-il si on supprime les carrés dans la somme définissant la fonctionnelle J ?

Exercice 13.15. On considère une famille $(g_i)_{i \in \mathbb{N}}$ infinie de vecteurs de \mathbb{R}^d , et le cône C engendré par cette famille, i.e. l'ensemble des combinaisons linéaires finies à coefficients positifs d'éléments de la famille. Montrer que, dès que $d \geq 2$, ce cône n'est pas nécessairement fermé.

Exercice 13.16. (La loi d'ohm comme conséquence de la loi des noeuds)

On considère un réseau électrique connexe constitués de fils $e \in E \subset V \times V$, où V est l'ensemble fini des sommets. On note Γ un sous-ensemble de points de V (au moins 2), en lesquels l'intensité sortante est supposée fixée. Écrire les conditions d'optimalité associées au problème de minimisation de l'énergie dissipée

$$J(I) = \frac{1}{2} \sum_e r_e I_e^2,$$

sous la contrainte de flux imposé en les points de Γ , et la loi des noeuds (ou de Kirchhoff) en chaque point intérieur au réseau. En déduire la loi d'Ohm sur chaque arête du réseau, où le potentiel électrique apparaît comme un multiplicateur de Lagrange de la loi des noeuds.

Exercice 13.17. (Mesure de Gibbs)

On s'intéresse à la minimisation de l'entropie

$$p = (p_1, \dots, p_N) \longmapsto S(p) = \sum p_i \log p_i,$$

parmi les lois de probabilité sur l'ensemble à N éléments telle que la moyenne des quantités E_1, \dots, E_N , est prescrite :

$$\sum_{i=1}^N p_i E_i = \bar{E}$$

- 1) Montrer que, si $\bar{E} \in [\max E_i, \min E_i]$, il existe un unique minimiseur.
- 2) Faire l'analyse du problème dans le cas où \bar{E} est l'une des extrémités de l'intervalle ci-dessus.

On fait désormais l'hypothèse $\bar{E} \in]\max E_i, \min E_i[$.

- 3) Faire l'analyse du problème dans le cas $N = 2$.

On fait désormais l'hypothèse que $N \geq 3$.

- 4) Montrer que, si le minimiseur est atteint sur $]0, +\infty[^N$, alors il existe $\beta \in \mathbb{R}$ tel que

$$p_i = \frac{1}{Z} \exp(-\beta E_i), \quad Z = \sum \exp(-\beta E_i).$$

- 5) Montrer qu'il existe un $\beta \in \mathbb{R}$ tel que

$$\frac{\sum \exp(-\beta E_i) E_i}{\sum \exp(-\beta E_i)} = \bar{E}.$$

- 6) En déduire que le $p = (p_i)$ défini à la question 4 est le minimiseur recherché.

Exercice 13.18. (Mesure de Gibbs : démonstration alternative)

On se place dans les hypothèses de l'exercice 13.17, dont on reprend les questions de 1 à 4.

- 5) Montrer que le minimiseur ne peut pas être atteint sur le bord de $]0, +\infty[^N$, et conclure.

Exercice 13.19. (Sous-différentiels)

Déterminer le sous différentiel des fonctions suivantes :

- a) $\varphi : x \in \mathbb{R}^d \longmapsto |x|$.
- b) φ convexe affine par morceaux sur \mathbb{R} .
- c) $\varphi = I_K$ où $K \subset \mathbb{R}$ est un intervalle.
- d) $\varphi = I_K$ où $K \subset \mathbb{R}^2$ est un polyèdre convexe (on pourra faire un dessin)

Chapitre 14

Transport optimal discret

Sommaire

14.1	Problème d'affectation et problème de Monge Kantorovich discret	286
14.2	Formulation duale du problème de MK discret	289
14.3	Exemples d'applications	291
14.4	Métriques induites sur l'ensemble des mesures de probabilités sur un espace métrique fini	292
14.5	Métriques induites sur l'ensemble des mesures atomiques	293
14.6	Complétion de l'espace de Wasserstein discret	295
14.7	Régularisation entropique	296
14.8	Calcul effectif par Régularisation entropique	298
14.9	Calcul effectif par l'algorithme des enchères	300
14.10	Compléments, extensions	303
14.11	Compléments sur la régularisation entropique	305
14.12	Exercices	309

14.1 Problème d'affectation et problème de Monge Kantorovich discret

Le problème d'affectation se formule comme suit :

Problème 14.1.1. 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 $N!$.

Nous allons considérer une version relaxée du problème ci-dessus¹, 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 cela coûte de transporter une quantité unitaire de matière 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 destinations. Cette relaxation du problème permet de lever la contrainte d'avoir le même nombre de points au départ et à l'arrivée. Nous considérons donc deux ensembles finis X et Y de cardinaux non nécessairement identiques. Pour $x \in X$ et $y \in Y$, on notera γ_{xy} la quantité de matière allant de x vers y . On appellera $\gamma = (\gamma_{xy})$ un *plan de transport*. On pourra se ramener à des index entiers en identifiant X à $\{1, \dots, N\}$ et Y à $\{1, \dots, M\}$.

Problème de Monge-Kantorovich discret

On considère 2 ensembles² finis X et Y , de cardinaux respectifs N et $M \in \mathbb{N}$ et l'on se donne une collection de coûts $c_{xy} \in \mathbb{R}$. On se donne deux mesures de probabilités discrètes μ et ν sur X et Y , respectivement (μ_x est la masse portée par x , avec $\sum \mu_x = 1$, de même pour ν). On supposera tous les poids strictement positifs³. Le problème s'écrit

$$\min_{\Pi_{\mu\nu}} C(\gamma) \quad (14.1)$$

avec

$$C(\gamma) = \sum_{x,y} c_{xy} \gamma_{xy}, \quad \Pi_{\mu,\nu} = \left\{ \gamma \in \mathbb{R}_+^{N \times M}, \quad \sum_y \gamma_{xy} = \mu_x \quad \forall i, \quad \sum_x \gamma_{xy} = \nu_y \quad \forall y \right\}$$

Remarque 14.1. 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" $c = (c_{xy})$ sur $X \times Y$.

Remarque 14.2. 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

$$\gamma_{xy} = \mu_x \nu_y.$$

Nous verrons plus loin que c'est le plan qui minimise l'entropie de la loi γ (voir définition 1.10, page 25).

Proposition 14.3. Le problème 14.1 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 fonctionnelle continue (car linéaire en dimension finie) $C(\cdot)$ admet donc un minimiseur sur Π . \square

Remarque 14.4. Dans le cas d'un coût du type $c_{xy} = a_x + b_y$, le problème est fortement dégénéré, puisque tout transport de μ vers ν réalise le même coût. Par ailleurs, 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 bijection⁴ soit l'unique minimiseur, en prenant par exemple $c_{x\varphi(x)} = -1$, et $c_{xy} = 0$ si $y \neq \varphi(x)$.

Lien avec le problème d'affectation

1. Cette approche a été proposée par L.V. Kantorovich en 1942. On trouvera une traduction du papier original sur <http://www.math.toronto.edu/mccann/assignments/477/Kantorovich42.pdf>

2. 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 concevoir le transport d'une essoreuse vers le *concept de néant* chez Sartre.

3. On peut toujours se ramener à cette situation en supprimant de X et Y les points non chargés.

4. C'est à dire : $\gamma_{x\varphi(x)} = 1/N$, et $\gamma_{xy} = 0$ si $y \neq \varphi(x)$.

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. Il ne s'agit donc plus de mesure de probabilité, mais on peut s'y ramener en divisant la mesure par le nombre de points.

Proposition 14.5. 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 transport associés à une affectation, i.e. $\gamma_{ij} = \delta_{i\varphi(j)}$, où φ est une permutation du groupe symétrique. L'ensemble des points extrémaux⁵ de Π s'identifie à Π_S .

Démonstration. Tout point de Π_S 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 points x pour lesquels γ_{xy} est nul pour tous les points $y \in Y$ sauf un (qui vaut donc 1). Cette sous-famille des points de départ est en bijection avec les points d'arrivées correspondants, pour lesquels, symétriquement, $\gamma_{xy'}$ est nul pour tous les x sauf y . On note I (resp. J) l'ensemble des points non concernés dans l'espace de départ (resp. d'arrivée). Les ensemble I et J sont de même cardinal, et non vides par hypothèse. La restriction du plan γ à $X_I \times Y_J$ est diffuse, au sens que pour tout x , $\gamma_{xy} \in]0, 1[$ pour au moins 2 points $y \in J$, et pour tout $y \in J$, on a $\gamma_{xy} \in]0, 1[$ pour au moins 2 points $x \in I$. On part d'un point $x_0 \in I$, et l'on choisit y_0 tel que $\gamma_{x_0 y_0} > 0$. On choisit ensuite $x_1 \neq x_0$ tel que $\gamma_{x_1 y_0} > 0$, puis $y_1 \neq y_0$ tel que $\gamma_{x_1 y_1} > 0$. On construit ainsi une suite de points alternant entre X et Y

$$x_0, y_0, i_1, \dots, x_{n-1}, y_n,$$

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 x_n correspond à un point $x_\ell \neq x_{n-1}$ déjà visité. On considère alors la variation

$$h = \sum_{k=\ell}^{n-1} (\pi_{x_k, y_k} - \pi_{x_{k+1}, y_k}),$$

avec $x_n = x_\ell$, et où π_{xy} est l'élément de \mathbb{R}^{NM} qui vaut 1 sur la composante (x, y) , et qui est nul pour les autres couples. Pour η suffisamment petit, $\gamma \pm \eta h$ est positif, et par construction $\gamma \pm \eta h$ vérifie 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 extrémaux correspondent donc aux permutations. \square

Corollaire 14.6. L'ensemble Π des plans de transport admissibles est l'enveloppe convexe de Π_S .

Démonstration. Il s'agit d'une conséquence du théorème de Krein-Milman en dimension finie (théorème 19.10, page 376), qui assure que tout convexe compact d'un espace affine de dimension finie est l'enveloppe convexe de ses points extrémaux. \square

Proposition 14.7. On se place comme précédemment dans la situation de mesures équidistribuées sur des ensembles de même cardinal. Le problème de Monge Kantorovich discret 14.1 admet au moins une solution dans Π_S , i.e. une solution optimale du type permutation.

Démonstration. D'après la proposition 14.3, le problème 14.1 admet un minimiseur γ . D'après la proposition 14.5, ce minimiseur s'écrit comme combinaison convexe de plans associés à des permutations $\varphi_1, \dots, \varphi_K$:

$$\gamma = \sum \theta_k \gamma^k$$

(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

$$C(\gamma) = \sum \theta_k C(\gamma^k).$$

5. On dit que $\gamma \in \Pi \subset \mathbb{R}^d$ est point extrémal de Π si $\gamma = (\gamma^1 + \gamma^2)/2$, avec $\gamma^1, \gamma^2 \in \Pi$, implique $\gamma^1 = \gamma^2 = \gamma$.

Comme chaque $C(\gamma^k)$ est supérieur ou égal à $C(\gamma)$, et que $\sum \theta_k = 1$ avec $\theta_k > 0$ 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. \square

14.2 Formulation duale du problème de MK discret

La formulation duale du problème 14.1 est basée sur l'expression duale des contraintes de marginales :

$$\sum_{y \in Y} \gamma_{xy} = \mu_x \quad \forall x \iff \sum_{x \in X} p_x \left(\mu_x - \sum_y \gamma_{xy} \right) = 0 \quad \forall p \in \mathbb{R}^X,$$

et l'on exprime de même les contraintes de destination à l'aide de $q \in \mathbb{R}^Y$. On introduit donc (conformément à la définition 13.31, page 270) le Lagrangien

$$(\gamma, p, q) \in V \times \Lambda \mapsto \sum_{x,y} c_{xy} \gamma_{xy} + \sum_{x=1} p_x \left(\mu_x - \sum_y \gamma_{xy} \right) + \sum_{y=1} q_y \left(\nu_y - \sum_x \gamma_{xy} \right), \quad (14.2)$$

avec $V = \mathbb{R}_+^{X \times Y}$ et $\Lambda = \mathbb{R}^X \times \mathbb{R}^Y$. 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 définition 13.28, page 269) est le problème consistant à minimiser la fonctionnelle

$$F(\gamma) = \sup_{p,q} L(\gamma, p, q) = \begin{cases} \sum_{x,y} c_{xy} \gamma_{xy} & \text{si } \gamma \in \Pi \\ +\infty & \text{sinon} \end{cases}$$

Minimiser cette fonctionnelle revient bien à résoudre le problème 14.1 de minimisation sous contrainte.

Le problème dual (voir toujours la définition 13.28, page 269) 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 $L(\gamma, p, q)$) :

$$\begin{aligned} G(p, q) &= \inf_{\gamma \in V} \left(\sum_{x,y} (c_{xy} - p_x - q_y) \gamma_{xy} + \sum_{x \in X} p_x \mu_x + \sum_{y \in Y} q_y \nu_y \right) \\ &= \sum_{x \in X} p_x \mu_x + \sum_{y \in Y} q_y \nu_y + \inf_{\gamma \in V} \left(\sum_{x,y} (c_{xy} - p_x - q_y) \gamma_{xy} \right). \end{aligned}$$

Comme γ parcourt $V = \mathbb{R}_+^{X \times Y}$, l'infimum ci-dessus vaut $-\infty$ à moins que l'on ait $p_x + p_y \leq c_{xy}$ pour tous x, y , et 0 dans ce dernier cas. On a donc

$$G(p, q) = \inf_{\gamma \in V} L(\gamma, p, q) = \begin{cases} \sum_{x \in X} p_x \mu_x + \sum_{y \in Y} q_y \nu_y & \text{si } p_x + q_y \leq c_{xy} \quad \forall i, j, \\ -\infty & \text{sinon}. \end{cases}$$

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

$$\sup_{p \oplus q \leq c} (p \cdot \mu + q \cdot \nu).$$

Il s'agit de montrer que le Lagrangien défini ci-dessus admet un point selle ou, de façon équivalente (voir proposition 13.30, page 269), que le problème dual admet une solution, et que sa valeur maximale est la valeur minimale du problème initial. La propriété suivante permet de se ramener à la construction de vecteurs de multiplicateurs de Lagrange vérifiant une propriété très simple. La démonstration en est élémentaire, mais vue son importance nous la présentons sous la forme d'une proposition.

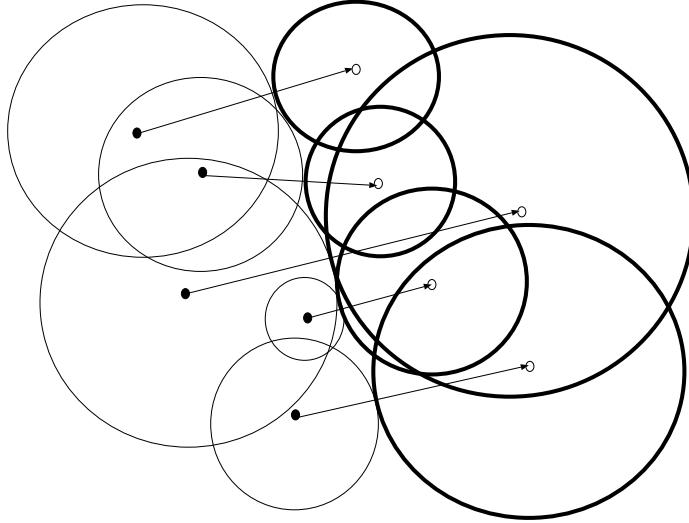


FIGURE 14.1 – Interprétation géométrique des potentiels de Kantorovich pour la distance 1.

Proposition 14.8. Soit γ un plan de transport entre μ et ν . Si (p, q) vérifie $p \oplus q \leq c$, avec égalité sur le support de γ , i.e.

$$\gamma_{xy} > 0 \implies p_x + q_y = c_{xy},$$

alors (γ, p, q) est point-selle pour le Lagrangien L (défini par (14.2)).

Démonstration. En effet, (p, q) vérifie alors la contrainte du problème dual, et on a

$$G(p, q) = \sum_x \mu_x p_x + \sum_y \mu_y q_y = \sum_{xy} \gamma_{xy} (p_x + q_y) = \sum_{xy} \gamma_{xy} c_{xy} = F(\gamma).$$

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) (voir proposition 13.30, page 269). \square

Remarque 14.9. Dans le cas où X et Y sont des collections d'un même nombre N de points de \mathbb{R}^d , et que $c_{xy} = |y - x|$, la remarque précédente peut s'interpréter géométriquement : pour trouver un minimiseur du coût, il suffit⁶ de trouver $2N$ cercles (ou sphères pour $d \geq 3$) Σ^x et Σ^y centrés en les points x et y , respectivement, de telle sorte qu'il existe une bijection φ telle que Σ^x est tangent à $\Sigma^{\varphi(x)}$, et que les autres couples de cercles (Σ^x, Σ^y) ne se chevauchent pas strictement. Selon cette vision du problème dual, les p_x (resp. q_y) sont les rayons des cercles Σ^x (resp. Σ^y). La figure 14.1 donne un exemple d'une telle construction, pour $d = 2$ et $N = 5$.

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 finie) de montrer rapidement l'existence d'un point-selle.

Proposition 14.10. Le Lagrangien $L(\cdot, \cdot, \cdot)$ admet un point selle (γ, p, q) ou, de façon équivalente,

$$G(p, q) = \max_{\tilde{p} \oplus \tilde{q} \leq c} G(\tilde{p}, \tilde{q}) = \min_{\tilde{\gamma} \in \Pi} F(\tilde{\gamma}) = F(\gamma).$$

6. 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.

Le couple (p, q) sature l'inégalité $p_x + q_y \leq c_{xy}$ sur le support de γ , c'est à dire

$$\forall (x, y) \text{ tel que } \gamma_{xy} > 0, p_x + q_y = c_{xy}.$$

Démonstration. On dualise ici les contraintes de positivité, au sens où l'on considère le problème consistant à minimiser $\gamma \mapsto C(\gamma)$ sous les contraintes de marginales, et les contraintes de positivité. Ce problème de minimisations sous contraintes (les contraintes sont toutes affines, donc qualifiées au sens de la définition 13.23, page 267) rentre dans le cadre du théorème 13.24, page 267 (en notant que les contraintes d'égalité affines peuvent se traiter comme deux contraintes d'inégalité affines, voire corollaire 13.25), et la condition d'optimalité s'écrit

$$c_{xy} - p_x - q_y - \lambda_{xy} = 0,$$

avec $\lambda_{xy} \geq 0$ pour tous x, y , et $\lambda_{xy} = 0$ dès que $\gamma_{xy} > 0$ (contrainte non activée). Le couple (p, q) vérifie donc la contrainte d'inégalité, avec égalité sur le support de γ , ce qui implique (voir proposition 14.8) que (γ, p, q) est point-selle du Lagrangien. Pour finir, $G(p, q) = F(\gamma)$ assure que (p, q) sature l'inégalité $p_x + q_y \leq c_{xy}$ sur le support de γ (voir preuve de la proposition 14.8). \square

Remarque 14.11. Soit (p, q) une solution du problème dual. On a alors

$$p_x = \min_y (c_{xy} - q_y).$$

En effet, pour tout x on a $p_i \leq c_{xy} - q_y$ pour tout y , d'après la contrainte, et si l'inégalité était stricte pour tous les y , on pourrait augmenter un peu le p_x , sans violer la contrainte, et en augmentant strictement la valeur du maximum.

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 14.7, page 296).

14.3 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_{xy} . Dans un grand nombre de situations, X et Y sont des ensembles de points de l'espace euclidien, et c_{xy} 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 du plan, la mesure d'arrivée ν est supportée par M points, et les coûts sont donnés par $c_{xy} = |y - x|$. Le problème envisagé par Monge concernait des déblais et des remblais, on peut étendre ce cadre à des lieux de production et de distribution : N boulangeries produisent des quantités de pain journalières (μ_x) destinées à M dépôts de pains dont les flux de vente s'écrivent (ν_y). Si l'on suppose que le coût de transport d'une quantité de pain peut être calculé en multipliant la quantité par un coût unitaire⁷, et que ce coût unitaire est lui-même proportionnel à la distance entre 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 $c_{xy} = |y - x|^p$, 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.

Marché du travail

6. On n'a bien sûr alors aucun information sur le signe du multiplicateur de Lagrange (ici p_i ou q_j), dont le signe final dépendra de laquelle des deux contraintes est réellement activée.

7. Cette hypothèse qui est assez discutable, et donc problématique puisque toute l'approche est basée sur cette hypothèse.

On considère une unité de production basée sur l'accomplissement d'un certain nombre de tâches, et l'on note ν_y le temps (par exemple hebdomadaire) qui doit être consacré à la tâche y . On a conjointement un ensemble X de travailleurs, chacun disposant d'un temps individuel μ_x . On note u_{xy} la productivité de x pour la tâche y , de telle sorte que si x consacre le temps γ_{xy} à la tâche x , la production effective s'écrira $\gamma_{xy}u_{xy}$. Maximiser la production globale correspond à un problème de type Monge-Kantorovich discret, c'est à dire à la recherche d'un plan d'affectation γ qui maximise la quantité

$$\sum_{xy} \gamma_{xy} u_{xy}.$$

Considérons un plan optimal γ . Le problème dual admet une solution (p, q) qui vérifie

$$p_x = \max_y (u_{xy} - q_y)$$

où q_y réalise le maximum ci-dessus dès que $\gamma_{xy} > 0$. On peut alors interpréter les q_y comme des rétributions (par unité de temps) associées aux tâches $y \in Y$. Pour une situation "équilibrée" (γ, p, q) point-selle du Lagrangien, si l'on considère que $u_{xy} - q_y$ est le gain effectif unitaire associé à l'affectation de x à la tâche y , le chef d'entreprise (ou planificateur) n'a, pour tout x , aucun intérêt à affecter x à une tâche à laquelle il n'est pas affecté.

Interprétation des q_y comme prix

Dans un esprit proche de ce qui précéde, on considère un ensemble d'agents X , un ensemble de biens Y de même cardinal, et l'on suppose que chaque agent doit se voir attribuer un bien de Y . On note u_{xy} l'utilité que représente le bien y pour l'agent x , u_{xy} mesure en quelque sorte la satisfaction apportée à x s'il se voit attribuer le bien y . Maximiser la satisfaction globale correspond à un problème de type Monge discret

$$\max_{\varphi \in S_N} \sum_x u_{x\varphi(x)},$$

que l'on peut écrire sous la forme relaxée d'un problème de type Monge-Kantorovich discret

$$\max_{\gamma \in \Pi} \sum_{xy} \gamma_{xy} u_{xy},$$

qui admet une formulation duale comme nous l'avons vu précédemment. Considérons un plan optimal γ^φ de type bijection. Une solution (p, q) du problème dual est telle que

$$p_x = \max_y (u_{xy} - q_y) \quad \forall x \in X$$

où y réalise le maximum ci-dessus dès que $\gamma_{xy} > 0$, i.e. dès que $\varphi(x) = y$ (x se voit attribuer x). On peut ainsi considérer q comme une collection de *prix* (que l'on peut prendre positifs), qui ont pour effet de diminuer l'utilité effective (ou l'attrait affectif) de y pour l'ensemble des agents, de telle sorte que, les prix étant ce qu'ils sont, chaque agent maximise son utilité effective en choisissant $y = \varphi(y)$. La problème dual peut donc s'interpréter comme la recherche d'une collection de prix afférents aux différents biens qui assure que, pour l'attribution optimale, le bien attribué à chaque agent a une utilité (corrigée par le prix) qui le rend optimal. L'algorithme des enchères, qui fait l'objet de la section 14.9 ci-après, est basé sur cette interprétation des potentiels de Kantorovich.

14.4 Métriques induites sur l'ensemble des mesures de probabilités sur un espace métrique fini

On considère ici un espace métrique (V, d) , où V est un ensemble fini. C'est un espace canoniquement mesuré par la mesure de comptage sur V . Nous définissons dans cette section une famille de distances sur l'ensemble $\mathcal{P}(V)$ des mesures de probabilité sur V , associées à la métrique $d(\cdot, \cdot)$ sous-jacente. Pour $p \in [1, +\infty]$ fixé, μ et ν dans $\mathcal{P}(V)$, on note

$$W_p(\mu, \nu) = \left(\inf_{\gamma \in \Pi(\mu, \nu)} \sum \gamma_{xy} d(x, y)^p \right)^{1/p},$$

où l'infimum correspond au problème de MK discret (14.1), pour lequel l'existence d'un plan minimisant est établie dans 14.3.

Théorème 14.12. La fonction $W_p(\cdot, \cdot)$ définie ci-dessus sur $\mathcal{P}(V) \times \mathcal{P}(V)$ est une distance.

14.5 Métriques induites 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

$$\mu = \sum_{i=1}^N \mu_i \delta_{x_i}, \quad \mu_i > 0, \quad \sum_{i=1}^N \mu_i = 1.$$

L'entier N n'est pas fixé, mais on ne considère ici que des sommes finies. Pour $p \in [1, +\infty[$ fixé, μ et ν dans \mathcal{A} , 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 (14.1), pour lequel l'existence d'un plan minimisant est établie dans 14.3. On se propose de montrer que W_p est une distance sur \mathcal{A}_d .

Théorème 14.13. 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 suivante⁸

$$\gamma_{i_1 i_2 i_3}^{123} = \frac{1}{\mu_{i_2}^2} \gamma_{i_1 i_2}^{12} \gamma_{i_2 i_3}^{23}.$$

On note γ^{13} le plan défini de façon naturelle par

$$\gamma_{i_1 i_3}^{13} = \sum_{i_2} \gamma_{i_1 i_2 i_3}^{123}.$$

On a

$$\begin{aligned} W_p(\mu^1, \mu^3) &\leq \left(\sum_{i_1 i_3} \gamma_{i_1 i_3}^{13} |x_{i_3}^3 - x_{i_1}^1|^p \right)^{1/p} = \left(\sum_{i_1 i_2 i_3} \gamma_{i_1 i_2 i_3}^{123} |x_{i_3}^3 - x_{i_1}^1|^p \right)^{1/p} \\ &\leq \left(\sum_{i_1 i_2 i_3} \gamma_{i_1 i_2 i_3}^{123} |x_{i_2}^2 - x_{i_1}^1|^p \right)^{1/p} + \left(\sum_{i_1 i_2 i_3} \gamma_{i_1 i_2 i_3}^{123} |x_{i_3}^3 - x_{i_2}^2|^p \right)^{1/p} \end{aligned}$$

d'après l'inégalité de Minkowski (proposition ??, page ??), d'où finalement

$$W_p(\mu^1, \mu^3) \leq \left(\sum_{i_1 i_2} \gamma_{i_1 i_2}^{12} |x_{i_2}^2 - x_{i_1}^1|^p \right)^{1/p} + \left(\sum_{i_2 i_3} \gamma_{i_2 i_3}^{23} |x_{i_3}^3 - x_{i_2}^2|^p \right)^{1/p} = W_p(\mu^2, \mu^3) + W_p(\mu^1, \mu^2),$$

ce qui termine la preuve. □

8. 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 .

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

$$\mu = \sum_{i=1}^N \mu_i \delta_{x_i}, \quad \mu_i > 0, \quad \sum_{i=1}^N \mu_i = 1.$$

Proposition 14.14. (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

$$W_1(\mu, \nu) = \inf_{\gamma \in \Pi(\mu, \nu)} \sum_{ij} \gamma_{ij} |y_j - x_i| = \max_{\varphi \in \text{Lip}_1} \left(\sum_i \mu_i \varphi(x_i) - \sum_j \nu_j \varphi(y_j) \right),$$

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,

$$\sum_i \mu_i \varphi(x_i) - \sum_j \nu_j \varphi(y_j) = \sum_{i,j} \gamma_{ij} (\varphi(x_i) - \varphi(y_j)) \leq \sum_{i,j} \gamma_{ij} |y_j - x_i| = W_1(\mu, \nu). \quad (14.3)$$

Réiproquement, considérons une solution (p, q) du problème dual :

$$\sum_i p_i \mu_i + \sum_j q_j \nu_j = W_1(\mu, \nu) \text{ avec } p_i + q_j \leq c_{ij}, \quad p_i + q_j = c_{ij} \text{ sur } \text{supp}(\gamma).$$

On a, pour tout i , $p_i \leq c_{ij} - q_j$ pour tout j , avec égalité pour au moins un indice j , donc (voir remarque 14.11)

$$p_i = \min_j (c_{ij} - q_j).$$

Considérons maintenant la fonction

$$\varphi : x \mapsto \inf_j (|y_j - x| - q_j).$$

Cette fonction est 1-Lipschitzienne comme infimum de fonctions 1-Lipschitziennes⁹. Par ailleurs φ prend les valeurs du potentiel de Kantorovitch sur le support de μ :

$$\varphi(x_i) = \inf_j (|y_j - x_i| - q_j) = p_i.$$

Enfin, on a

$$\varphi(y_j) = \inf_k (|y_k - y_j| - q_k) \leq -q_j,$$

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 (14.3),

$$\sup_{\varphi \in \text{Lip}_1} \left(\sum_i \mu_i \varphi(x_i) - \sum_j \nu_j \varphi(y_j) \right) = W_1(\mu, \nu),$$

ce qui termine la preuve. □

9. 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 $\varphi(x) - \varphi(y) \leq |y - x|$.

14.6 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

$$\mu = \sum_{i=1}^N \mu_i \delta_{x_i}, \quad \mu_i > 0, \quad \sum_{i=1}^N \mu_i = 1, \quad x_1, \dots, x_N \in K,$$

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

$$W_p(\mu, \nu) = \left(\inf_{\gamma \in \Pi(\mu, \nu)} \sum \gamma_{ij} |y_j - x_i|^p \right)^{1/p}.$$

Lemme 14.15. Soient μ et ν deux mesures atomiques sur K . Pour toute fonction φ L -Lipschitzienne, on a

$$|\langle \nu - \mu, \varphi \rangle| \leq LW_p(\mu, \nu).$$

Démonstration. Soit γ un plan de transport optimal entre μ et ν . On a

$$\begin{aligned} \langle \nu - \mu, \varphi \rangle &= \sum_j \nu_j \varphi(y_j) - \sum_i \mu_i \varphi(x_i) = \sum_j \sum_i \gamma_{ij} (\varphi(y_j) - \varphi(x_i)) \\ &\leq L \sum_j \sum_i \gamma_{ij} |y_j - x_i| \leq L \left(\sum_j \sum_i \gamma_{ij} |y_j - x_i|^p \right)^{1/p} \end{aligned}$$

par convexité de la fonction $\alpha \mapsto \alpha^p$. Comme on a la même inégalité pour $-\varphi$, on a l'estimation sur $|\langle \nu - \mu, \varphi \rangle|$. \square

Proposition 14.16. 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

$$(\mu^n) \sim (\nu^n) \iff W_p(\mu^n, \nu^n) \rightarrow 0.$$

muni de la métrique induite, en notant $\bar{\mu}$ (respectivement $\bar{\nu}$) la classe d'équivalence de la suite (μ^n) (resp. (ν^n))

$$W_p(\bar{\mu}, \bar{\nu}) = \lim_{n \rightarrow +\infty} W_p(\mu^n, \nu^n).$$

Montrons que l'on peut associer de façon univoque un élément de $\mathcal{P}(K)$ à toute suite de Cauchy dans $\mathcal{A}(K)$, et que cette limite est la même pour tout autre représentant de la même classe.

Soit (μ^n) une suite de Cauchy dans \mathcal{A} . On peut extraire une sous-suite $(\mu^{g(n)})$ qui converge faiblement¹⁰ vers une mesure μ dans $\mathcal{P}(K)$. Montrons que toute la suite (μ^n) converge faiblement vers μ . On montre dans un premier temps la convergence contre des fonctions-test lipschitzien, puis simplement continues. D'après le lemme 14.15 on a, pour toute fonction φ lipschitzienne,

$$|\langle \mu^n - \mu^{g(n)}, \varphi \rangle| \leq LW_p(\mu^n, \mu^{g(n)}).$$

Comme μ^n est de Cauchy, toute suite extraite lui est adjacente, la quantité ci-dessus tend donc vers 0, d'où

$$\langle \mu^n, \varphi \rangle \rightarrow \langle \mu, \varphi \rangle.$$

10. 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).

Soit maintenant une fonction φ continue sur K . Par densité des fonctions Lipschitziennes dans les fonctions continues (K est compact), φ est limite uniforme d'une suite (φ^k) de fonctions L -Lipchitzienne. On a

$$\langle \mu^n - \mu, \varphi \rangle = \langle \mu^n - \mu, \varphi - \varphi^k \rangle + \langle \mu^n - \mu, \varphi^k \rangle. \quad (14.4)$$

Comme $\mu^n - \mu$ est une mesure uniformément bornée, le premier terme est inférieur à ε pour k assez grand. Le second terme peut être rendu arbitrairement petit pour n assez grand, d'où la convergence de $\langle \mu^n, \varphi \rangle$ vers $\langle \mu, \varphi \rangle$.

Soit (ν^n) un autre représentant de la classe $\bar{\mu}$. Toujours par le même lemme on a pour toute fonction φ Lipschitzienne,

$$\langle \mu^n - \nu^n, \varphi \rangle \leq LW_p(\mu^n, \nu^n),$$

d'où l'on déduit comme précédemment la convergence de $\langle \mu^n - \nu^n, \varphi \rangle$ vers 0 pour toute fonction φ continue.

Montrons maintenant que toute mesure de probabilité $\mu \in \mathcal{P}(K)$ peut être approchée faiblement par une suite de Cauchy de $\mathcal{A}(K)$. 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. \square

Remarque 14.17. 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 le lemme 14.15 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

$$\langle \mu - \mu_k, \varphi \rangle \longrightarrow 0$$

pour toute fonction φ continue sur K .

14.7 Régularisation entropique

On propose ici une démonstration alternative de l'existence d'un point-selle, plus laborieuse, mais qui permet d'étudier une méthode utilisée en pratique pour l'approximation effective du coût de transport entre deux mesures. Cette méthode est basée sur la *régularisée entropique* de la fonctionnelle $C(\gamma)$, définie par

$$\gamma \in \mathbb{R}_+^{NM} \longmapsto C_\varepsilon(\gamma) = \sum_{i,j} c_{ij} \gamma_{ij} + \varepsilon \sum_{i,j} \gamma_{ij} \log \gamma_{ij} = C(\gamma) + \varepsilon S(\gamma), \quad (14.5)$$

où S est l'entropie de la probabilité γ sur $\mathbb{R}^N \times \mathbb{R}^M$ (voir définition 1.10, page 25).

Lemme 14.18. On suppose que μ et ν chargent tous les points de X et Y , respectivement. La fonctionnelle C_ε définie par (14.10) admet un minimiseur γ^ε unique sur Π (défini par (20.12)), 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 admissibles¹¹.

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 $1 \rightarrow j$ et $i \rightarrow 1$. Pour le terme d'entropie correspondant à $1 \rightarrow 1$, 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. \square

Cette régularisation entropique 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 14.19. On se donne deux mesures (μ_i) et (ν_j) , et une collection de coûts (c_{ij}) . Il existe un unique minimiseur $\bar{\gamma}$ de l'entropie parmi les minimiseurs de la fonction coût $C(\cdot)$ sur $\Pi_{\mu,\nu}$.

Démonstration. L'ensemble des minimiseurs du coût est un fermé borné, l'infimum de la fonction continue $S(\cdot)$ est donc atteint. Comme l'ensemble des minimiseurs est convexe, et que l'entropie est strictement convexe, ce minimiseur est unique. \square

Proposition 14.20. On se donne deux mesures (μ_i) et (ν_j) , et une collection de coûts (c_{ij}) . On note γ^ε le minimiseur du problème régularisé (voir lemme 14.30), qui minimise

$$C_\varepsilon(\gamma) = \sum_{ij} \gamma_{ij} c_{ij} + \varepsilon \sum_{ij} \gamma_{ij} \log \gamma_{ij},$$

sur $\Pi_{\mu,\nu}$ (défini par (20.12)). Alors γ^ε converge vers $\bar{\gamma}$, plan qui minimise l'entropie parmi tous les minimiseurs admissibles de $C(\cdot)$ (voir proposition 14.19).

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

$$S_\varepsilon(\gamma) = \frac{1}{\varepsilon} (C(\gamma) - C_{opt}) + S(\gamma)$$

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 S_ε , on a la chaîne d'inégalité suivante

$$S(\gamma^\varepsilon) \leq S_\varepsilon(\gamma^\varepsilon) \leq S_\varepsilon(\bar{\gamma}) = S(\bar{\gamma}),$$

où $\bar{\gamma}$ est le minimiseur de l'entropie parmi les minimiseurs du coût (voir proposition 14.19). On a donc à la limite $S(\gamma^0) \leq S(\bar{\gamma})$. Par ailleurs, d'après l'inégalité $S_\varepsilon(\gamma^\varepsilon) \leq S(\bar{\gamma})$ ci-dessus, la quantité

$$\frac{1}{\varepsilon} (C(\gamma^\varepsilon) - C_{opt}) + S(\gamma)$$

11. 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.

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 les autres minimiseurs. On en déduit la convergence de toute la suite γ^ε vers $\bar{\gamma}$. \square

14.8 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 14.1, 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 équation (20.12)), 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

$$\gamma_{ij} c_{ij} = -\varepsilon \gamma_{ij} \log \left(e^{-c_{ij}/\varepsilon} \right),$$

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 loi¹² η . Cette entropie relative est aussi appelée *divergence de Kullback-Leibler*, et notée en conséquence $\text{KL}(\gamma|\eta)$. 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 13.33, page 271) il peut se mettre sous la forme

$$\gamma_{ij} = a_i b_j \eta_{ij}, \quad a_i > 0, \quad b_j > 0, \tag{14.6}$$

tout en vérifiant bien sûr les conditions de marginales :

$$a_i \sum_j b_j \eta_{ij} = \mu_i, \quad b_j \sum_i a_i \eta_{ij} = \nu_j. \tag{14.7}$$

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 sur X ou sur Y). Considérons une loi de probabilité $\bar{\gamma}$ sur $X \times Y$ (sans hypothèse sur les marginales). On cherche à minimiser l'entropie relative de γ relativement à $\bar{\gamma}$, sous la contrainte de marginale sur X :

$$\inf_{\gamma \in \Pi_\mu} \left(\sum_{i,j} \gamma_{ij} \log \left(\frac{\gamma_{ij}}{\bar{\gamma}_{ij}} \right) \right), \quad \Pi_\mu = \left\{ \gamma \in \mathbb{R}_+^{NM}, \quad \sum_j \gamma_{ij} = \mu_i \quad \forall i \right\}.$$

12. La densité η n'est pas nécessairement 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.

Du fait de la présence du log, les contraintes $\gamma_{ij} \geq 0$ ne sont pas activées (voir démonstration du lemme 14.30), et l'on a des multiplicateurs de Lagrange p_1, \dots, 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 14.21. On construit de façon itérative

$$\gamma^0 = \eta, \gamma^{1/2}, \gamma^1, \dots, \gamma^k, \gamma^{k+1/2}, \gamma^{k+1}, \dots$$

de la façon suivante :

$$\begin{aligned} \gamma_{ij}^{k+1/2} &= \gamma_{ij}^k \frac{\mu_i}{\sum_j \gamma_{ij}^k} \quad \left(\gamma^{k+1/2} = \arg \min_{\Pi_\mu} KL(\gamma|\gamma^k) \right) \\ \gamma_{ij}^{k+1} &= \gamma_{ij}^{k+1/2} \frac{\nu_j}{\sum_i \gamma_{ij}^{k+1/2}} \quad \left(\gamma^{k+1} = \arg \min_{\Pi_\nu} KL(\gamma|\gamma^{k+1/2}) \right). \end{aligned}$$

On peut voir cet algorithme de "projections"¹³ alternées comme un algorithme de point fixe sur le problème en a_i, b_j donné par les équations (14.6)-(14.7). En effet, si l'on prend pour a^0 et b^0 des vecteurs qui ne contiennent que des 1, et qu'on pose

$$\gamma_{ij}^0 = a_i^0 b_j^0 \eta_{ij}, \quad \gamma_{ij}^k = a_i^k b_j^k \eta_{ij}$$

une étape de l'algorithme précédent peut s'écrire

$$\begin{aligned} \gamma_{ij}^{k+1/2} &= \gamma_{ij}^k \frac{\mu_i}{\sum_j \gamma_{ij}^k} = a_i^k b_j^k \eta_{ij} \frac{\mu_i}{\sum_j a_i^k b_j^k \eta_{ij}} = b_j^k \underbrace{\left(\frac{\mu_i}{\sum_j b_j^k \eta_{ij}} \right)}_{a_i^{k+1}} \eta_{ij}, \\ \gamma_{ij}^{k+1} &= \gamma_{ij}^{k+1/2} \frac{\nu_j}{\sum_i \gamma_{ij}^{k+1/2}} = a_i^{k+1} b_j^k \eta_{ij} \frac{\nu_j}{\sum_i a_i^{k+1} b_j^k \eta_{ij}} = a_i^{k+1} \underbrace{\left(\frac{\nu_j}{\sum_i a_i^{k+1} \eta_{ij}} \right)}_{b_j^{k+1}} \eta_{ij}. \end{aligned}$$

L'algorithme se ramène finalement au calcul des $a^1, b^1, \dots, a^k, b^k, \dots$, selon la procédure

$$a_i^{k+1} = \frac{\mu_i}{\sum_j b_j^k \eta_{ij}}, \quad b_j^{k+1} = \frac{\nu_j}{\sum_i a_i^{k+1} \eta_{ij}}.$$

Remarquons en premier lieu que, si la suite des couples (a^k, b^k) converge vers (a, b) , alors le plan limite $\gamma_{ij} = a_i b_j \eta_{ij}$ vérifie (14.6)-(14.7), c'est donc le minimiseur recherché.

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)`.

13. Il ne s'agit pas à strictement parler de projection, car la divergence de Kulback-Leibler n'est pas une distance.

14.9 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

$$u_{i\varphi(i)} - q_{\varphi(i)} = \max_j (u_{ij} - q_j), \quad (14.8)$$

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 proposition 14.8) que le plan γ^φ associé à φ est optimal.

Algorithme 14.22. (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 (14.8), c'est terminé. Dans le cas contraire, on sélectionne un i^* pour lequel la relation est invalidée, i.e. tel que

$$u_{i^*\varphi^n(i^*)} - q_{\varphi^n(i^*)} < \max_j (u_{i^*j} - q_j).$$

On note j^* un indice qui réalise le max ci-dessus¹⁴ :

$$u_{ij^*} - q_{j^*} = \max_j (u_{i^*j} - q_j).$$

On attribue alors j^* à i^* , et $\varphi^n(i^*)$ à $(\varphi^n)^{-1}(j^*)$, i.e.

$$\varphi^{n+1}(i^*) = j^*, \quad \varphi^{n+1}((\varphi^n)^{-1}(j^*)) = \varphi^n(i^*)$$

ou, exprimé différemment,

$$\varphi^{n+1} = \varphi^n \circ \tau_{i^*, (\varphi^n)^{-1}(j^*)},$$

où τ_{i_1, i_2} est la transposition qui échange i_1 et i_2 . On augmente enfin le prix de j^* d'une quantité qui ramène l'attrait de j^* pour i^* au niveau du second produit le plus attractif :

$$q_{j^*}^{n+1} = q_{j^*}^n + \underbrace{\max_j (u_{i^*j} - q_j^n)}_{u_{i^*j^*} - q_{j^*}^n} - \max_{j \neq j^*} (u_{i^*j} - q_j^n).$$

Cet algorithme est susceptible de patiner dans certains cas, lorsque plusieurs produits 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

$$u_{i\varphi(i)} - q_{\varphi(i)} \geq \max_j (u_{ij} - q_j) - \varepsilon. \quad (14.9)$$

Algorithme 14.23. (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 (14.9), on s'arrête. Dans le cas contraire, on sélectionne un i^* pour lequel la relation est invalidée, i.e. tel que

$$u_{i^*\varphi^n(i^*)} - q_{\varphi^n(i^*)} < \max_j (u_{i^*j} - q_j) - \varepsilon.$$

14. L'agent i^* préférerait l'objet j^* qui, en l'état courant des prix, lui apporterait plus de satisfaction (= utilité - prix) que $\varphi^n(i^*)$.

On note j^* un indice qui réalise le max ci-dessus

$$u_{ij^*} - q_{j^*} = \max_j (u_{i^*j} - q_j).$$

On attribue alors j^* à i^* , et $\varphi^n(i^*)$ à $(\varphi^n)^{-1}(j^*)$, i.e.

$$\varphi^{n+1}(i^*) = j^*, \quad \varphi^{n+1}((\varphi^n)^{-1}(j^*)) = \varphi^n(i^*).$$

On augmente enfin le prix de j^* du montant maximum qui préserve son ε -satisfaction :

$$q_{j^*}^{n+1} = q_{j^*}^n + \max_j (u_{i^*j} - q_j) - \max_{j \neq j^*} (u_{i^*j} - q_j) + \varepsilon \geq q_{j^*}^n + \varepsilon.$$

Remarque 14.24. Noter que, dans cette ε -version de l'algorithme, le bien j^* choisi par i^* après une étape n'est pas forcément son meilleur choix (après augmentation du prix de j^*), mais l'agent est tout de même ε -satisfait avec son j^* , 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 faire qu'il renonce à j^* 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 14.25. L'algorithme 14.23 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. \square

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 13.27, page 269, i.e.

$$F(\gamma) \leq G(p, q) \quad \forall \gamma \in (\mathbb{R}_+)^{N^2}, \quad (p, q) \in \mathbb{R}^N \times \mathbb{R}^N.$$

Du fait de l'existence d'un point selle démontré au début de cette section (proposition 14.10), on a bien sûr

$$\sup F(\gamma) = \max F(\gamma) = \inf G(p, q) = \min G(p, q).$$

Proposition 14.26. Pour tout $\varepsilon > 0$, on considère une bijection φ de S_N et un système de prix (q_j) qui vérifient¹⁵

$$u_{i\varphi(i)} - q_{\varphi(i)} \geq \max_j(u_{ij} - q_j) - \varepsilon.$$

Alors l'utilité associée à la bijection φ approche l'utilité maximale à $N\varepsilon$ près, i.e.

$$U(\gamma^\varphi) \geq \max_{\Pi} U(\gamma) - N\varepsilon.$$

Démonstration. On définit

$$p_i = u_{i\varphi(i)} - q_{\varphi(i)}.$$

On a par hypothèse

$$p_i \geq u_{ij} - q_j - \varepsilon \quad \forall j$$

de telle sorte que le couple $(p + \varepsilon, q)$ est admissible. On a donc

$$\begin{aligned} \max F = \min G \leq G(p + \varepsilon, q) &= \sum (p_i + \varepsilon) + \sum q_j = \sum_i (p_i + q_{\varphi(i)}) + N\varepsilon \\ &= \sum_i u_{i\varphi(i)} + N\varepsilon = F(\gamma^\varphi) + N\varepsilon. \end{aligned}$$

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

$$u_{ij} = -|y_j - x_i|^p.$$

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 = 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` correspondant au i^* 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)`.

Le recherche d'un *argmax* peut se faire de la façon suivante. Considérons par exemple une bijection φ de S_N , encodée par un tableau d'entiers `phi`. On se donne un j , on cherche à déterminer l'antécédent i de j par φ . On peut utiliser pour cela l'instruction `np.where(phi==1)`, qui construit un *tuple de arrays* (ici un seul tableau) qui contiennent les indices pour lesquels la condition est satisfaite. On écrira donc

15. On écrit exactement ici que (φ, q) est un point d'arrêt de l'algorithme des enchères modifié.

```
i = np.where(phi==j)[0][0]
```

pour obtenir l'indice i tel que $\varphi(i) = j$. Noter que s'il en existait plusieurs, `np.where(phi==j)[0]` serait un tableau de taille le nombre d'antécédents.

Remarque 14.27. On prendra garde au fait que, à chaque itération, l'agent i^* choisit le (ou un) bien j^* 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 ε – satisfait, mais *pas mieux*. On aura toujours (mathématiquement), du fait de l'augmentation du prix,

$$u_{i^* \varphi^{n+1}(i^*)} - q_{\varphi^{n+1}(i^*)} = \max_j (u_{i^* j} - q_j) - \varepsilon,$$

où i^* , 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 -satisfaisants¹⁶, 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^* soit fausse, alors qu'elle devrait être vraie, du fait des erreurs d'arrondis. Même si la réalité mathématique 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 0.99ε . 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 sensibles¹⁷.

14.10 Compléments, extensions

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

$$\mu = \sum_{i=1}^N \mu_i \delta_{x_i}, \quad \mu_i > 0, \quad \sum_{i=1}^N \mu_i = 1, \quad \mu_i \geq 0.$$

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 entre 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. McCann¹⁸ en 1997, et on parle parfois d'interpolation *au sens de McCann*.

16. Il est naturel d'arrêter l'algorithme lorsque ce nombre vaut le nombre total d'agents.

17. Dans le cas présent il est assez aisés 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 1, on a de l'ordre d'une chance sur 10^{14} de tomber sur un cas ambigu de quasi-égalité. On aurait tort de négliger le problème sur la base de sa faible probabilité d'occurrence : c'est en fait beaucoup plus vicieux, puisque le problème risque de ne se poser qu'après un très grand nombre de tests de l'algorithme, et donc de ne pas être révélé par des batteries de tests préliminaires.

18. 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>

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 transport 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 14.28. 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 littérature 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).

L'ensemble des mesures de probabilités atomiques sur \mathbb{R}^d reste convexe pour cette nouvel acceptation 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 14.29. Si Ψ est une fonction strictement convexe de \mathbb{R}^d dans \mathbb{R} , régulière¹⁹, 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 propriétés de convexité de la fonction Ψ .

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 dernier²⁰. 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

$$|x_1 - x_0|^2 \leq \ell^2 = \left(\int_0^1 |v(s)| ds \right)^2 \leq \int_0^1 |v(s)|^2 ds.$$

Par ailleurs, si l'on prend la vitesse constante égale à $(x_1 - x_0)$, on a égalité entre les deux extrémités de la chaîne précédente d'inégalités. On a donc

$$|x_1 - x_0|^2 = \min_{V(x_0, x_1)} \int_0^1 |v(s)|^2 ds,$$

où $V(x_0, x_1)$ est l'espace des vitesses continues de $[0, 1]$ dans \mathbb{R}^d qui conduisent x_0 en x_1 , i.e. telles que $x_1 = x_0 + \int_0^1 v$. 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

19. À strictement parler il n'est pas nécessaire d'expliquer cette hypothèse car toute fonction convexe sur \mathbb{R}^d est localement Lipschitzienne, donc en particulier continue, ce qui permet de donner un sens au produit de dualité ci-dessous.

20. 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>

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

$$\min_{v \in W, \gamma \in \Pi} \left(\sum_{ij} \int_0^1 \gamma_{ij} |v_{ij}(s)|^2 ds \right)$$

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 (au sens de la définition 4.9, page 88) 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 .

14.11 Compléments sur la 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

$$\gamma \in \mathbb{R}_+^{NM} \longmapsto C_\varepsilon(\gamma) = \sum_{i,j} c_{ij} \gamma_{ij} + \varepsilon \sum_{i,j} \gamma_{ij} \log \gamma_{ij} = C(\gamma) + \varepsilon S(\gamma), \quad (14.10)$$

où S est l'entropie de la probabilité γ sur $\mathbb{R}^N \times \mathbb{R}^M$ (voir définition 1.10, page 25).

Lemme 14.30. On suppose que μ et ν chargent tous les points de X et Y , respectivement. La fonctionnelle C_ε définie par (14.10) admet un minimiseur γ^ε unique sur Π (défini par (20.12)), 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 admissibles²¹.

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 $1 \rightarrow j$ et $i \rightarrow 1$. Pour le terme d'entropie correspondant à $1 \rightarrow 1$, on a une

21. Celà pourrait ne pas être le cas que l'illustre l'exemple suivant. Un problème classique consiste à minimiser l'entropie de la densité d'un 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.

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. \square

Lemme 14.31. 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 14.30.

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 13.12, page 263, existence de multiplicateurs de Lagrange $(p^\varepsilon, q^\varepsilon) \in \mathbb{R}^N \times \mathbb{R}^M$ tels que

$$c_{ij} + \varepsilon(1 + \log \gamma_{ij}^\varepsilon) - p_i^\varepsilon - q_j^\varepsilon = 0. \quad (14.11)$$

On applique alors le corollaire 13.34 du théorème 13.33, page 271, qui assure que $(\gamma^\varepsilon, p^\varepsilon, q^\varepsilon)$ est point-selle du Lagrangien L_ε . \square

Lemme 14.32. 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

$$G_\varepsilon(p, q) = \sum_{i=1}^N p_i \mu_i + \sum_{j=1}^M q_j \nu_j + \inf_{\gamma \in V} \left(\sum_{i,j} (c_{ij} - p_i - q_j + \varepsilon \log \gamma_{ij}) \gamma_{ij} \right). \quad (14.12)$$

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

$$G_\varepsilon(p, q) = \sum_{i=1}^N p_i \mu_i + \sum_{j=1}^M q_j \nu_j - \varepsilon e^{-1} \sum_{i,j} e^{-\frac{c_{ij} - p_i - q_j}{\varepsilon}}. \quad (14.13)$$

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 $(p, q) \mapsto \sum e^{p_i + q_j}$ (on prend momentanément $\varepsilon = 1$ pour alléger l'écriture). Cette matrice H peut se décrire par blocs : 2 blocs diagonaux du type

$$D_p = \text{diag} \left(e^{p_i} \sum_j e^{q_j} \right)_i, \quad D_q = \text{diag} \left(e^{q_j} \sum_i e^{p_i} \right)_j,$$

et un bloc extra-diagonal supérieur $B = (e^{p_i + q_j})_{ij}$ (le bloc inférieur est ${}^t B$). On a

$$(\bar{p}, \bar{q}) \cdot H \begin{pmatrix} \bar{p} \\ \bar{q} \end{pmatrix} = \sum_i e^{p_i} \bar{p}_i^2 \sum_j e^{q_j} l + \sum_j e^{q_j} \bar{q}_j^2 \sum_i e^{p_i} + 2 \sum_{ij} \bar{p}_i \bar{q}_j e^{p_i + q_j}.$$

On a $2\bar{p}_i \bar{q}_j \geq -\bar{p}_i^2 - \bar{q}_j^2$, avec inégalité stricte dès que $\bar{q}_j \neq -\bar{p}_i$. Si l'on prend (\bar{p}, \bar{q}) non nul dans l'orthogonal de $(1, -1)$, on aura nécessairement $\bar{q}_j \neq -\bar{p}_i$ pour au moins l'un des couples (i, j) , d'où

$$(\bar{p}, \bar{q}) \cdot H \begin{pmatrix} \bar{p} \\ \bar{q} \end{pmatrix} > 0.$$

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 nul, c'est ce minimiseur particulier que nous noterons $(p^\varepsilon, q^\varepsilon)$ dans la suite.

□

Lemme 14.33. 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 (14.11), 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 (14.13), 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 $(p^\varepsilon, q^\varepsilon)/|(p^\varepsilon, q^\varepsilon)|$ 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 14.34. Le minimiseur γ^ε construit au lemme 14.30 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 (14.11), on obtient $p^0 \oplus q^0 \leq c$, avec

$$\gamma_{ij}^0 > 0 \implies p_i + q_j = \gamma_{ij},$$

d'où la conclusion. □

Remarque 14.35. 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 14.36. 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 14.37. 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 (14.1), i.e. γ minimise

$$C(\gamma) = \sum_{ij} \gamma_{ij} c_{ij},$$

sur $\Pi_{\mu,\nu}$ (défini par (20.12)), et γ^ε le minimiseur du problème régularisé (voir lemme 14.30), qui minimise

$$C_\varepsilon(\gamma) = \sum_{ij} \gamma_{ij} c_{ij} + \varepsilon \sum_{ij} \gamma_{ij} c_{ij},$$

sur $\Pi_{\mu,\nu}$. Alors γ^ε converge vers $\bar{\gamma}$, plan qui minimise l'entropie parmi tous les minimiseurs 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)

$$C_\varepsilon(\gamma) = \frac{1}{\varepsilon} (C(\gamma) - C_{opt}) + S(\gamma)$$

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 (où $\bar{\gamma}$ est le minimiseur de l'entropie parmi les minimiseurs du coût)

$$S(\gamma^\varepsilon) \leq C_\varepsilon(\gamma^\varepsilon) \leq C_\varepsilon(\bar{\gamma}) = S(\bar{\gamma}).$$

On a donc à la limite $S(\gamma^0) \leq S(\bar{\gamma})$. Par ailleurs, d'après l'inégalité $C_\varepsilon(\gamma^\varepsilon) \leq S(\bar{\gamma})$ ci-dessus, la quantité

$$\frac{1}{\varepsilon} (C(\gamma) - C_{opt}) + S(\gamma)$$

est bornée, avec $S(\gamma)$ minoré, et $C(\gamma^\varepsilon) - C_{opt} \geq 0$. 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 $\bar{\gamma}$.

□

14.12 Exercices

Exercice 14.1. 1) a) On considère deux mesures de $\mathcal{A}(\mathbb{R})$

$$\mu = \sum_{i=1}^N \delta_{x_i}, \quad \mu = \sum_{j=1}^N \delta_{y_j}, \quad x_i, y_j \in \mathbb{R}.$$

Décrire l'ensemble des plans de transport optimaux pour le coût $c_{ij} = |y_j - x_i|^p$, avec $p \in [1, +\infty[$ (on pourra distinguer les cas $p = 1$ et $p > 1$). On pourra aussi supposer les mesures linéairement séparées, au sens où les enveloppes convexes des supports ne se chevauchent pas, sauf éventuellement en un point.

b) Décrire de la même manière l'ensemble des plans optimaux dans le cas de mesures non uniformes (mais de même masse)

$$\mu = \sum_{i=1}^N \mu_i \delta_{x_i}, \quad \mu = \sum_{j=1}^N \nu_j \delta_{y_j}, \quad x_i, y_j \in \mathbb{R}.$$

2) On se place maintenant sur \mathbb{R}^2 , on considère deux mesures atomiques μ et ν supportées par deux collections finies de points X et Y .

a) On considère γ un plan de transport optimal pour le coût $c_{xy} = |y - x|$. On considère l'ensemble des segments $[x, y]$ tels que $\gamma_{xy} > 0$. Montrer que ces segments ne se croisent pas (on parle ici de croisement quand les segments ne sont pas inclus dans une même droite).

b) dans le contexte de la question précédente, montrer que l'on peut avoir croisement dès que le coût est strictement convexe (on pourra considérer le coût quadratique $c_{xy} = |y - x|^2$).

Exercice 14.2. (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.

Exercice 14.3. 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.

Exercice 14.4. Décrire, dans $(\mathcal{A}(\mathbb{R}^d), W_1)$, la sphère S dont le centre est un Dirac centré à l'origine, et de rayon 1.

Exercice 14.5. On considère X et Y deux ensembles finis, de cardinaux respectifs N et M , deux mesures discrètes $\mu = (\mu_i)$ et $\nu = (\nu_j)$ sur X et Y , respectivement, et une famille de coûts $(c_{ij}) \in \mathbb{R}^{NM}$ donnée. On supposera toutes les masses élémentaires μ_i et ν_j strictement positives. On suppose que la masse totale de ν est supérieure ou égale à celle de μ , qui vaut 1 :

$$|\mu| = \sum_{i=1}^N \mu_i = 1 \leq |\nu| = \sum_{j=1}^M \nu_j. \tag{14.14}$$

On s'intéresse aux plans γ qui transportent μ vers une mesure portée par Y dominée par ν . L'ensemble des plans de transports admissibles est donc

$$\Pi = \left\{ \gamma = (\gamma_{ij}) \in \mathbb{R}_+^{NM}, \sum_j \gamma_{ij} = \mu_i \quad \forall i, \sum_i \gamma_{ij} \leq \nu_j \quad \forall j \right\}. \tag{14.15}$$

On s'intéresse à la minimisation sur Π du coût associé à un plan γ , défini classiquement par

$$C(\gamma) = \sum_{i,j} c_{ij} \gamma_{ij}.$$

Pour tout $\gamma \in \Pi$, on notera $\gamma \sharp \mu$ la mesure image, définie sur Y par

$$(\gamma \sharp \mu)_j = \sum_i \gamma_{ij}.$$

- 1) On suppose dans cette première question que ν_j est égal à une constante $\beta > 0$ pour tout j .
 - a) Que devient le problème de minimisation décrit ci-dessus lorsque β tend vers $+\infty$?
 - b) Pour toute valeur de β admissible, on note C_β le coût optimal correspondant. Que peut-on dire de la manière dont C_β dépend de β ?
- 2) On se replace dans le cas général. Montrer que, sous l'hypothèse (20.11), Π n'est pas vide, et qu'il existe $\gamma \in \Pi$ qui minimise C sur Π .
- 3) Montrer que γ minimise C sur Π si et seulement s'il existe $p = (p_i) \in \mathbb{R}^N$ et $q = (q_j) \in \mathbb{R}^M$ tels que $p_i + q_j \leq c_{ij}$ pour tous i, j , avec égalité sur le support de γ , et $q_j = 0$ pour tous les j correspondant à des points non saturés, i.e. tels que $(\gamma \sharp \mu)_j < \nu_j$.
- 5) On considère la situation où $\nu_j = 1$ pour tout j (en supposant que la masse portée par Y est supérieure à la masse portée par X). Pour tout $\varepsilon > 0$, on considère la fonctionnelle

$$C^\varepsilon : \gamma \in \mathbb{R}_+^{N \times M} \mapsto \sum c_{ij} \gamma_{ij} + \varepsilon \sum_{j=1}^M \eta_j^{1/\varepsilon}.$$

où l'on a noté pour alléger l'écriture η_j la masse transportée en j , i.e. $\eta_j = (\gamma \sharp \mu)_j$ (il s'agit donc bien d'une fonction de γ).

- a) Montrer que C^ε admet un minimiseur γ^ε sur Π_μ , ensemble des plans de transports qui admettent μ comme première marginale ($\sum_j \gamma_{ij} = \mu_i$ pour tout i), sans aucune contrainte sur la mesure image.
- b) Montrer que, quand ε tend vers 0, γ^ε converge (à sous-suite extraite près) vers une solution du problème initial, c'est à dire un minimiseur de C sur Π (défini par l'équation (20.12)).

Chapitre 15

Analyse de sensibilité

Sommaire

15.1	Introduction	311
15.2	Sensibilité pour les problèmes d'optimisation	312
15.3	Méthode de l'état adjoint	315
15.3.1	Méthode de l'adjoint : principe général et exemples d'application	315
15.3.2	Cadre des équations aux dérivées partielles	318
15.3.3	Cadre abstrait	320
15.4	Exercices	321

15.1 Introduction

Tous les problèmes abordés dans cet ouvrage dépendent de données / paramètres. Par exemple la solution d'une EDO dépend de la condition initiale et du second membre, que l'on peut supposer dépendre de paramètres, en nombre fini ou infini. Dans un autre contexte, la fonctionnelle intervenant dans un problème de minimisation peut dépendre de paramètres (comme par exemple la position des atomes pour un problème de transport optimal discret). Si la minimisation est contrainte, les contraintes elles même peuvent dépendre de paramètres, dont la valeur va affecter le minimiseur.

Ce chapitre aborde sous différents points de vue la dépendance de la solution (ou d'une fonctionnelle dépendant de la solution) d'un problème vis-à-vis de paramètres.

Les motivations sont de divers types. En premier lieu, tout modèle mathématiques est basé sur des paramètres, et maîtriser la manière dont la solution, ou une observable du système (une fonction de la solution) dépend de ces différents paramètres, de façon en particulier à distinguer leur importance relative, jusqu'à potentiellement considérer que certains ne jouent en fait aucun rôle significatif. Inversement, si la solution à un problème dépend de façon très sensible d'un paramètre, à tel point qu'une perturbation de ce paramètre inférieure à la précision à laquelle on le connaît conduit à une modification significative de la solution, le modèle considéré est inutilisable en pratique. Dans un autre contexte, on se trouve couramment dans la situation suivante : on dispose d'un modèle (dit *direct*) qui dépend de paramètres dont on ne connaît pas la valeur, mais on observe la solution de ce problème. La démarche d'*identification de paramètres* consiste à inférer la valeur des paramètres à partir de l'observation dont on dispose. Dans cette optique, il est important de maîtriser la manière dont les différents paramètres affectent la solution.

N.B. : nous utilisons dans les sections qui suivent les notation traditionnelles dans les domaines de l'étude de la sensibilité ou du contrôle, en notant y la *variable d'état*, et u ou v la collection de paramètres.

15.2 Sensibilité pour les problèmes d'optimisation

On considère dans un premier temps un problème d'optimisation impliquant une fonctionnelle J qui dépend d'un inconnue principale y et d'une collection de paramètres u . La première propriété précise, sous des hypothèses fortes, la dépendance d'un minimiseur local y_u vis-à-vis de u .

La suite de la section précise la manière dont on peut s'affranchir de la connaissance de la dépendance $y \mapsto y_u$ pour préciser la manière dont la valeur du minimum $\Phi(u) = J(y_u, u)$ dépend localement de u .

Proposition 15.1. Soit J une fonctionnelle définie d'un ouvert $U \times W \subset \mathbb{R}^n \times \mathbb{R}^m$ dans \mathbb{R} . On suppose que, pour tout $u \in W$, $y \mapsto J(y, u)$ est deux fois continûment différentiable, de matrice hessienne $H(y, u)$ symétrique définie positive. On suppose que J et son gradient par rapport à y sont de plus continûment différentiable par rapport à u . Soit $u_0 \in W$ tel que J admette un minimiseur y_{u_0} sur U . Alors il existe un voisinage $B(y_{u_0}, \eta) \times B(u_0, \eta)$ de (y_{u_0}, u_0) tel que, pour tout $u \in B(u_0, \eta)$, $J(\cdot, u)$ admette un minimiseur unique y_u dans $B(y_{u_0}, \eta)$. La correspondance $u \mapsto y_u$ est différentiable, avec

$$D_u y_u = -H(y_u, u)^{-1}(D_u \nabla_y J)(y_u, u).$$

Démonstration. On a, par condition nécessaire d'optimalité,

$$f(u_0, y_{u_0}) = \nabla_y J(y_{u_0}, u_0) = 0.$$

Du fait que la différentielle de f par rapport à y (qui s'exprime matriciellement par la hessienne de J), les conditions du théorème des fonctions implicite 19.11, page 376 sont vérifiées. Il existe donc un voisinage $B(y_{u_0}, \eta) \times B(u_0, \eta)$ et une application $u \mapsto y_u$ tel que $f(u, y) = 0$, avec $(u, y) \in B(y_{u_0}, \eta) \times B(u_0, \eta)$ si et seulement $y = y_u$, et l'on a

$$D_u y_u = -H(y_u, u)^{-1}(D_u \nabla_y J)(y_u, u).$$

Comme la fonctionnelle $J(\cdot, u)$ est convexe pour tout u , la condition d'annulation du gradient de $J(\cdot, u)$ en y_u assure que y_u minimise J sur $B(y_{u_0}, \eta)$. \square

Proposition 15.2. Soit J une fonctionnelle C^1 sur un ouvert $U \times W \subset V \times X$, où V et X sont des espaces de Hilbert. On suppose que $y \mapsto J(y, u)$ admet un minimiseur y_u , sur U pour tout u dans W . On note $\Phi(u) = J(y_u, u)$ la valeur du minimum associée à u . On a alors, au voisinage de tout $u \in W$,

$$\Phi(u) + [D_u J](y_u, u) \cdot \delta u \leq \Phi(u + \delta u) + o(\delta u). \quad (15.1)$$

De plus, si $v \mapsto \Phi(v)$ est différentiable en u , alors

$$D\Phi(u) = [D_u J](y_u, u).$$

Démonstration. On a

$$\begin{aligned} \Phi(u) = J(y_u, u) \leq J(y_{u+\delta u}, u) &= J(y_{u+\delta u}, u + \delta u) - (J(y_{u+\delta u}, u + \delta u) - J(y_{u+\delta u}, u)) \\ &= J(y_{u+\delta u}, u + \delta u) - [D_u J](y_{u+\delta u}, u) + o(\delta u) \\ &= \Phi(u + \delta u) - [D_u J](y_u, u) + o(\delta u), \end{aligned}$$

qui établit (15.1).

Si Φ est différentiable en u , on a

$$\begin{aligned} \Phi(u + \delta u) &\geq \Phi(u) + [D_u J](y_u, u) \cdot \delta u + o(\delta u) \\ \Phi(u + \delta u) &= \Phi(u) + D\Phi(u) \cdot \delta u + o(\delta u). \end{aligned}$$

d'où

$$([D_u J](y_u, u) - D\Phi(u)) \cdot \delta u + o(\delta u) \geq 0$$

où δu est dans un voisinage de 0, d'où $D\Phi(u) = [D_u J](y_u, u)$. \square

Remarque 15.3. L'équation (15.1) exprime l'appartenance de $[D_u J](y_u, u)$ au *sous-différentiel* au sens de Fréchet (voir définition 13.45, page 274) de Φ .

Remarque 15.4. Il peut sembler étonnant que la différentielle du minimum ne dépende que de la dérivée partielle de la fonctionnelle vis-à-vis des paramètres. En effet on s'intéresse à la dépendance de $\Phi(u) = J(y_u, u)$, qui dépend de u à deux titres, dépendance directe de J vis-à-vis de u , et dépendance indirecte du fait que J dépend de y_u , qui dépend lui-même de u . Mais y_u n'est pas quelconque : il minimise $y \mapsto J(y, u)$. Dans le cas où tout est régulier, en particulier si $u \mapsto y_u$ est différentiable (ce que nous n'avons pas supposé ci-dessus, il n'est même pas supposé que y_u soit l'unique minimiseur), on a

$$DJ(u) = D_u [J(y_u, u)] = [D_u J](y_u, u) + [D_y J](y_u, u) \circ (D_u y_u).$$

Comme y_u minimise $J(\cdot, u)$, on a $(D_y J)(y_u, u) = 0$, le second terme ci-dessus est donc nul, et il ne reste que le premier. L'intérêt de la proposition 15.2 est qu'elle ne nécessite pas d'identifier la différentielle de y_u vis à vis de u .

On se reportera à l'exercice 15.1, page 321, pour une illustration de la propriété précédente pour un système masse-ressorts.

Problèmes avec contraintes

Nous montrons ci-dessous que la proposition 15.2 reste valide, sous certaines hypothèses, dans le cas d'une minimisation avec contraintes.

On considère une fonctionnelle J , continûment différentiable sur un ouvert U d'un espace de Hilbert V . On s'intéresse au problème de minimisation de J sur $U \cap K$, où l'ensemble admissible K est défini par

$$K = \{y \in V, \varphi_i(y) \leq 0, i = 1, \dots, m\} = \left\{ v \in V, \sum_{i=1}^m \varphi_i(y) \mu_i \leq 0 \quad \forall \mu \in \mathbb{R}_+^m \right\}.$$

Nous noterons $\varphi(y) \in \mathbb{R}^m$ le vecteur des contraintes, de telle sorte que l'appartenance à K s'exprime de façon duale $\langle \varphi(y) | \mu \rangle \leq 0$ pour tout $\mu \in \mathbb{R}_+^p$.

Nous notons $u \in X$ la variable de contrôle, qui contient les paramètres, et nous considérerons la situation où la fonctionnelle J dépend de u , puis la situation où φ dépend de u .

Nous noterons $\Phi(u)$ la valeur du minimum de J sur $U \cap K$, qui dépend de u au travers de la dépendance à J ou φ .

Minimisation sous contrainte, fonctionnelle dépendant des paramètres

Proposition 15.5. Soit J une fonctionnelle C^1 sur un ouvert $U \times W \subset V \times X$, où V et X sont des espaces de Hilbert. On suppose que $y \mapsto J(y, u)$ admet un minimiseur y_u , sur $U \cap K$ pour tout u dans W , et que ces problèmes de minimisation admettent une formulation point-selle, c'est à dire qu'il existe, pour tout $u \in B$, $\lambda_u \in \mathbb{R}_+^m$ tel que

$$J(y_u, u) + \langle \varphi(y_u) | \mu \rangle \leq J(y_u, u) + \langle \varphi(y_u) | \lambda_u \rangle \leq J(z, u) + \langle \varphi(z) | \lambda_u \rangle \quad \forall z \in U, \mu \in \mathbb{R}_+^p.$$

On note $\Phi(u) = J(y_u, u)$ la valeur du minimum associée à u . On a alors

$$\Phi(u) + D_u J(y_u, u) \cdot \delta u \leq \Phi(u + \delta u) + o(\delta). \tag{15.2}$$

De plus, si $v \mapsto \Phi(v)$ est différentiable en u , alors¹

$$D\Phi(u) = [D_u J](y_u, u).$$

Démonstration. On écrit que (y_u, λ_u) est point-selle. L'inégalité de droite pour $z = y_{u+\delta u}$ s'écrit

$$\begin{aligned} J(y_u, u) + \langle \varphi(y_u) | \lambda_u \rangle &\leq J(y_{u+\delta u}, u) + \langle \varphi(y_{u+\delta u}) | \lambda_u \rangle \\ &\leq J(y_{u+\delta u}, u + \delta u) + \langle \varphi(y_{u+\delta u}) | \lambda_u \rangle + J(y_{u+\delta u}, u) - J(y_{u+\delta u}, u + \delta u). \end{aligned}$$

On majore la sommes des deux premiers termes grâce à l'inégalité de gauche du point-selle pour le couple $(y_{u+\delta u}, \lambda_{u+\delta u})$, et on approche la seconde différence par la différentielle de J par rapport à u . On a donc

$$\begin{aligned} \Phi(u) = J(y_u, u) + \langle \varphi(y_u) | \lambda_u \rangle &\leq J(y_{u+\delta u}, u + \delta u) + \langle \varphi(y_{u+\delta u}) | \lambda_{u+\delta u} \rangle - D_u J(y_{u+\delta u}, u) \cdot \delta u + o(\delta u) \\ &= \Phi(u + \delta u) - D_u J(y_u, u) \cdot \delta u + o(\delta u), \end{aligned}$$

d'où l'inégalité (15.2). Si $v \mapsto \Phi(v)$ est différentiable en u , alors

$$(D_u J(y_u, u) - D\Phi(u)) \cdot \delta u \geq 0 + o(\delta u),$$

où δu décrit un voisinage de 0, cela impose donc la nullité du terme d'ordre 1 dans le membre de gauche, d'où l'égalité des différentielles. \square

Contraintes dépendant des paramètres

Nous supposons maintenant que ce sont les contraintes qui dépendent du champ de paramètres $u \in X$, on écrira donc

$$K = K_u = \{y \in V, \varphi_i(y, u) \leq 0, i = 1, \dots, m\} = \left\{ v \in V, \sum_{i=1}^m \mu_i \varphi_i(y, u) \leq 0 \quad \forall \mu \in \mathbb{R}_+^m \right\}.$$

Comme précédemment, nous noterons plus simplement la contrainte $\langle \varphi(y, u) | \mu \rangle \leq 0$ pour tout $\mu \in \mathbb{R}_+^m$.

Proposition 15.6. Soit J une fonctionnelle d'un ouvert $U \subset V$, où V est un espace de Hilbert. On considère u dans un espace de Hilbert X , et W un voisinage de u . On suppose que les fonctions $(y, v) \mapsto \varphi_i(y, v)$ sont différentiables sur $U \times W$. On suppose que $y \mapsto J(y)$ admet un minimiseur y_v , sur $U \cap K_v$ pour tout v dans W , et que ces problèmes de minimisation admettent une formulation point-selle, c'est à dire qu'il existe, pour tout $v \in W$, $\lambda_v \in \mathbb{R}_+^p$ tel que

$$J(y_v, v) + \langle \varphi(y_v, v) | \mu \rangle \leq J(y_v, v) + \langle \varphi(y_v, v) | \lambda_v \rangle \leq J(z) + \langle \varphi(z, v) | \lambda_v \rangle \quad \forall z \in U, \mu \in \mathbb{R}_+^m.$$

On note $\Phi(v) = J(y_v)$ la valeur du minimum associée à v . On a alors

$$\Phi(u) + \langle (D_u \varphi(y_u, u))^* \lambda_u | \delta u \rangle \leq \Phi(u + \delta u) + o(\delta u). \tag{15.3}$$

De plus, si $v \mapsto \Phi(v)$ est différentiable en u , alors

$$D\Phi(u) = ([D_u \varphi](y_u, u))^* \lambda_u.$$

Démonstration. On utilise l'inégalité de droite du point selle pour $v = u$, $z = y_{u+\delta u}$

$$\begin{aligned} J(y_u, u) + \langle \varphi(y_u, u) | \lambda_u \rangle &\leq J(y_{u+\delta u}) + \langle \varphi(y_{u+\delta u}, u) | \lambda_u \rangle \\ &\leq J(y_{u+\delta u}) + \langle \varphi(y_{u+\delta u}, u + \delta u) | \lambda_u \rangle + \langle \varphi(y_{u+\delta u}, u) | \lambda_u \rangle - \langle \varphi(y_{u+\delta u}, u + \delta u) | \lambda_u \rangle. \end{aligned}$$

1. On utilise dans l'équation qui suit des crochets : $[D_u J](y_u, u)$, pour bien mettre en évidence qu'il s'agit de la dérivée partielle de J par rapport à u prise en (y_u, u) , et non pas la différentielle par rapport à u de la quantité $J(y_u, u)$, que l'on noterait $D_u[J(y_u, u)]$. Il est important de bien faire la distinction même si, dans le cas présent, si tout est régulier, ces deux notions s'identifient (voir remarque 15.4).

On majore la sommes des deux premiers termes grâce à l'inégalité de gauche du point-selle pour le couple $(u + \delta u, \lambda_{u+\delta u})$, et on approche la seconde somme par la différentielle de J par rapport à u . On a donc

$$\begin{aligned}\Phi(u) = J(y_u) + \langle \varphi(y_u, u+) | \lambda_u \rangle &\leq J(y_{u+\delta u}) + \langle \varphi(y_{u+\delta u}, u + \delta u) | \lambda_{u+\delta u} \rangle - \langle D_u \varphi(y_{u+\delta u}, u) \cdot \delta u | \lambda_u \rangle + o(\delta u) \\ &= \Phi(u + \delta u) - \langle (D_u \varphi(y_{u+\delta u}, u))^* \lambda_u | \delta u \rangle + o(\delta u) \\ &= \Phi(u + \delta u) - \langle (D_u \varphi(y_u, u))^* \lambda_u | \delta u \rangle + o(\delta u),\end{aligned}$$

d'où l'inégalité (15.3). Si $v \mapsto \Phi(v)$ est différentiable en u , alors

$$(D_u J(y_u, u) - D\Phi(u)) \cdot \delta u \geq 0 + o(\delta u),$$

où δu décrit un voisinage de 0, cela impose donc la nullité du terme d'ordre 1 dans le membre de gauche, d'où l'égalité des différentielles. \square

15.3 Méthode de l'état adjoint

Nous nous intéressons ici à la minimisation de fonctionnelles du type

$$\Phi(u) = J(y_u),$$

où u est une variable dite *de contrôle*, et y_u une variable d'état associée univoquement à u . Dans les situations auxquelles nous nous intéresserons, la correspondance $u \mapsto y_u$ n'est pas donnée sous forme explicite, mais au travers d'une relation implicite, typiquement équation différentielle ordinaire ou équation aux dérivées partielles. La variable u joue le rôle d'un paramètre pour le problème considéré, et y_u est la solution associée à ce paramètre.

Motivation(s)

L'approche que nous allons présenter est notamment motivée par des considérations *numériques* : considérons par exemple le cas où u vit dans un espace de grande dimension m , et le calcul du y_u associé est "cher". Si l'on cherche à calculer ou approcher le gradient de J en u , une première méthode consiste à utiliser une approche de type différences finies : si l'on note (e_i) une base de l'espace dans lequel vit u , on estime les dérivées partielles de J par rapport aux composantes de u par

$$\frac{\partial \Phi}{\partial u_i}(u) \approx \frac{J(y_{u+\varepsilon e_i}) - J(y_u)}{\varepsilon}.$$

Un telle approche nécessite la résolution de $m + 1$ problèmes $u \mapsto y_u$. En outre, le choix d'un $\varepsilon > 0$ adapté (suffisamment petit pour que l'approximation soit précise, mais pas trop petit pour éviter des phénomènes d'instabilité numérique liés au calcul de la différence de deux quantités voisines). Nous verrons que l'introduction d'un problème dit *adjoint* permet de se ramener à la résolution de seulement 2 problèmes $u \mapsto y_u$, et de contourner le problème du choix d'un ε .

Cette approche peut aussi être motivée par des considérations plus *théoriques*. Dans le cas où la correspondance $u \mapsto J(y_u)$ est régulière, un minimiseur de Φ vérifie $\nabla \Phi(u) = 0$. Une identification de ce gradient en tout point permet ainsi d'écrire des conditions nécessaires d'optimalité.

15.3.1 Méthode de l'adjoint : principe général et exemples d'application

Comme indiqué au début de ce chapitre, nous nous intéressons ici à l'identification de la différentielle (que nous chercherons à identifier à un gradient) de fonctionnelles du type

$$u \mapsto \Phi(u) = J(y_u),$$

où u est une variable dite *de contrôle*, et y_u une variable d'état associée univoquement à u . Nous introduirons un Lagrangien associé à ce problème,

$$(y, u, p) \mapsto L(y, u, p)$$

somme de $J(y)$ (noter que dans cette définition la variable y est *dissociée* de u) et d'une expression duale de la relation entre u et y_u . Il s'agira d'une expression du type $\langle \Psi(y, u), p \rangle$, où $\Psi(y, u) = 0$ est la relation implicite qui permet de définir y_u à partir de u :

$$L(y, u, p) = J(y) + \langle \Psi(y, u), p \rangle.$$

Pour tout y associé à u , le lagrangien prend la valeur de la fonctionnelle, i.e.

$$L(y_u, u, p) = J(y_u) = \Phi(u).$$

On a donc², quel que soit p ,

$$DJ(u) = D_u(L(y_u, u, p)) = D_y L(y_u, u, p) \circ D_u y_u + D_u L(y_u, u, p). \quad (15.4)$$

L'idée générale consiste à choisir un p particulier qui annule $D_y L(y_u, u, p)$, donc le premier terme, ce qui permet de contourner le problème d'identification de $D_u y_u$. La différentielle est alors donnée par le second terme, estimé en (y_u, u, p) où p est ce p bien choisi.

Le problème adjoint sur p est obtenu en demandant précisément que $D_y L(y_u, u, p) = 0$.

Précisons cette dernière équation. On a

$$D_y L(y, u, p) \cdot \delta y = D_y J \cdot \delta y + \langle D_y \Psi \cdot \delta y \mid p \rangle = D_y J \cdot \delta y + (D_y \Psi)^* p \cdot \delta y.$$

Le problème adjoint s'écrit donc

$$(D_y \Psi(y_u, u))^* p = -D_y J(y_u).$$

Cadre linéaire

On considère ici le cas $y \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, et l'on cherche à minimiser

$$\Phi(u) = \frac{1}{2} |Cy_u - \bar{z}|^2,$$

où y_u est défini par

$$Ay_u = Bu,$$

avec $A \in \mathcal{M}_n(\mathbb{R})$ (supposée inversible), $B \in \mathcal{M}_{n,m}(\mathbb{R})$, $C \in \mathcal{M}_{p,n}(\mathbb{R})$, $\bar{z} \in \mathbb{R}^p$.

Bien que le caractère linéaire de la correspondance $u \mapsto y_u$ rende l'approche un peu artificielle (on peut ici se passer de la notion de problème adjoint³ pour identifier le gradient de Φ), nous décrivons dans ce cas simplifié la démarche qui sera généralisée à d'autres situations.

On définit le Lagrangien comme suit :

$$(y, u, p) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n \mapsto \frac{1}{2} |Cy - \bar{z}|^2 + (Bu - Ay) \cdot p.$$

2. On prendra garde à bien distinguer $D_u(L(y_u, u, p))$, différentielle de l'application qui à u associe $L(y_u, u, p)$, de l'expression visuellement voisine $D_u L(y_u, u, p)$, qui est la différentielle de $u \mapsto L(y, u, p)$ prise au point (y_u, u, p) . La variable y est figée dans ce second cas, alors qu'elle varie en fonction de u dans le premier cas.

3. On a en effet $y_u = A^{-1}Bu$, d'où

$$\Phi(u + \tilde{u}) = C^*(Cy_u - \bar{z}) \cdot \tilde{y} = C^*(Cy_u - \bar{z}) \cdot A^{-1}B\tilde{u} = B^*(A^*)^{-1}C^*(Cy_u - \bar{z}) \cdot \tilde{u}.$$

l'estimation du gradient de G en u peut donc se faire par la résolution d'un premier problème $Ay_u = Cu$, puis d'un second $A^* = C^*(Cy_u - \bar{z})$.

On applique la démarche générale décrite précédemment (autour de l'équation (15.15)), basée sur l'expression

$$D\Phi(u) = D_u(L(y_u, u, p)) = D_y L \circ D_u y_u + D_u L.$$

On a

$$D_y L(y_u, u, p) \tilde{y} = C^*(C y_u - \bar{z}) \cdot \tilde{y} - A^* p \cdot \tilde{y}.$$

le problème adjoint s'écrit donc

$$A^* p = C^T (C y_u - \bar{z}). \quad (15.5)$$

On note maintenant p la solution de ce problème adjoint. On a alors, pour ce p particulier,

$$D\Phi(u) = D_u(L(y_u, u, p)) = \underbrace{D_y L \circ D_u y_u}_{=0} + D_u L = D_u L$$

pris en (y_u, u, p) , d'où $D\Phi(u) \tilde{u} = p \cdot B \tilde{u} = B^* p \cdot \tilde{u}$. On a donc finalement

$$\nabla \Phi = B^* p,$$

où p est la solution de (15.5).

Problème adjoint dans le cas d'une EDO

On considère l'équation différentielle suivante, dans \mathbb{R}^n ,

$$\begin{cases} \dot{y} &= f(y, u, t) \\ y(0) &= y_0 \end{cases} \quad (15.6)$$

où u est un paramètre de contrôle qui vit dans l'espace $U = \mathbb{R}^m$. On s'intéresse à la dépendance d'une fonction de y (et éventuellement de u lui-même) vis-à-vis de la variable de contrôle u .

Contrôle de l'état final.

On s'intéresse dans un premier temps au cas où la fonctionnelle mesure l'écart entre l'état final et un point cible donné :

$$\Phi(u) = J(y_u) = \frac{1}{2} |y_u(T) - \bar{y}_T|^2.$$

L'objectif est de calculer la différentielle de J .

On introduit le Lagrangien

$$L(y, u, p) = \frac{1}{2} |y(T) - \bar{y}_T|^2 + \int_0^T (\dot{y}(t) - f(y, u, t)) \cdot p(t) dt,$$

où p est une fonction définie sur $[0, T]$.

Lorsque y est associé à u par (15.6), on le note y_u . On applique la démarche générale décrite autour de l'équation (15.15). L'approche consiste à trouver un p particulier qui annule $D_y L$.

On a

$$D_u L \tilde{u} = - \int_0^T (D_u f(y, u, t) \tilde{u}) \cdot p dt,$$

où $D_u f(y, u, t)$ est linéaire de \mathbb{R}^m dans \mathbb{R}^n . On peut identifier $D_u L$ à un vecteur de \mathbb{R}^m , qui s'identifie donc au gradient de J :

$$\nabla J(u) = - \int_0^T (D_u f(y, u, t))^* p.$$

Pour la différentielle par rapport à y , on réécrit tout d'abord le Lagrangien en intégrant par partie le second terme :

$$L(y, u, p) = \frac{1}{2} |y(T) - \bar{y}_T|^2 + \int_0^T (-f(y, u, t) \cdot p(t) - y(t) \cdot \dot{p}(t)) dt + y(T) \cdot p(T) - y(0) \cdot p(0).$$

On a donc

$$\langle D_y L, \tilde{y} \rangle = (y(T) - \bar{y}_T) \cdot \tilde{y} + \int_0^T (-\dot{p} - D_y f(y, u, t)^* p(t)) \cdot \tilde{y} + \tilde{y}(T) \cdot p(T).$$

On introduit maintenant le problème adjoint, à valeur *finale* prescrite :

$$\begin{cases} -\dot{p} &= D_y f(y, u, t)^* p(t) \\ p(T) &= -(y(T) - \bar{y}_T). \end{cases} \quad (15.7)$$

Pour un tel p , $D_y L = 0$, et donc

$$D_u J = D_y L \circ D_u y_u + D_u L = D_u L = \int_0^T (D_u f(y, u, t))^* p,$$

où p est solution de (15.7).

Fonctionnelle plus générale

On considère maintenant le cas

$$J(u) = \int_0^T |y - \bar{y}|^2 w(t) dt,$$

où $w(t) \geq 0$ est une fonction de poids, qui quantifie l'importance que l'on donne à la mesure au temps t .

On vérifie que le problème adjoint (rétrograde en temps) s'écrit

$$\begin{cases} -\dot{p} &= D_y f(y, u, t) p(t) - (y - \bar{y}) w \\ p(T) &= 0. \end{cases} \quad (15.8)$$

Le gradient de J s'écrit en fonction de la solution de ce problème

$$\nabla J = - \int_0^T (D_u f(y, u, t))^* p.$$

15.3.2 Cadre des équations aux dérivées partielles

Contrôle au travers du terme source

On considère ici un domaine Ω bornée régulier, de frontière Γ . La variable de contrôle u est une fonction définie sur un sous-domaine $\omega \subset \Omega$, la variable d'état la solution du problème de Poisson "chauffé" par u , et la fonction coût basée sur l'écart entre le variable d'état et une fonction observée \bar{y} sur un sous-domaine $\mathcal{O} \subset \Omega$. On a plus précisément

$$\begin{cases} -\Delta y_u &= u \mathbf{1}_\omega \\ y_u &= 0 \quad \text{sur } \Gamma. \end{cases} \quad (15.9)$$

La fonction d'observation est

$$J(u) = \frac{1}{2} \int_{\mathcal{O}} |y_u - \bar{y}|^2.$$

On introduit le lagrangien du problème

$$L(y, u, p) = \frac{1}{2} \int_{\mathcal{O}} |y - \bar{y}|^2 + \int_{\Omega} \nabla y \cdot \nabla p - \int_{\omega} up,$$

où l'on impose que y soit nul sur le bord du domaine. On a, comme dans les autres cas, que pour tout y associé à u , le lagrangien prend la valeur de la fonctionnelle, i.e.

$$L(y_u, u, p) = G(y_u) = J(u).$$

On a donc, quel que soit p ,

$$DJ(u) = D_u(L(y_u, u, p)) = D_y L \circ D_u y_u + D_u L,$$

pris en (y_u, u, p) avec p quelconque. L'approche consiste à trouver p tel que $D_y L(y_u, u, p)$ de façon à ce que seul le second terme (qui dépendra bien sûr du p particulier) demeure. Dans le cas considéré ici, on a

$$D_u L \tilde{u} = \int_{\omega} p \tilde{u}.$$

Par ailleurs

$$D_y L \tilde{y} = \int_{\mathcal{O}} (y - \bar{y}) \tilde{y} + \int_{\Omega} \nabla \tilde{y} \cdot \nabla p = \int_{\mathcal{O}} (y - \bar{y}) \tilde{y} + \int_{\Omega} (-\Delta p) \tilde{y}.$$

Le problème adjoint s'écrit donc

$$-\Delta p = -(y - \bar{y}) \mathbf{1}_{\mathcal{O}},$$

avec conditions de Dirichlet $p = 0$ sur Γ , et l'on a, pour ce p particulier, $\nabla J = p \mathbf{1}_{\omega}$.

Contrôle par le champ de conductivité

On suppose ici que la variable de contrôle (ou l'ensemble des paramètres que l'on cherche à identifier) est le champ de conductivité $u(x)$ au sein du domaine. On suppose la valeur de la variable d'état imposée au bord du domaine. Le problème définissant la variable d'état est donc

$$\begin{cases} -\nabla \cdot u \nabla y_u &= 0 \\ y_u &= y_{\Gamma} \quad \text{sur } \Gamma. \end{cases} \quad (15.10)$$

Nous considérerons le cas où la fonction coût mesure l'écart entre la variable d'état et un champ connu \bar{y} sur un sous-domaine \mathcal{O} :

$$J(u) = \frac{1}{2} \int_{\mathcal{O}} |y_u - \bar{y}|^2.$$

Le Lagrangien s'écrit

$$L(y, u, p) = \frac{1}{2} \int_{\mathcal{O}} |y - \bar{y}|^2 + \int_{\Omega} u \nabla y \cdot \nabla p,$$

où y est supposé vérifier la condition aux limites sur le bord de Ω , et p est nul sur le bord du domaine.

On a

$$D_u L \tilde{u} = \int_{\Omega} \nabla y \cdot \nabla p \tilde{u},$$

et

$$D_y L \tilde{y} = \int_{\mathcal{O}} (y - \bar{y}) \tilde{y} + \int_{\Omega} u \nabla \tilde{y} \cdot \nabla p = \int_{\mathcal{O}} (y - \bar{y}) \tilde{y} - \int_{\Omega} (\nabla \cdot u \nabla p) \tilde{y} + \int_{\Gamma} u \frac{\partial p}{\partial n} \tilde{y}.$$

Le terme de bord s'annule car, la valeur de y étant fixée sur Γ , sa variation y est nulle. Le problème adjoint est donc

$$\begin{cases} -\nabla \cdot u \nabla p &= -(y_u - \bar{y}) \mathbf{1}_{\mathcal{O}} \\ p &= 0 \quad \text{sur } \Gamma. \end{cases} \quad (15.11)$$

et

$$\nabla J = \nabla y_u \cdot \nabla p,$$

où p est la solution du problème adjoint.

Contrôle sur la frontière

On suppose maintenant que la variable de contrôle (ou l'ensemble des paramètres que l'on cherche à identifier) est la valeur de la variable d'état sur le bord du domaine. Le problème définissant la variable d'état est donc

$$\begin{cases} -\Delta y_u &= 0 \\ y_u &= u \quad \text{sur } \Gamma. \end{cases} \quad (15.12)$$

Nous considérerons encore ici le cas où la fonction coût mesure l'écart entre la variable d'état et un champ connu \bar{y} sur un sous-domaine \mathcal{O} :

$$J(u) = \frac{1}{2} \int_{\mathcal{O}} |y_u - \bar{y}|^2.$$

Le Lagrangien s'écrit

$$L(y, u, p, \lambda) = \frac{1}{2} \int_{\mathcal{O}} |y - \bar{y}|^2 + \int_{\Omega} \nabla y \cdot \nabla p + \int_{\Gamma} (y - u) \lambda,$$

où p est nul sur le bord Γ . On a

$$D_u L \tilde{u} = - \int_{\Gamma} \lambda \tilde{u}.$$

et

$$D_y L \tilde{y} = \int_{\mathcal{O}} (y - \bar{y}) \tilde{y} + \int_{\Omega} \nabla \tilde{y} \cdot \nabla p + \int_{\Gamma} \lambda \tilde{y} = \int_{\mathcal{O}} (y - \bar{y}) \tilde{y} + \int_{\Omega} (-\Delta p) \tilde{y} + \int_{\Gamma} \frac{\partial p}{\partial n} \tilde{y} + \int_{\Gamma} \lambda \tilde{y}.$$

Le problème adjoint est donc

$$\begin{cases} -\Delta p &= -(y_u - \bar{y}) \mathbb{1}_{\mathcal{O}} \\ p &= 0 \quad \text{sur } \Gamma, \end{cases} \quad (15.13)$$

et $\lambda = -\partial p / \partial n$. Pour ce λ particulier, on a donc

$$\nabla J = -\lambda = -\frac{\partial p}{\partial n}.$$

15.3.3 Cadre abstrait

Nous nous intéressons ici à une fonctionnelle qui dépend d'une variable de contrôle $u \in U$ par l'intermédiaire d'une variable d'état $y \in Y$, univoquement associée à u , i.e.

$$u \mapsto J(u) = G(y_u) \in \mathbb{R},$$

où y_u est reliée à u par une relation implicite

$$\Psi(y_u, u) = 0,$$

où $\Psi(y, u)$ appartient à un espace vectoriel normé P . Pour simplifier cette première présentation, nous supposerons que les espaces U , Y , et P sont des espaces de Hilbert, identifiés à leurs espaces duals respectifs, et nous noterons $\langle \cdot, \cdot \rangle$ la dualité correspondante sur chacun de ces espaces.

On écrit la contrainte (lien entre u et y) de façon duale

$$\langle \Psi(y_u, u), p \rangle = 0,$$

pour tout $p \in P$.

On introduit le lagrangien, défini sur l'espace produit entre variables d'état, variables de contrôle, et ce nouvel espace qui permet d'exprimer la contrainte de façon duale :

$$(y, u, p) \longmapsto L(y, u, p) = G(y) + \langle \Psi(y, u), p \rangle \in \mathbb{R}, \quad (15.14)$$

qui est défini pour des couples (y, u) quelconques (i.e. qui ne vérifient pas nécessairement le lien $\Psi(y_u, u) = 0$). Pour tout y associé à u , le lagrangien prend la valeur de la fonctionnelle, i.e.

$$L(y_u, u, p) = G(y_u) = J(u),$$

quel que soit p . La différentielle de J par rapport à u s'identifie donc à la différentielle par rapport u de l'application qui à u associe $L(y_u, u, p)$. Cette différentielle est donc (en supposant que toutes les dépendances sont régulières) somme d'un premier terme $D_y L \circ D_u y_u$, et d'un deuxième terme du fait de la dépendance explicite de L par rapport à u (second terme de (15.14)). On a

$$\langle \Psi(y, u + \tilde{u}), p \rangle - \langle \Psi(y, u), p \rangle = \langle (D_u \Psi(y, u)) \tilde{u}, p \rangle + o(\tilde{u}) = \langle (D_u \Psi(y, u))^* p, \tilde{u} \rangle + o(\tilde{u}),$$

la contribution est donc $(D_u \Psi(y, u))^* p$ (exprimé au travers de la dualité⁴ $\langle \cdot, \cdot \rangle$ sur $U \times U$).

On a donc

$$D_u J = D_y L \circ D_u y_u + (D_u \Psi)^* p, \quad (15.15)$$

avec

$$D_y L = D_y G + (D_y \Psi)^* p. \quad (15.16)$$

L'idée est alors de construire un p particulier qui annule $D_y L$, et donc le premier terme de (15.15). Il n'est alors plus nécessaire de connaître la différentielle de y_u par rapport à u pour exprimer $D_u J$: on obtient, la dualité choisie sur U étant celle du produit scalaire,

$$\nabla J = (D_u \Psi)^* p,$$

où p a été construit de façon à annuler $D_y L$ (expression donnée par (15.16)).

15.4 Exercices

Exercice 15.1. a) On considère (comme dans la section 13.9, page 281) $n + 1$ masses sur l'axe réel, on note x_0, x_1, \dots, x_n les positions de ces masses, que l'on suppose reliées par n ressort de même raideur k . L'énergie potentielle du système s'écrit

$$J(x) = \frac{1}{2} k \sum_{i=1}^n |x_i - x_{i-1}|^2 = \frac{1}{2} k \langle Ax | x \rangle.$$

On fixe les extrémités en 0 et 1, respectivement, de façon *essentielle*, c'est à dire qu'on ne traite pas cette condition comme une contrainte (comme on l'avait fait dans la section 13.9), mais que l'on fixe x_0 à 0 et x_N à 1. Le vecteur des variables est noté $x = (x_1, \dots, x_{n-1})$. Montrer que J admet un minimiseur unique sur \mathbb{R}^{n-1} , et écrire les conditions d'optimalité sous la forme d'un système linéaire en x .

b) On considère maintenant que les raideurs des n ressorts sont susceptibles de varier, indépendamment les unes des autres i.e. qu'elles deviennent $k + u_1, k + u_2, \dots, k + u_n$. Montrer que, si les perturbations sont petites, la nouvelle fonctionnelle d'énergie qui dépend de $u = (u_1, \dots, u_n)$ admet toujours un unique minimiseur $x_u \in \mathbb{R}^{n-1}$, et écrire les conditions d'optimalité sous la forme d'un nouveau système linéaire en x . On note $E(u) = J(x_u)$ la nouvelle valeur de l'énergie potentielle. Expliquer pourquoi il est délicat d'exprimer x_u en fonction de u , et montrer que le gradient de E (vis-à-vis de u) s'exprime en revanche très simplement.

Exercice 15.2. Vue d'ensemble : on introduit dans ce problème une méthode d'estimation du gradient d'une fonctionnelle $F(u)$ à valeurs réelles, définie comme $g(y_u)$, où y_u est solution d'un problème de type

$$f(u, y) = 0. \quad (\star)$$

4. Dans l'hypothèse, que nous avons faite, où $\langle \cdot, \cdot \rangle$ est le produit scalaire sur l'espace de Hilbert U , il s'agit en fait d'un gradient, mais nous conserverons la notion de différentielle pour souligner le caractère générique de la démarche.

considéré comme un problème en y , paramétré par u . On pourra dans la suite assimiler la différentielle d'une application à la matrice jacobienne qui la représente dans les bases canoniques sur les espaces d'arrivée et de départ.

On considère une application f de $\mathbb{R}^n \times \mathbb{R}^m$ dans \mathbb{R}^m , continument différentiable sur un ouvert W de $\mathbb{R}^n \times \mathbb{R}^m$. On notera $u \in \mathbb{R}^n$ la variable dite de contrôle, et $y \in \mathbb{R}^m$ la variable d'état.

On considère $(u_0, y_0) \in \mathbb{R}^n \times \mathbb{R}^m$ solution de $f(u_0, y_0) = 0$, et tel que $\partial_y f(u_0, y_0)$ est inversible.

1) Montrer qu'il existe un ouvert U contenant u_0 et un ouvert V contenant y_0 tels que, sous l'hypothèse que $f(u, y) = 0$, on puisse exprimer y fonction de u sur $U \times V$, plus précisément qu'il existe une application Ψ de U dans V telle que

$$\forall (u, y) \in U \times V, \quad f(u, y) = 0 \iff y = \Psi(u),$$

avec Ψ continument différentiable sur U .

Préciser l'expression de la différentielle de Ψ en $u \in U$.

Pour désigner le y associé à u comme solution de l'équation (\star) , on pourra utiliser indifféremment $\Psi(u)$ ou y_u .

2) On considère maintenant une application g de \mathbb{R}^m dans \mathbb{R} , et l'on définit la fonction F de $U \subset \mathbb{R}^n$ dans \mathbb{R} par

$$F(u) = g(y_u) = g \circ \Psi(u),$$

où y_u est la solution de (\star) associée à u .

a) Exprimer $dF(u)$.

b) Si l'on souhaite estimer la différentielle de F en u_0 , sans avoir à inverser la matrice $\partial_y f(u, y_u)$, on peut approcher chacune des dérivées partielles par

$$\partial_{u_j} F(u) \approx \frac{F(u + he_j) - F(u)}{h},$$

pour un certain $h > 0$ petit, où les e_i sont les vecteurs de la base canonique de \mathbb{R}^n . Combien cette approche nécessite-t-elle de résolutions de problèmes (consistant à trouver y à u fixé), de type (\star) ?

3) On considère dans cette question l'exemple suivant : on suppose $n = 1$, $m \geq 1$, et l'on note $A(u)$ une matrice de $\mathcal{M}_m(\mathbb{R})$ dont les coefficients dépendent de la variable scalaire u :

$$A(u) = (a_{ij}(u))_{1 \leq i, j \leq n}.$$

On suppose que les $u \mapsto a_{ij}(u)$ sont des fonctions continument différentiables sur \mathbb{R} . Le problème définissant y fonction de u s'écrit sous la forme d'un système matriciel en y

$$A(u) \cdot y = b,$$

où b est un vecteur donné de \mathbb{R}^m . On suppose que $A(u_0)$ est inversible pour un certain $u_0 \in \mathbb{R}$. On définit g comme

$$y \in \mathbb{R}^m \longmapsto g(y) = \frac{1}{2} \|y - \bar{y}\|^2,$$

où $\bar{y} \in \mathbb{R}^m$ est donné.

a) On pose $f(u, y) = A(u) \cdot y - b$.

Préciser les expressions de $\partial_y f$ et $\partial_u f$ en fonction de la matrice A et de sa dérivée $A' = (a'_{ij})$.

N.B. : On pourra effectuer pour cela les développements limités de $f(u + \delta u, y)$ et $f(u, y + \delta y)$

b) Montrer que ce problème rentre dans le cadre général de la question 1, en déduire que l'on peut exprimer y en fonction de u sur un voisinage U de u_0 , sous la forme $y = \Psi(u)$, et donner l'expression de la différentielle de Ψ .

N.B. : cette différentielle étant une application de \mathbb{R} dans \mathbb{R}^m , on pourra la représenter par une matrice colonne.

- c) Préciser la différentielle de l'application $y \in \mathbb{R}^m \mapsto g(y) = \frac{1}{2} \|y - \bar{y}\|^2$, en expliquant pourquoi on peut l'identifier à une matrice ligne. Quel est le lien entre cette matrice ligne et le gradient de g (pour le produit scalaire canonique sur \mathbb{R}^m) ?

Expliquer pourquoi la différentielle de $u \mapsto F(u)$ peut être assimilée à un réel, et préciser l'expression de cette différentielle en u_0 .

Chapitre 16

Méthode des différences finies

Sommaire

16.1 La méthode	324
16.2 Consistance, stabilité, convergence	326
16.3 Analyse des principaux schémas numériques	330
16.4 Symboles discret et continu des opérateurs différentiels	332
16.5 Interprétation probabiliste de schémas explicites	336
16.6 Extensions, développements	339
16.7 Implémentation effective	339
16.8 Exercices	342

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 $[0, T]$:

$$\partial_t u - D \partial_{xx} u = 0, \quad u(\cdot, 0) = u^0(\cdot) \text{ donné.}$$

Nous considérerons ici, à titre d'illustration, le cas de conditions de Dirichlet homogènes :

$$u(0, t) = u(1, t) = 0.$$

On introduit une discrétisation uniforme de l'intervalle I , de pas $\Delta x = 1/J$:

$$0 = x_0, \quad x_1 = \Delta x, \quad \dots, \quad x_j = j\Delta x, \quad \dots, \quad x_{J-1} = (J-1)\Delta x, \quad x_J = J\Delta x, \quad (16.1)$$

et de même pour l'intervalle en temps (de pas $\Delta t = T/N$)

$$0 = t_0, \quad t_1 = \Delta t, \quad t_n = n\Delta t, \quad t_N = N\Delta t = T.$$

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

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} - D \frac{u_{j-1}^n - 2u_j^n + u_{j+1}^n}{(\Delta x)^2} = 0 \quad \forall j = 1, \dots, J-1, \quad (16.2)$$

ce qui peut s'écrire matriciellement, avec des notations évidentes

$$u^{n+1} = \left(\text{Id} - \frac{D\Delta t}{(\Delta x)^2} A \right) u^n,$$

où A est la matrice du Laplacien discret (avec condition de Dirichlet) définie par (19.16). 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

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} - D \frac{u_{j-1}^{n+1} - 2u_j^{n+1} + u_{j+1}^{n+1}}{(\Delta x)^2} = 0 \quad \forall j = 1, \dots, J-1. \quad (16.3)$$

Ce schéma peut s'écrire de façon matricielle :

$$\left(\text{Id} + \frac{D\Delta t}{(\Delta x)^2} A \right)^{-1} u^{n+1} = u^n.$$

On peut vérifier qu'il s'agit bien d'un schéma qui permet de construire u^{n+1} sans ambiguïté à partir de u^n , du fait que la matrice ci-dessus est à diagonale strictement dominante, donc inversible.

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 cette définition porte sur le schéma lui-même, dans sa version native : si l'on connaît l'inverse de la matrice impliquée dans le schéma, il devient de fait explicite, et l'on peut montrer que la matrice associé est pleine (tous ses éléments sont non nul), ce qui exprime au niveau discret le caractère non-local de l'inverse du Laplacien, et la propagation à vitesse infinie de la matière. Sous cette forme explicitée du schéma, le 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

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + V \frac{u_j^n - u_{j-1}^n}{\Delta x} = 0 \quad \forall j = 1, \dots, J \quad (\text{avec } 0 \equiv J), \quad (16.4)$$

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 densités négatives, et la densité maximale augmente au fil des itérations.

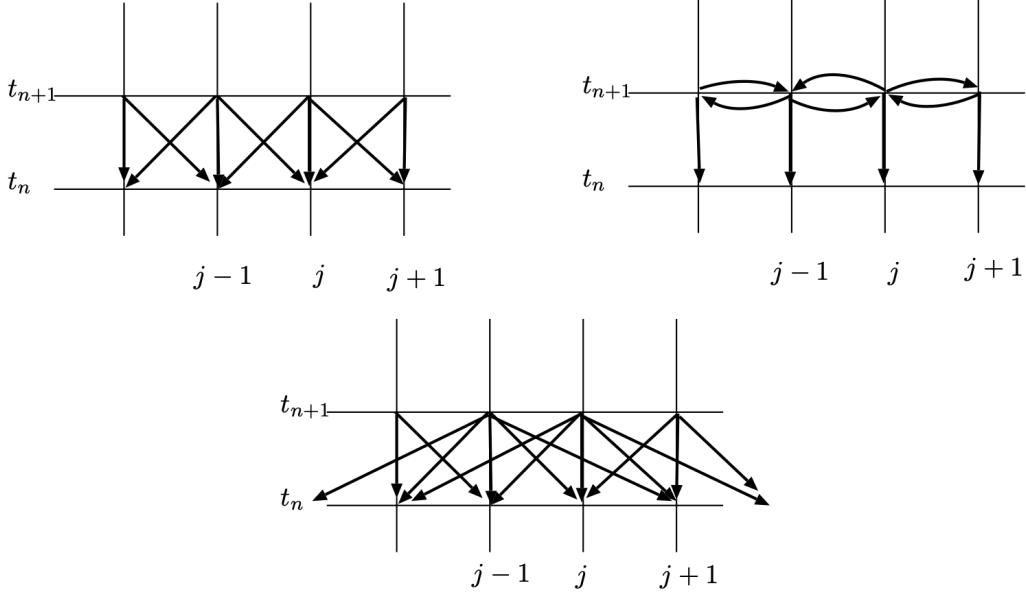


FIGURE 16.1 – Graphes de dépendance associés aux schémas explicite (gauche) et implicite (droite) pour l'équation de la chaleur. Le graphe du bas correspond au graphe effectif du schéma explicite après le “peignage” du graphe par inversion de la matrice.

16.2 Consistance, stabilité, convergence

On considère ici une équation aux dérivées partielles d'ordre 1 en temps :

$$\partial_t u + L(u) = 0.$$

où L est un opérateur différentiel en espace, linéaire (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 $(u_j^n)_j$ et $(u_j^{n+1})_j$, qui permet de calculer de façon unique les secondes à partir des premières. Comme c'est l'usage pour définir la notion de consistance, nous nous limiterons ici au cas de conditions aux limites périodique, en gardant à l'esprit que la prise en compte d'autres conditions devra faire l'objet d'une étude particulière. De façon à écrire le schéma de façon concise, pour toute partie finie Λ_0 de \mathbb{Z} , on note $u_{j+\Lambda_0}^n$ la collection des valeurs u_{j+k}^n pour k parcourant Λ_0 . Par exemple pour $\Lambda_0 = \{-1, 0, 1\}$ (qui permet d'écrire le schéma explicite pour l'équation de la chaleur), on a $u_{j+\Lambda}^n = \{u_{j-1}^n, u_j^n, u_{j+1}^n\}$. On définit de même $u_{j+\Lambda_1}^{n+1}$ associé à $\Lambda_1 \subset \mathbb{Z}$.

Les schémas numériques que nous considérons s'écrivent alors de la façon suivante :

$$F(u_{j+\Lambda_1}^{n+1}, u_{j+\Lambda_0}^n, \Delta t, \Delta x) = 0, \quad (16.5)$$

pour tout $j = 1, \dots, J$ (nombre de points de discréttisation en espace), tout $n = 0, \dots, N - 1$ (pas de temps). Nous ne considérerons ici que des schémas *linéaires*, qui peuvent s'écrire de façon matricielle¹

$$u^{n+1} = Au^n, \quad (16.6)$$

mais la définition pourrait s'appliquer à des schémas non linéaires.

1. 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).

On parlera de schéma numérique lorsque, pour tout n , les relations ci-dessus pour $j = 1, \dots, J$ permettent de déterminer $u^{n+1} = (u_j^{n+1})_j \in \mathbb{R}^J$ de façon unique à partir de u^n .

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 discréttisation (16.5) pour une équation aux dérivées partielles. Soit u une solution exacte, régulière, de l'équation. Pour une discréttisation donnée, on note $\tilde{u} \in \mathbb{R}^{J \times (N+1)}$ l'interpolée d'une solution exacte aux points de discréttisation, i.e.

$$\tilde{u}_j^n = u(j\Delta x, n\Delta t).$$

S'il existe une constante C (qui dépend de normes uniformes de dérivées en temps et en espace de la solution exacte) telle que

$$F(\tilde{u}_{j+\Lambda_1}^{n+1}, \tilde{u}_{j+\Lambda_0}^n, \Delta t, \Delta x) \leq C((\Delta x)^q + (\Delta t)^r)$$

uniformément en j et n , on dit que le schéma est consistant, d'ordre q en espace, et r en temps².

Remarque 16.3. Pour lever le flou sur la régularité requise, précisons la démarche qui permet d'établir l'ordre de consistance d'un schéma : 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 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 formelle, elle est afférente au schéma lui-même, on pourrait imaginer un schéma d'ordre très élevé qui discréttise 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 tant 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 indépassable³.

Remarque 16.4. On peut écrire le schéma à l'aide de la matrice A sous la forme (16.6)), qui est obtenue en multipliant le schéma écrit sous forme canonique par Δt , en regroupant les termes implicite, et en inversant la matrice correspondante. L'injection de la solution exacte dans le schéma écrit sous cette forme vérifie donc

$$\max_{n,j} |(\tilde{u}^{n+1} - A\tilde{u}^n)_j| \leq C\Delta t ((\Delta x)^q + (\Delta t)^r),$$

pour un schéma d'ordre (q, r) . On notera la présence du facteur Δt , qui vient simplement du fait que, pour obtenir cette écriture, le schéma natif a été multiplié par Δt .

Nous aurons besoin pour comparer la solution approchée 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 discréttisation en espace-temps) une fonction définie partout (ou au moins presque partout). On associe ainsi à une collection $u = (u_j)$ de valeurs aux points de discréttisation x_j la fonction constante, égale à u_j sur l'intervalle $]x_j - \Delta x/2, x_j + \Delta x/2[$. On notera \bar{u} cette fonction.

On peut alors exprimer la norme $\|\bar{u}\|_p$ en fonction des valeur discrètes, par exemple pour $p = 1, 2, +\infty$,

$$\|\bar{u}\|_1 = \Delta x \sum_j |u_j|, \quad \|\bar{u}\|_2 = \left(\Delta x \sum_j |u_j|^2 \right)^{1/2}, \quad \|\bar{u}\|_\infty = \max_j |u_j|.$$

2. 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) une 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 .

3. Sous réserve que les développements de Taylor aient été effectués de façon optimale.

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 16.5. (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 K telle que

$$\|\bar{u}^n\|_p \leq K \|\bar{u}^0\|_p \quad \forall n = 1, \dots, N = T/\Delta t,$$

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 16.6. Noter que la notion de stabilité n'est pas liée à l'équation discrétisée, mais au schéma elle-même. On pourrait imaginer selon cette définition des schémas stables qui n'ont aucun lien avec une EDP.

Remarque 16.7. Insistons ici sur l'abus de notation qui est couramment pratiqué par souci de lisibilité. Comme précédemment, u^n (ou \bar{u}^n) est ambigu, puisque ce vecteur dépend aussi de Δx et Δt (sa taille en particulier dépend de Δx). On devrait en toute rigueur noter $u_{\Delta x, \Delta t}^n$, ce que l'on ne fait pas pour alléger les notations.

Remarque 16.8. 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 $(0, 0)$, 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 $(0, 0)$.

Remarque 16.9. On a, pour tout $u = (u_j)$ et \bar{u} la fonction en escalier associée,

$$\|\bar{u}\|_p = C \|u\|_{\ell^p},$$

où la constante C dépend de Δx (on a par exemple $C = \sqrt{\Delta x}$ pour $p = 2$). La stabilité peut donc s'exprimer à l'aide de la matrice qui intervient dans l'écriture matricielle du schéma linéaire (voir (16.6)), elle revient à une majoration uniforme de la norme (subordonnée à la norme p) de A^n :

$$\|A^n\|_p \leq K.$$

La remarque 16.7 s'applique bien sûr à la matrice A , qui devrait en toute rigueur être notée $A_{\Delta t, \Delta x}$.

Le théorème suivant⁴ établit qu'un schéma consistant et stable est convergent, à l'ordre de consistance.

Théorème 16.10. (Lax)

On considère une équation 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 (avec q et r strictement positifs), 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.

Pour alléger les notations on considère que (u^n) (ainsi que e^n introduit ci-dessous) désigne à la fois, selon le contexte, la famille de vecteurs des inconnues aux points de discrétisation, ainsi que la famille de fonctions constantes par morceaux obtenues à partir de ces valeurs, avec $u^0 = \bar{u}^0$ (interpolée de U_0 aux points de discrétisation). On introduit l'erreur $e^n = \bar{u}^n - u^n$. On a la convergence de la méthode numérique au sens suivant

$$\lim_{\Delta t, \Delta x \rightarrow 0} \sup_n \|e^n\|_p = 0$$

(où e^n représente la fonction en escalier associée au champ de valeurs). On a plus précisément

$$\sup_n \|e^n\|_p \leq C ((\Delta x)^q + (\Delta t)^r).$$

Il est entendu ici que, dans le cas où la stabilité est conditionnelle, les paramètres Δx et Δt tendent vers 0 en vérifiant la condition de stabilité.

4. On pourra se reporter à l'article original :
<https://pdfs.semanticscholar.org/59b8/d99b13931ceb08a43700f6719760f1c35881.pdf>

Démonstration. Le schéma s'écrit $u^n = Au^{n-1}$. Comme il est consistant, la solution exacte le vérifie approximativement, plus précisément (voir remarque 16.4)

$$\tilde{u}^n = A\tilde{u}^{n-1} + \Delta t \varepsilon^n \quad \|\varepsilon^n\|_\infty \leq C ((\Delta x)^q + (\Delta t)^r)$$

(la consistance implique 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, $e^n = Ae^{n-1} + \Delta t \varepsilon^n$, et donc (d'après la remarque 16.9)

$$\|e^n\|_p = \left\| A^n \underbrace{e^0}_{=0} + \Delta t \sum_{k=0}^n A^k \varepsilon^{n-k} \right\|_p \leq CKT ((\Delta x)^q + (\Delta t)^r),$$

car $\Delta t = T/N$. □

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 $(u_j)_j$ on associe comme précédemment une fonction \bar{u} constante par morceaux sur les intervalles centrés en

$$0, \Delta x, 2\Delta x, \dots, J\Delta x = 1,$$

(avec identification du dernier intervalle au premier). Cette fonction de L^2 peut s'écrire comme la somme de sa série de Fourier

$$\bar{u}(x) = \sum_{k \in \mathbb{Z}} \hat{u}_k \exp(2i\pi kx) \text{ p.p. avec } \hat{u}_k = \int_0^1 \exp(-2i\pi kx) \bar{u}(x) dx,$$

et la formule de Parseval s'écrit

$$\|\bar{u}\|_{L^2}^2 = \int_0^1 |\bar{u}(x)|^2 dx = \sum_{k \in \mathbb{Z}} |\hat{u}_k|^2.$$

Maintenant, pour $x = j\Delta x$, on a $u_j^n = \bar{u}^n(x)$,

$$u_{j+1}^n = \sum_{k \in \mathbb{Z}} \hat{u}_k^n \exp(2i\pi kx) \exp(2i\pi k\Delta x),$$

Et une expression similaire pour u_{j-1}^n . Considérons par exemple le schéma explicite (16.2) pour l'équation de la chaleur, il peut s'écrire

$$\frac{\bar{u}^{n+1}(x) - \bar{u}^n(x)}{\Delta t} - D \frac{\bar{u}^n(x + \Delta x) - 2\bar{u}^n(x) + \bar{u}^n(x - \Delta x)}{(\Delta x)^2} = 0. \quad (16.7)$$

En remplaçant les \bar{u}^n et \bar{u}^{n+1} par leurs expressions en 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

$$\begin{aligned} \hat{u}_k^{n+1} &= \hat{u}_k^n \left(1 + \frac{D\Delta t}{(\Delta x)^2} (\exp(2i\pi k\Delta x) - 2 + \exp(-2i\pi k\Delta x)) \right) \\ &= \hat{u}_k^n \left(1 + \frac{D\Delta t}{(\Delta x)^2} (\exp(i\pi k\Delta x) - \exp(-i\pi k\Delta x))^2 \right) = \underbrace{\left(1 - 4 \frac{D\Delta t}{(\Delta x)^2} \sin^2(\pi k\Delta x) \right)}_{A(k)} \hat{u}_k^n. \end{aligned}$$

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 (suffisante) de stabilité

$$\Delta t \leq \frac{(\Delta x)^2}{2D}.$$

Remarque 16.11. 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

$$(1 + c\Delta t)^n = (1 + cT/N)^n \leq (1 + cT/N)^N \leq e^{cT}.$$

16.3 Analyse des principaux schémas numériques

Équation de transport

Proposition 16.12. 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

$$\hat{u}_k^{n+1} = \hat{u}_k^n \left(1 - \frac{V\Delta t}{\Delta x} (1 - \exp(-2i\pi k\Delta x))\right)$$

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 particulier⁵, bizarrement stable (conditionnellement) pour la norme L^2 , mais instable pour la norme L^∞ .

Proposition 16.13. Le schéma centré pour l'équation de transport

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + V \frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} = 0 \tag{16.8}$$

est instable en norme L^∞ , mais stable en norme L^2 sous la condition $\Delta t = \mathcal{O}((\Delta x)^2)$.

5. Il est souvent indiqué comme inconditionnellement instable dans la littérature.

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 le principe du maximum. Cela n'exclut pas à strictement parler la stabilité L^∞ (on ne demande pas que la constante K de la définition de stabilité 16.5, page 328 soit inférieure à 1, mais on peut vérifier numériquement que le schéma est effectivement instable. L'étude de stabilité L^2 conduit à

$$\hat{u}_k^{n+1} = \hat{u}_k^n(1 - \lambda \exp(2i\pi k\Delta x) + \lambda \exp(-2i\pi k\Delta x)) = \hat{u}_k^n(1 - 2i\lambda \sin(2\pi k\Delta x))$$

Le coefficient d'amplification a donc un module de carré 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 16.11). \square

Équation de la chaleur

Proposition 16.14. 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

$$\Delta t \leq \frac{(\Delta x)^2}{2D}.$$

Démonstration. Montrons d'abord la consistance du schéma. Soit $u(\cdot, \cdot)$ une solution exacte de l'équation. On a (les fonctions et dérivées sont prises en (x_j, t_n) sauf mention contraire) :

$$\begin{aligned} u(x_{j+1}, t_n) &= u + \Delta x \partial_x u + \frac{(\Delta x)^2}{2} \partial_{xx} u + \frac{(\Delta x)^3}{6} \partial_{xxx} u + \frac{(\Delta x)^4}{24} \partial_{xxxx} u(x_j + \theta^+ \Delta x, t_n) \\ -2u(x_j, t_n) &= -2u(x_j, t_n) \\ u(x_{j-1}, t_n) &= u - \Delta x \partial_x u + \frac{(\Delta x)^2}{2} \partial_{xx} u - \frac{(\Delta x)^3}{6} \partial_{xxx} u + \frac{(\Delta x)^4}{24} \partial_{xxxx} u(x_j - \theta^- \Delta x, t_n), \end{aligned}$$

avec $\theta^-, \theta^+ \in]0, 1[$. On a de même

$$u(x_j, t_{n+1}) = u + \Delta t \partial_t u + \frac{(\Delta t)^2}{2} \partial_{tt} u(x_j, t_n + \theta \Delta t),$$

d'où (\tilde{u} désigne l'interpolée de la solution exacte)

$$\begin{aligned} \frac{\tilde{u}_j^{n+1} - \tilde{u}_j^n}{\Delta t} - D \frac{\tilde{u}_{j-1}^n - 2\tilde{u}_j^n + \tilde{u}_{j+1}^n}{(\Delta x)^2} &= \underbrace{(\partial_t u - D \partial_{xx} u)(x_j, t_n)}_{=0} \\ -D \frac{(\Delta x)^4}{24} (\partial_{xxxx} u(x_j + \theta^+ \Delta x, t_n) + \partial_{xxxx} u(x_j - \theta^- \Delta x, t_n)) &+ \frac{(\Delta t)^2}{2} \partial_{tt} u(x_j, t_n + \theta \Delta t), \end{aligned}$$

d'où une erreur de consistance majorée par

$$C \left(\Delta t \sup_{I \times [0, T]} |\partial_{tt} u(x, t)| + (\Delta x)^2 \sup_{I \times [0, T]} |\partial_{xxxx} u(x, t)| \right).$$

Stabilité L^∞ .

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 valeurs précédentes sous la condition CFL $\Delta t \leq (\Delta x)^2/2D$.

Stabilité L^2 .

On écrit

$$\hat{u}_k^{n+1} = \hat{u}_k^n \left(1 + \frac{D\Delta t}{(\Delta x)^2} (\exp(2i\pi k\Delta x) - 2 + \exp(-2i\pi k\Delta x)) \right)$$

$$= \hat{u}_k^n \left(1 + \frac{D\Delta t}{(\Delta x)^2} (\exp(i\pi k\Delta x) - \exp(-i\pi k\Delta x))^2 \right) = \hat{u}_k^n \left(1 - \frac{4D\Delta t}{(\Delta x)^2} \sin(\pi k\Delta x)^2 \right)$$

qui est bien de module ≤ 1 sous la même condition sur le pas de temps. \square

Proposition 16.15. 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 ,

$$u_j^{n+1} + \lambda(u_j^{n+1} - u_{j-1}^{n+1}) + \lambda(u_j^{n+1} - u_{j+1}^{n+1}) = u_j^n,$$

avec $\lambda = D\Delta t/(\Delta x)^2 > 0$. 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

$$\hat{u}_k^{n+1} = \hat{u}_k^n \left(1 + \frac{4D\Delta t}{(\Delta x)^2} \sin(\pi k\Delta x)^2 \right)^{-1},$$

d'où l'inconditionnelle stabilité L^2 . \square

Exercice 16.1. Étudier (consistance et stabilité L^2) le θ -schéma pour l'équation de la chaleur

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + \theta D \frac{-u_{j-1}^{n+1} + 2u_j^{n+1} - u_{j+1}^{n+1}}{(\Delta x)^2} + (1-\theta) D \frac{-u_{j-1}^n + 2u_j^n - u_{j+1}^n}{(\Delta x)^2} = 0 \quad (16.9)$$

en fonction de la valeur de θ . Montrer en particulier que le schéma est inconditionnellement stable pour tout $\theta \in [1/2, 1]$.

Exercice 16.2. Faire l'étude complète (consistance, stabilité, convergence) du schéma de Lax-Wendroff pour l'équation de transport :

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + \frac{V}{2} \frac{u_{j+1}^n - u_{j-1}^n}{\Delta x} - \frac{V^2 \Delta t}{2} \frac{u_{j-1}^n - 2u_j^n + u_{j+1}^n}{(\Delta x)^2} = 0 \quad (16.10)$$

On montrera en particulier que ce schéma est d'ordre 2 en temps et en espace, qu'il est stable en norme L^2 sous la condition CFL usuelle, mais qu'il ne vérifie pas le principe du maximum.

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

$$u(x, t) = \sum_{\mathbb{Z}} \hat{u}_k^t \exp(2i\pi kx),$$

avec, pour chaque coefficient de Fourier, l'équation différentielle

$$\frac{d}{dt} \hat{u}_k^t + \hat{L}(k) \hat{u}_k^t = 0,$$

où $\hat{L}(k)$ est appelé *symbole* de l'opérateur L . Il s'agit d'un polynôme en k à coefficients complexes (coefficients d'ordre impair imaginaires purs, et coefficients d'ordre pair réels), tel que

$$L(\exp(2i\pi kx)) = \hat{L}(k) \exp(2i\pi kx).$$

Pour l'équation de la chaleur, on a par exemple

$$Lu = -D\partial_{xx}u, \quad \hat{L}(k) = 4D\pi^2k^2,$$

et pour le transport

$$Lu = V\partial_xu, \quad \hat{L}(k) = 2i\pi kV.$$

Si l'on discrétise en temps (par un schéma d'Euler explicite) l'équation différentielle sur \hat{u}_k^t , on obtient

$$\hat{u}_k^{n+1} = \hat{u}_k^n \left(1 - \Delta t \hat{L}(k)\right).$$

Il apparaît qu'un tel schéma est génériquement instable pour les modes de haute fréquence ($\hat{L}(k)$ 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 tronque 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

$$\Lambda(k) = \frac{D}{(\Delta x)^2} (-\exp(2i\pi k\Delta x) + 2 - \exp(-2i\pi k\Delta x)) = 4\frac{D}{(\Delta x)^2} \sin(\pi k\Delta x)^2$$

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 en k , 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 $1 - \Delta t\Lambda(k)$ est dans le disque unité pour tout k (cette condition n'est pas nécessaire à strictement parler, voir remarque 16.11, 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 $|1 - \Delta t \hat{L}(k)| > 1$ 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

$$\Lambda(k) = \frac{V}{\Delta x} (1 - \exp(-2i\pi k\Delta x))$$

qui est bien équivalent au symbole continu, à k fixé, quand Δx tend vers 0. Mais il n'est pas imaginaire pur, et l'on a bien

$$|1 - \Delta t\Lambda(k)| \leq 1 \text{ dès que } \Delta t \leq V/\Delta x.$$

Cette stabilisation par discrétisation s'accompagne d'un phénomène dit de *diffusion numérique*, qui apparaît clairement au niveau spectral. Le symbole 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

$$\frac{d}{dt} \hat{u}_k^t = -\hat{L}(k) \hat{u}_k^t = -2i\pi kV \hat{u}_k^t$$

est bien de module constant. Plus précisément, 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 .

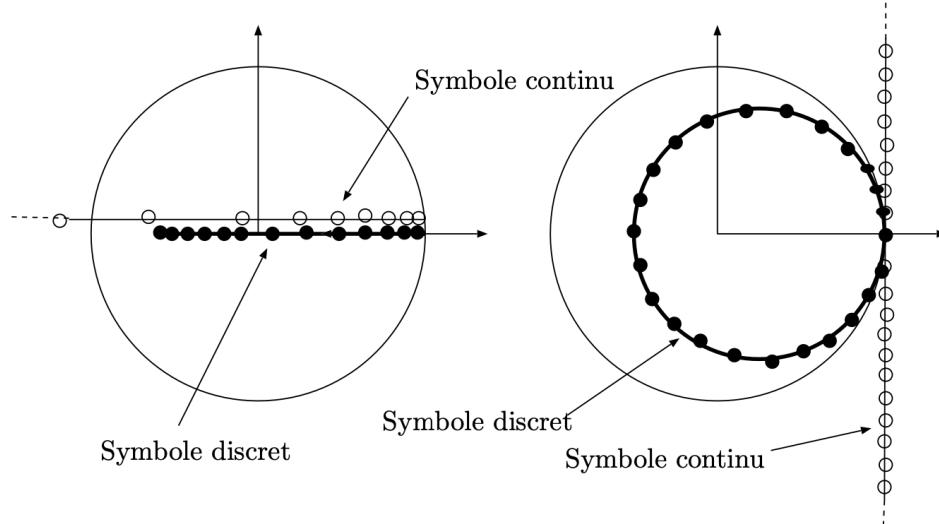


FIGURE 16.2 – Image des symboles discrets (ronds noirs) et continus (ronds blancs) pour l'équation de la chaleur (gauche) et l'équation de transport (droite). Plus précisément, la figure représente les symboles après transformation $z \mapsto 1 - \Delta t z$.

Par discréétisation en espace, chaque mode $2i\pi kV$ est remplacé par un mode tourné $V(1 - \exp(-2i\pi k\Delta x)) / \Delta x$, qui stabilise l'évolution, mais qui n'est plus imaginaire pur, on a une partie réelle non triviale

$$\text{Re}(\Lambda) = \frac{V}{\Delta x} (1 - \cos(2\pi k\Delta x)).$$

Le pendant discréétisé en espace de l'équation différentielle ci-dessus est

$$\frac{d}{dt} \hat{u}_k^t = -\Lambda(k) \hat{u}_k^t = -\frac{V}{\Delta x} (1 - \exp(-2i\pi k\Delta x)) \hat{u}_k^t,$$

qui correspond à une décroissance exponentielle vers 0 pour les modes non constants (i.e. pour $k \neq 0$) : tous les modes oscillants sont amortis.

Dans le processus d'évolution des modes de Fourier de la solution discrète, cela conduit au fait que les coefficients d'amplification $A(k) = (1 - \Delta t \Lambda(k))$ sont de module strictement inférieur à 1 (pour k différent de 0, et d'un multiple du nombre total de points), ce qui entraîne 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). Pour $k = 0$ (mode constant), on a un coefficient d'amplification égal à 1, ce qui exprime la conservation de la masse totale. Ainsi, la solution discrète va converger vers une constante, de telle sorte que la masse totale soit conservée. Ce comportement très éloigné de la solution exacte peut sembler en contradiction avec le résultat de convergence de la méthode. Il n'en est rien : le résultat de convergence porte sur un intervalle de temps $[0, T]$ fixé, sur lequel on va en effet avoir convergence vers le profil initial transporté sans déformation. En revanche, quels que soient les paramètres (en dehors du cas très particulier $V\Delta t/\Delta x = 1$), on s'éloigne de la solution exacte en temps long.

On peut quantifier plus précisément ce phénomène de diffusion numérique, ainsi que la manière dont la discréétisation en espace modifie la vitesse de transport des modes de Fourier de haute fréquence. Pour mettre en lumière le rôle joué par la discréétisation en espace, on s'intéresse ici au problème semi-discréétisé en espace :

$$\frac{d}{dt} \hat{u}_k^t = -\Lambda(k) \hat{u}_k^t = -\frac{V}{\Delta x} (1 - \exp(-2i\pi k\Delta x)) \hat{u}_k^t.$$

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

$$\operatorname{Re}(-\Lambda(k)) = -\frac{V}{\Delta x} (1 - \cos(2\pi k \Delta x)),$$

est strictement négative pour tous les modes en dehors de $k = 0$ (ou k multiple de J), qui correspond au fonctions constantes. Cette négativité des parties réelles pour les modes oscillants 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 :

$$\operatorname{Im}(-\Lambda(k)) = -\frac{V}{\Delta x} \sin(2\pi k \Delta x).$$

La partie de la solution associée à ce mode imaginaire s'écrit en effet

$$\exp\left(-i\frac{V}{\Delta x} \sin(2\pi k \Delta x)t\right) \exp(2i\pi kx) = \exp\left(2i\pi k \underbrace{\left(x - \frac{V}{2\pi k \Delta x} \sin(2\pi k \Delta x)t\right)}_{=x-V_k t}\right),$$

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 une 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 16.16. 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 16.17. (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 $\operatorname{Re}(z) \geq 0$ (ce qui est bien le cas pour les opérateurs de diffusion et de transport). On a en effet, pour le mode k ,

$$\frac{d}{dt} \hat{u}_k^t = -\hat{L}(k) \hat{u}_k^t,$$

et donc décroissance du (module du) coefficient correspondant au mode k dès que $\operatorname{Re}(\hat{L}(k)) \geq 0$. Pour le problème semi-discrétisé en temps, l'approche explicite s'écrit

$$\hat{u}_k^{n+1} = \left(1 - \Delta t \hat{L}(k)\right) \hat{u}_k^n$$

d'où, comme on l'a vu précédemment, une instabilité inconditionnelle sauf dans les cas triviaux. Le schéma implicite s'écrit

$$\hat{u}_k^{n+1} = \left(1 + \Delta t \hat{L}(k)\right)^{-1} \hat{u}_k^n,$$

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.

$\text{Re}(\Lambda(k)) \geq 0$, 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 conditionnelle⁶. Toujours sous l'hypothèse $\text{Re}(\Lambda(k)) \geq 0$, le schéma implicite

$$\hat{u}_k^{n+1} = (1 + \Delta t \Lambda(k))^{-1} \hat{u}_k^n,$$

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

$$\hat{u}_k^{n+1} = (1 - \Delta t \Lambda(k)) \hat{u}_k^n, \quad \Lambda(k) = \frac{V}{\Delta x} (\exp(2i\pi k \Delta x) - 1),$$

on a cette fois instabilité inconditionnelle : le symbole discret pointe dans la mauvaise direction (vers les parties réelles positives), la situation est donc désespérée. Le schéma implicite s'écrirait

$$\hat{u}_k^{n+1} = (1 + \Delta t \Lambda(k))^{-1} \hat{u}_k^n$$

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 16.8).

16.5 Interprétation probabiliste de schémas explicites

Certains schémas de discrétisation par différences finies peuvent s'interpréter de façon probabiliste. 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 325, peut s'écrire

$$u_j^{n+1} = \left(1 - 2 \frac{D \Delta t}{(\Delta x)^2}\right) u_j^n + \frac{D \Delta t}{(\Delta x)^2} u_{j-1}^n + \frac{D \Delta t}{(\Delta x)^2} u_{j+1}^n \quad \forall j = 0, \dots, J-1,$$

(avec la convention naturelle $0 \equiv J$ et $-1 \equiv J-1$). Considérons $u^n = (u_j^n)_{0 \leq j \leq J-1}$ comme une mesure discrète de probabilité, le schéma s'écrit

$$u^{n+1} = {}^t P u^n,$$

6. 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 $\text{Re}(z) \leq 1$ (comme pour le schéma centré explicite, voir proposition 16.8).

avec⁷

$${}^t P = \begin{pmatrix} 1 - 2\lambda & \lambda & 0 & \cdot & \cdot & \cdot & \lambda \\ \lambda & 1 - 2\lambda & \lambda & 0 & \cdot & \cdot & \cdot \\ 0 & \lambda & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 1 - 2\lambda & \cdot & \cdot \\ \lambda & \cdot & \cdot & 0 & \lambda & 1 - 2\lambda & \end{pmatrix}$$

Pour $\lambda \leq 1/2$ (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 probabilités 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 chaîne de Markov est irréductible et réversible, et la mesure stationnaire associée est la mesure discrète uniforme, qui minimise l'entropie (voir section 1.4, page 25).

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 325, peut s'écrire

$$u^{n+1} = {}^t P u^n,$$

avec

$${}^t P = \begin{pmatrix} 1 - \lambda & 0 & 0 & \cdot & \cdot & \cdot & \lambda \\ \lambda & 1 - \lambda & 0 & 0 & \cdot & \cdot & \cdot \\ 0 & \lambda & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 1 - \lambda & \cdot & \cdot \\ 0 & \cdot & \cdot & 0 & \lambda & 1 - \lambda & \end{pmatrix}$$

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, considérons 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 discréétisation par différences finies, du type (explicite)

$$u^{n+1} = (\text{Id} + \Delta t A) u^n,$$

où A est une discréétisation 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 conserve⁸, alors⁹ la somme des éléments d'une colonne de A vaut 0 : le schéma se met sous la forme

$$u^{n+1} = {}^t P u^n,$$

7. nous écrivons ${}^t P$ bien que la matrice soit symétrique, car c'est bien ${}^t P$ qui interviendra dans les cas non symétriques.

8. Cette condition est vérifiée de fait par tous les schémas consistants usuels, même si la consistance n'implique pas, à strictement parler, la vérification exacte de cette propriété de conservation.

9. 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.

où P est une matrice stochastique.

Dans les cas considérés précédemment, la matrice $\text{Id} + \Delta t A = {}^t P$ 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 ${}^t P$ associées 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 de cardinal J . Un tel plan de transport peut être représenté par une matrice (γ_{ij}) (on se reportera au chapitre 14, page 286, pour plus de détails), qui précise quelle quantité provenant de i est transportée vers j .

Pour l'équation de transport, à partir d'une densité discrète u^0 , le schéma construit ainsi une nouvelle densité selon le plan de transport

$$\gamma_{j,j} = \left(1 - \frac{V\Delta t}{\Delta x}\right) u_j^0, \quad \gamma_{j,j+1} = \frac{V\Delta t}{\Delta x} u_j^0,$$

les autres coefficients étant nuls. La nouvelle densité u^1 est alors définie comme seconde marginale de γ :

$$u_j^1 = \sum_i \gamma_{i,j} u_i^0.$$

Remarque 16.18. (Liens avec le schéma Lagrangien projeté)

On notera que, sauf dans le cas d'un nombre CFL $(\Delta t V / \Delta x)$ exactement égal à un, il s'agit bien d'un appariement diffus, et pas d'une application, alors que le phénomène sous-jacent n'est pas de nature à disperser la matière, ni à la mélanger. On peut tout de même faire un lien entre ce plan et un véritable transport de matière. En effet, si l'on note \bar{u}^0 la fonction constante par morceaux associée à la densité courante, et \bar{u}^1 la nouvelle densité, on peut vérifier que, quand $V\Delta t / \Delta x \leq 1$,

$$\bar{u}^1 = P(\text{Id} + \Delta t V)_{\sharp} \bar{u}^0,$$

où $T_{\sharp} \bar{u}$ désigne la densité transportée par l'application T , et P la projection (L^2) sur l'espace des fonctions constantes par morceaux. Le phénomène de diffusion numérique déjà évoqué est alors associé à l'étape de projection. La relation ci-dessus correspond à un schéma numérique utilisé en pratique, qui se distingue en général du schéma aux différences finies. Il peut en particulier être utilisé sur des maillages très généraux (il n'est pas basé sur une discréttisation de l'EDP eulérienne, qui fait intervenir des opérateurs de différentiel, mais plutôt sur une expression Lagrangienne du transport). Par ailleurs, il n'est pas limité par une condition CFL stricte.

Pour l'équation de la chaleur, on a

$$\gamma_{j,j} = \left(1 - \frac{2D\Delta t}{(\Delta x)^2}\right) u_j^0, \quad \gamma_{j,j\pm 1} = \frac{2D\Delta t}{(\Delta x)^2} u_{j\pm 1}^0,$$

qui est bien un plan de transport sous réserve que la condition $\Delta t \leq (\Delta x)^2 / 2D$ soit vérifiée.

Remarque 16.19. On peut vérifier une certaine consistance du schéma vis-à-vis du mouvement brownien sous-jacent à l'équation de la chaleur elle-même. En effet, on peut estimer le second moment du déplacement, pour une quantité de matière initialement en x_j . On trouve

$$0 \times \left(1 - \frac{2D\Delta t}{(\Delta x)^2}\right) + 2 \times \frac{D\Delta t}{(\Delta x)^2} (\Delta x)^2 = 2D\Delta t,$$

qui correspond bien au déplacement quadratique moyen d'un particule brownienne X_t issue de x_j , et évoluant suivant $dX_t = \sigma dW_t$, avec $\sigma^2 = 2D$ (voir remarque 4.15, page 91).

Exercice 16.3. (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 16.4. On considère le schéma décentré amont appliquée 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

$$\partial_{tt}u - c^2\partial_{xx}u = 0.$$

Un schéma couramment utilisé est le schéma de Crank-Nicholson, i.e.

$$\frac{u_j^{n+1} - 2u_j^n + u_j^{n-1}}{(\Delta t)^2} + \theta c^2 \frac{-u_{j-1}^{n+1} + 2u_j^{n+1} - u_{j+1}^{n+1}}{(\Delta x)^2} + (1-\theta)c^2 \frac{-u_{j-1}^n + 2u_j^n - u_{j+1}^n}{(\Delta x)^2} = 0 \quad (16.11)$$

avec $\theta = 1/2$, qui peut s'écrire matriciellement

$$\left(\text{Id} + \frac{c^2(\Delta t)^2}{2(\Delta x)^2} A \right) u^{n+1} = 2u^n - u^{n-1} - \frac{c^2(\Delta t)^2}{2(\Delta x)^2} A u^n,$$

où A est la matrice du Laplacien discret.

16.7 Implémentation effective

Les schémas explicites ne nécessitent en général pas l'assemblage de la matrice. On pourra utiliser avantageusement les opérateurs de shift à droite S_R et shift à gauche S_L définis, dans un cadre périodique, par

$$S_R(u_1, u_2, \dots, u_J) = (u_J, u_1, \dots, u_{J-1}), \quad S_L(u_1, u_2, \dots, u_J) = (u_2, u_3, \dots, u_J, u_1).$$

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

$$u^{n+1} = u^n - \frac{V\Delta t}{\Delta x} (u^n - S_R u^n),$$

et le schéma centré :

$$u^{n+1} = u^n - \frac{V\Delta t}{2\Delta x} (S_L u^n - S_R u^n).$$

Diffusion

Le schéma explicite pour l'équation de la chaleur peut être implémenté (cas périodique) en utilisant les opérateurs de shift :

$$u^{n+1} = u^n + \frac{D\Delta t}{(\Delta x)^2} (S_R u^n - 2u^n + S_L u^n),$$

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 extrémiales (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 naturel¹⁰ 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
```

La manière la plus simple d'assembler les matrices résultant d'une discrétisation par différences finies est de passer par la commande `ssp.diags`, qui prend en argument 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.

$$A = \begin{pmatrix} 1 & \beta & 0 & \cdot & \cdot & \cdot & -\beta \\ -\beta & 1 & \beta & 0 & \cdot & \cdot & \cdot \\ 0 & -\beta & \cdot & \cdot & & & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & & \cdot \\ \cdot & & -\beta & 1 & \beta & & \\ \beta & \cdot & \cdot & 0 & -\beta & 1 & \end{pmatrix}$$

avec $\beta = \Delta t V / (2\Delta x)$, de la façon suivante

```
beta = 0.5*V*dt/dx
```

10. Cet assemblage n'est pas nécessaire à strictement parler. On peut être amené à utiliser, pour résoudre le système linéaire, des méthodes dites itératives, 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 pré-assembler la matrice. Cette approche permet d'économiser de l'espace mémoire dans le cas où la matrice contient très peu d'éléments différents, ce qui est le cas des matrices résultant de la discrétisation d'opérateurs différentiels invariants par translation, sur un maillage régulier.

```

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`¹¹ 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

$$A_1 = \begin{pmatrix} 2 & -1 & 0 & \cdot & \cdot & \cdot & 0 \\ -1 & 2 & -1 & 0 & \cdot & \cdot & \cdot \\ 0 & -1 & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & & \cdot & 2 & -1 & & \\ 0 & \cdot & \cdot & 0 & -1 & 2 & \end{pmatrix}$$

En dimension 2 d'espace, le Laplacien discret agit sur les valeurs au point $(i\Delta x, j\Delta x)$ de la discrémination 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 16.5. Généraliser la construction décrite ci-dessus au cas de la dimension 3.

Exercice 16.6. 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 problème d'évolution par un schéma implicite conduit à la résolution de multiples systèmes linéaires impliquant la même matrice, pour des seconds membres différents. 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)`.

11. Voir <http://perso.univ-perp.fr/langlois/images/pdf/mp/scipy.pdf>

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)
```

16.8 Exercices

Exercice 16.7. On s'intéresse à la résolution numérique de l'équation de transport

$$\partial_t u + V \partial_x u = 0,$$

en géométrie périodique, où $V > 0$ est une vitesse constante fixée. En utilisant les notations usuelles, on introduit le schéma dit de Lax-Friedrich :

$$\frac{2u_j^{n+1} - u_{j-1}^n - u_{j+1}^n}{2\Delta t} + V \frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} = 0$$

- 1) Écrire ce schéma sous forme matricielle, i.e. préciser la matrice A tel que le schéma s'écrive $u^{n+1} = Au^n$, où u^n est le vecteur $(u_j^n)_j$ (avec périodicité des indices).
- 2) Montrer que ce schéma est d'ordre 1 en temps et en espace, sous réserve de supposer qu'il existe une constante $C > 0$ telle que $\Delta x \leq C\Delta t$.
- 3) Montrer que le schéma est stable en norme L^∞ , sous une condition reliant Δt et Δx que l'on précisera.
- 4) Montrer que le schéma est stable en norme L^2 , sous la même condition que dans la question précédente.
- 5) Décrire le comportement de ce schéma lorsque l'on a exactement $\Delta t = \Delta x/V$. Si maintenant $\Delta t \in]0, \Delta x/V[$, montrer que u^n converge vers un vecteur colinéaire à $(1, 1, \dots, 1)$ lorsque n tend vers l'infini. Expliquer pourquoi ce comportement est très éloigné du comportement de la solution exacte de l'équation, mais qu'il n'est pourtant pas contradictoire avec le théorème de convergence du schéma numérique.
- 6) Comment se positionne ce schéma de Lax-Friedrich (avantage(s), inconvénient(s) ?) par rapport au schéma explicite centré usuel (pour lequel la dérivée par rapport au temps est approchée par $(u_j^{n+1} - u_j^n)/\Delta t$).

Exercice 16.8. On considère l'équation de convection-diffusion

$$\frac{\partial u}{\partial t} + V \frac{\partial u}{\partial x} - D \frac{\partial^2 u}{\partial x^2} = 0 \text{ pour } x \in \mathbb{R}, t > 0,$$

avec $u(x, 0) = u^0$, u et u^0 périodiques de période 1.

- 1) Schéma décentré amont. On considère le schéma
$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + V \frac{u_j^n - u_{j-1}^n}{\Delta x} - D \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{(\Delta x)^2} = 0.$$
 - a) Montrer que le schéma est au moins d'ordre un en temps et en espace.
 - b) Donner des conditions suffisantes de stabilité L^∞ du schéma lorsque $V > 0$ et $V < 0$.
 - c) Montrer que le schéma est convergent sous la condition CFL introduite précédemment.
- 2) Schéma centré.

Pour la même équation, on considère le schéma

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + V \frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} - \frac{D}{2} \left(\frac{u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}}{(\Delta x)^2} + \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{(\Delta x)^2} \right) = 0.$$

a) Étudier la stabilité L^2 du schéma. Que se passe-t-il lorsque $D \rightarrow 0$?

b) Montrer que le schéma est consistant, déterminer son ordre.

c) Établir un résultat de convergence.

Chapitre 17

Méthode des éléments finis

Sommaire

17.1	Formulation variationnelle du problème de Poisson	344
17.2	Méthode des éléments finis	349
17.3	Estimation d'erreur pour la méthode des Éléments Finis	351
17.4	Éléments finis et réseaux résistifs	356

17.1 Formulation variationnelle du problème de Poisson

On considère le problème de Poisson dans un domaine Ω de \mathbb{R}^d , supposé borné et régulier.

$$\begin{cases} -\Delta u &= f \quad \text{dans } \Omega \\ u &= 0 \quad \text{sur } \Gamma \end{cases} \quad (17.1)$$

L'approche présentée dans la suite consiste à exprimer ce problème de façon duale : on écrit le produit de dualité L^2 de chacun de ses membres contre une fonction-test générique. Un choix approprié de l'espace dans lequel on fait vivre la fonction inconnue et la fonction-test permet de transformer ce problème en un énoncé de type Riez-Fréchet : l'inconnue joue alors le rôle de l'élément d'un espace de Hilbert qui s'identifie à une forme linéaire donnée (qui résulte du terme de forçage, i.e. du second membre) au travers d'un produit scalaire particulier. Le résultat d'existence est d'unicité prend ainsi la forme d'une *identité* entre l'inconnue u et la donnée f , qui expriment le même objet de façons différentes.

Formulation variationnelle

On obtient¹ 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

$$\int_{\Omega} \nabla u \cdot \nabla v - \int_{\Gamma} v \frac{\partial u}{\partial n} = \int f v$$

1. 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.

d'où (les termes de bord s'annulent sur Γ du fait de la nullité de v)

$$\int_{\Omega} \nabla u \cdot \nabla v = \int f v.$$

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 chapitre 12) des fonctions de L^2 dont les dérivées partielles sont aussi dans L^2 , et qui sont nulles² sur Γ :

Dans le cas où la forme bilinéaire $a(\cdot, \cdot)$ est coercive, c'est à dire (voir définition 18.20) 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 18.25) assure l'existence et l'unicité d'une solution dans V .

Cette solution peut être caractérisée comme unique minimiseur de la fonctionnelle

$$J(v) = \frac{1}{2} a(v, v) - \langle \varphi, v \rangle = \frac{1}{2} \int_{\Omega} |\nabla v|^2 - \int_{\Omega} f v.$$

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 12.33, page 12.33).

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 régulière³ 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 12.10, et la section 12.5 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 domaine sont contenues dans l'appartenance de u à l'espace V

2. Le sens que l'on peut donner à l'expression $u|_{\Gamma} = 0$ est précisé dans la section 19.7.2, page 384.

3. Précisons que ce résultat de régularité interviendra de façon essentielle dans l'analyse d'erreur de la méthode de discréétisation.

Conditions de Neuman Les conditions de Neuman portent sur la dérivée normale de la solution sur la frontière, que l'on fixe à 0 pour le cas de conditions homogènes (ce qui correspond à un flux nul). Ces conditions posent des difficultés particulières, parmi lesquelles

1. Le problème de Poisson avec conditions de Neuman ne fait intervenir que des dérivées de la fonction inconnue, on ne peut donc espérer avoir au mieux qu'une solution définie à une constante additive près. On verra qu'effectivement ce problème est mal posé en général, y compris en termes d'existence. On contournera cette difficulté dans un premier temps en rajoutant un terme de masse au Laplacien⁴.
2. La condition de Neuman implique la trace de la dérivée normale de la fonction inconnue. Le cadre mathématique naturel est le théorème de Lax-Milgram appliqué dans l'espace de Hilbert H^1 , or la trace de la dérivée normale d'une fonction de H^1 n'est pas définie. De fait, la formulation variationnelle sur laquelle repose le théorème d'existence et d'unicité ne fait pas apparaître explicitement cette dérivée normale. On parle de condition *naturelle*, qui disparaît en tant que telle de la formulation⁵.

Nous considérons en premier lieu le problème suivant

$$\begin{cases} u - \Delta u = f & \text{in } \Omega \\ \frac{\partial u}{\partial n} = 0 & \text{sur } \Gamma \end{cases} \quad (17.2)$$

On obtient la formulation variationnelle en multipliant par une fonction-test v . Le terme de bord disparaît du fait de la condition homogène.

$$\int_{\Omega} uv + \int_{\Omega} \nabla u \cdot \nabla v = \int_{\Omega} fv.$$

Ce problème se ramène donc à la recherche de $u \in V$ tel que

$$a(u, v) = \langle \varphi, v \rangle \quad \forall v \in V.$$

On vérifie immédiatement la continuité et la coercivité de $a(\cdot, \cdot)$ (qui est en fait le produit scalaire canonique sur H^1). Le problème admet donc une unique solution $u \in V$.

Retour à l'équation de départ

Si l'on souhaite donner un statut précis à l'équation de départ, avec des identités entre fonctions⁶, il est nécessaire de montrer que la solution est dans $H^2(\Omega)$. Cette propriété peut être délicate à établir rigoureusement, en particulier dans le cas de domaines peu réguliers. Nous supposerons ici le domaine régulier, et nous admettrons la régularité H^2 de la solution.

La démarche consiste dans un premier temps à considérer des fonctions-test régulières à support compact. On utilise alors la formule de Green, ce qui autorise la régularité H^2 de la solution u , pour obtenir

$$\int_{\Omega} (u - \Delta u - f) v = 0$$

d'où l'on déduit par densité dans L^2 des fonctions régulières que $-\Delta u = 0$ presque partout. On considère dans un second temps des fonctions régulières non nécessairement nulles au bord, pour obtenir

$$\int_{\Omega} (u - \Delta u) v + \int_{\Gamma_N} \frac{\partial u}{\partial n} v = \int_{\Omega} fv$$

4. En terme de modélisation, cela correspondrait à prendre en compte un terme de disparition ou transformation pour l'espèce concernée). Une approche permettant de donner un cadre rigoureux au problème sans le terme de masse, en prescrivant la valeur moyenne de la fonction sur le domaine, est proposée plus loin.

5. On trouve parfois dans la littérature non mathématique le terme de *Do nothing approach*. Il s'agit en effet de la condition implémentée lorsque l'on considère la formulation variationnelle discrète sans termes de bord, en laissant libre les degrés de liberté sur la frontière.

6. Il existe une autre manière (que nous ne privilierons pas ici) de donner un sens à l'équation de Poisson sans l'aide d'aucun théorème de régularité en passant par la notion de divergence faible 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 $\nabla u \in H_{div}$. Cette trace est alors définie dans un sens faible, ce qui interdit par exemple l'écriture $\partial_n u = g$ p.p.

Comme l'équation de Poisson est vérifiée presque partout, il reste

$$\int_{\Gamma} \left(\frac{\partial u}{\partial n} \right) v = 0.$$

La fonction v pouvant être choisie arbitrairement, on en déduit $\partial_n u = 0$ presque partout sur Γ (la dérivée normale de u est dans $L^2(\Gamma)$).

Considérons maintenant le problème

$$\begin{cases} -\Delta u &= f & \text{in } \Omega \\ \frac{\partial u}{\partial n} &= 0 & \text{sur } \Gamma \end{cases} \quad (17.3)$$

La solution éventuelle à ce problème est manifestement définie au mieux à une constante additive près. Par ailleurs, si l'on suppose que le problème admet une solution régulière, l'intégration de l'équation sur le domaine donne, après intégration par parties,

$$0 = \int_{\Omega} f.$$

Il ne saurait donc y avoir de solution si f n'est pas à moyenne nulle. Nous supposerons donc f est à moyenne nulle. La formulation variationnelle s'écrit

$$\int_{\Omega} \nabla u \cdot \nabla v = \int_{\Omega} f v.$$

La forme bilinéaire $a(\cdot, \cdot)$ du membre de gauche est bien définie et continue sur $H^1 \times H^1$, mais manifestement pas coercive. On introduit alors l'espace

$$K = \left\{ v \in H^1(\Omega), \int_{\Omega} v = 0 \right\}.$$

Il s'agit d'un espace de Hilbert comme sous-espace fermé de l'espace de Hilbert H^1 , et la forme bilinéaire $a(\cdot, \cdot)$ est coercive sur cet espace. On donc existence et unicité dans K d'une solution u à la formulation variationnelle

$$\int_{\Omega} \nabla u \cdot \nabla v = \int_{\Omega} f v \quad \forall v \in K.$$

On remarquera que l'hypothèse $\int f = 0$ n'a pas été utilisée pour l'instant, et de fait elle n'est pas nécessaire pour démontrer le caractère bien posé du problème. On notera par ailleurs que f n'intervient plus qu'à une constante additive près, puisque les fonctions-test sont à moyenne nulle.

De fait, on va voir qu'il est impossible de revenir à l'équation de départ si f n'est pas à moyenne nulle. En effet, pour revenir à l'équation, on doit pouvoir disposer de fonctions tests qui engendrent un sous-espace dense dans L^2 . Soit une fonction test régulière φ . On note $\bar{\varphi}$ sa moyenne, de telle sorte que $\varphi - \bar{\varphi}$ est à moyenne nulle, on peut donc l'utiliser dans la formulation variationnelle, et, du fait que $\int f = 0$, on obtient

$$\int_{\Omega} \nabla u \cdot \nabla \varphi = \int_{\Omega} f \varphi \quad \forall \varphi \in C_c^{\infty}(\Omega),$$

et on peut donc procéder comme précédemment, en admettant la régularité H^2 de la solution, pour retrouver l'équation et la condition aux limites.

Conditions de Robin On s'intéresse maintenant au problème

$$\begin{cases} -\Delta u &= f & \text{in } \Omega \\ \beta u + \frac{\partial u}{\partial n} &= 0 & \text{sur } \Gamma \end{cases} \quad (17.4)$$

avec $\beta > 0$. La formulation variationnelle s'écrit

$$\underbrace{\int_{\Omega} \nabla u \cdot \nabla v + \beta \int_{\Gamma} uv}_{a(u,v)} = \int_{\Omega} f v.$$

L'inégalité de Poincaré généralisée permet d'établir la coercivité de la forme bilinéaire $a(\cdot, \cdot)$ sur $H^1 \times H^1$, d'où l'existence et l'unicité d'une solution. Pour reconstruire l'équation et la condition aux limites (en admettant la régularité H^2 de la solution), on procède comme précédemment.

Exercice 17.1. On considère le problème de Poisson avec conditions de Robin, et l'on note u_β la solution associée au paramètre β . Étudier la limite de u_β quand β tend vers l'infini, et quand β tend vers 0.

Extension à des conditions aux limites plus générales

Obstacle de conductivité infinie

On considère un domaine Ω du plan, et ω un sous-domaine fortement inclus dans Ω , c'est-à-dire que $\bar{\omega} \subset \Omega$. 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 $\bar{\omega} \subset \Omega$, 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 $L^2(\Omega \setminus \bar{\omega})$, et l'on s'intéresse au problème suivant :

$$\left\{ \begin{array}{lcl} -\Delta u & = & f \quad \text{dans } \Omega \setminus \bar{\omega} \\ u & = & 0 \quad \text{sur } \partial\Omega \\ u & = & U \quad \text{sur } \partial\omega \\ \int_{\partial\omega} \frac{\partial u}{\partial n} & = & 0, \end{array} \right. \quad (17.5)$$

où U est une constante réelle dont la valeur est inconnue.

On introduit l'espace

$$H_C^1(\Omega \setminus \bar{\omega}) = \{u \in H^1(\Omega \setminus \bar{\omega}), u = 0 \text{ sur } \partial\Omega, u = \text{cste sur } \partial\omega\}.$$

L'approche variationnelle directe est basée sur la fonctionnelle

$$\begin{aligned} H_C^1(\Omega \setminus \bar{\omega}) &\longrightarrow \mathbb{R} \\ v &\mapsto J(v) = \frac{1}{2} \int_{\Omega \setminus \bar{\omega}} |\nabla v|^2 - \int_{\Omega \setminus \bar{\omega}} fv, \end{aligned}$$

Le problème 17.6 consiste donc à minimiser J sur $H_C^1(\Omega \setminus \bar{\omega})$. 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 17.1. Soit $u \in H_C^1(\Omega \setminus \bar{\omega})$ la fonction qui minimise la fonctionnelle J sur $H_C^1(\Omega \setminus \bar{\omega})$. Alors u est solution du problème (17.5).

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 $H_0^1(\Omega \setminus \bar{\omega})$, et c'est la solution faible de l'équation

$$-\Delta w = f + \Delta \tilde{U},$$

avec conditions de Dirichlet homogènes. C'est donc un élément de $H^2(\Omega \setminus \bar{\omega})$, et par suite u lui-même a une régularité H^2 . On considère maintenant des fonctions-test dans $H_0^1(\Omega \setminus \bar{\omega})$. Par intégration par parties, on obtient $-\Delta u = f$ dans $\Omega \setminus \bar{\omega}$. 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

$$-\int_{\Omega \setminus \bar{\omega}} v \Delta u + \int_{\partial\omega} \frac{\partial u}{\partial n} v = \int f v,$$

d'où

$$\int_{\partial\omega} \frac{\partial u}{\partial n} = 0,$$

ce qui termine la preuve. \square

17.2 Méthode des éléments finis

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 354, 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 :

$$V_h = \{v_h \in V, v_h|_K \text{ est affine sur tout } K \in T_h\}.$$

Le problème discret s'écrit

$$\left\{ \begin{array}{l} \text{Trouver } u_h \in V_h \text{ tel que} \\ \int_{\Omega} \nabla u_h \cdot \nabla v_h = \int_{\Omega} f v_h \quad \forall v_h \in V_h. \end{array} \right. \quad (17.6)$$

Formulation matricielle

On numérote $i = 1, 2, \dots, N_h$ 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 (17.6) se ramène au système matriciel (on garde la notation u_h pour désigner le vecteur (u^1, \dots, u^{N_h}))

$$A u_h = b_h,$$

où A est une matrice carrée d'ordre N_h , et $b_h \in \mathbb{R}^{N_h}$:

$$A = (a_{ij}) = \left(\int_{\Omega} \nabla w_i \cdot \nabla w_j \right), \quad b_h = \left(\int_{\Omega} f w_i \right)_i.$$

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, qui sera détaillée dans la section 17.3, 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 discrétilisation 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 17.3), 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 17.3.

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 12.45, page 257 (ou plus précisément de la proposition 12.47 qui s'applique au cas d'un polyèdre convexe), du théorème d'approximation 17.18, et du lemme de Céa 17.3.

Remarque 17.2. On prendra garde au fait que le lemme de Céa, contrairement à l'estimation de l'erreur d'approximation détaillée ci-après, 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.

17.3 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 éléments 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 . \square

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

$$\hat{K} = \{(x_1, \dots, x_N) \in \mathbb{R}_+^N, x_1 + \dots + x_N \leq 1\}.$$

On se placera dans ce qui suit en dimension 2 d'espace, où \hat{K} est le triangle de référence

$$\hat{K} = \{(x_1, x_2) \in \mathbb{R}_+^2, x_1 + x_2 \leq 1\}.$$

Notation 17.5. Pour toute fonction w définie sur K (ou sur tout autre domaine), on notera (lorsque ces quantités sont définies)

$$|w|_{0,K} = \|w\|_{L^2(K)}, |w|_{1,K} = \|\nabla w\|_{L^2(K)^2}, |w|_{2,K} = \|D^2 w\|_{L^2(K)^{N^2}} = \left(\sum_{i,j} |\partial_{ij} w|^2 \right)^{1/2}.$$

Notation 17.6. On note $P^k(K)$ l'espace des fonctions polynomiales 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 . \square

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.

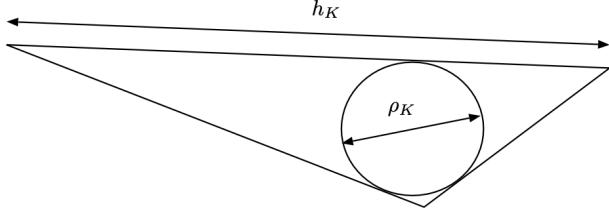


FIGURE 17.1 – Définition de h et ρ pour un triangle

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 \hat{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) :

$$\hat{x} \mapsto x = \Phi(\hat{x}) = B\hat{x} + b$$

On a

$$\|\nabla\Phi\| = \|{}^t\nabla\Phi\| = \|B\| \leq \frac{1}{\hat{\rho}}h_K, \quad \|\nabla\Phi^{-1}\| = \|{}^t\nabla\Phi^{-1}\| = \|B^{-1}\| \leq \frac{1}{\rho_K}\hat{h}.$$

Démonstration. Soit $\tilde{\xi} \in \mathbb{R}^N$ de norme $\tilde{\rho}$. Il existe \tilde{x}_1 et \tilde{x}_2 dans \hat{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 . \square

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 partielles 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_K et de diamètre intérieur ρ_K , et \hat{u} définie sur \hat{K} par

$$\hat{u}(\hat{x}) = u \circ \Phi(\hat{x}).$$

Soit $\alpha = (\alpha_1, \alpha_2)$ un multi-indice, avec $|\alpha| = \alpha_1 + \alpha_2 = s \in \mathbb{N}$. On a

$$\left| \frac{\partial^s \hat{u}}{\partial \hat{x}^\alpha} \right| \leq Ch_K^s \sum_{|\alpha'|=s} \left| \frac{\partial^s u}{\partial x^{\alpha'}} \right|, \quad \left| \frac{\partial^s u}{\partial x^\alpha} \right| \leq C \frac{1}{\rho_K^s} \sum_{|\alpha'|=s} \left| \frac{\partial^\alpha \hat{u}}{\partial \hat{x}^{\alpha'}} \right|.$$

Démonstration. Soit u une fonction régulière définie sur K . On a

$$\frac{\partial \hat{u}}{\partial \hat{x}_i} = \nabla u \cdot \frac{\partial \Phi}{\partial \hat{x}_i} = ((\nabla \Phi)^T \nabla u) \cdot \hat{e}_i,$$

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 élevé, ainsi que les estimations inverses (à partir de $u(x) = \hat{u} \circ \Phi^{-1}$), se démontrent de la même manière. \square

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

$$\begin{aligned} |I_K u - u|_{1,K} &\leq C \frac{h^2}{\rho} |u|_{2,K} \quad \forall u \in H^2(K) \\ |I_K u - u|_{0,K} &\leq Ch^2 |u|_{2,K} \quad \forall u \in H^2(K) \\ |I_K^0 u - u|_{0,K} &\leq Ch |u|_{1,K} \quad \forall u \in H^1(K) \end{aligned}$$

Démonstration. Ces estimations se démontrent à partir de l'estimation de stabilité (proposition 17.7) appliquée au simplexe de référence. On transporte $|I_K u - u|_{1,K}^2$ sur le triangle de référence, ce qui fait apparaître $\widehat{|I_K u - u|_{1,K}^2} = |I_{\hat{K}} \hat{u} - \hat{u}|_{1,\hat{K}}^2$ 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 $|\hat{u}|_{2,\hat{K}}^2$. 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'inégalité obtenue donne la première inégalité, les autres se démontrent de la même manière. \square

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 noeuds 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 néanmoins 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 $h = \sup_{K \in T_h} h_K \rightarrow 0$.

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

$$|u - I_h u|_{1,\Omega} \leq C\sigma h |u|_{2,\Omega}, \quad |u - I_h u|_{0,\Omega} \leq Ch^2 |u|_{2,\Omega}$$

Démonstration. On a

$$\int_{\Omega} |u - I_h u|^2 = \sum_{K \in T_h} \int_K |u - I_h u|^2 \leq C^2 h^4 \sum_{K \in T_h} |u|_{2,K}^2 \leq C^2 h^2 |u|_{2,\Omega}^2.$$

On raisonne de la même manière pour estimer $|u - I_h u|_{0,\Omega}$. \square

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 $H_0^1(\Omega)$ dont la restriction à chaque triangle de T_h est affine. Pour $f \in L^2(\Omega)$, on note $u \in H_0^1(\Omega)$ la solution faible de

$$-\Delta u = f,$$

et u_h la solution du problème discréteisé

$$\int_{\Omega} \nabla u_h \cdot \nabla v_h = \int_{\Omega} f v_h \quad \forall v_h \in V_h.$$

Il existe une constante $C > 0$ telle que

$$|u - u_h|_1 \leq Ch |f|_0.$$

Démonstration. D'après la proposition 12.47, page 258, la solution exacte est de régularité H^2 , avec $|u|_2 \leq C |f|_0$. On a donc

$$|u - I_h u|_1 \leq Ch |u|_2 \leq C' h |f|_0.$$

Le lemme de Céa 17.3 permet de conclure. \square

Proposition 17.20. (Lemme de Aubin-Nitsche)

Sous les hypothèses de la proposition précédente, il existe une constante $C > 0$ telle que

$$|u - u_h|_0 \leq Ch^2 |f|_0.$$

Démonstration. On considère le problème aux limites suivant

$$-\Delta w = u - u_h.$$

On prend $u - u_h$ comme fonction-test dans la formulation variationnelle de ce problème. Il vient

$$\int_{\Omega} |u - u_h|^2 = \int_{\Omega} \nabla w \cdot \nabla(u - u_h) = \int_{\Omega} \nabla(w - I_h w) \cdot \nabla(u - u_h)$$

car $\int \nabla(u - u_h) \cdot \nabla v_h = 0$ pour tout $v_h \in V_h$. On a donc

$$|u - u_h|_0^2 \leq |w - I_h w|_1 |u - u_h|_1.$$

Le premier facteur du produit se majore de la façon suivante

$$|w - I_h w|_1 \leq Ch |w|_2 \leq Ch |u - u_h|_0,$$

et le second par $C_4 h |f|_0$, d'où l'estimation en $\mathcal{O}(h^2)$ sur la norme L^2 de l'erreur. \square

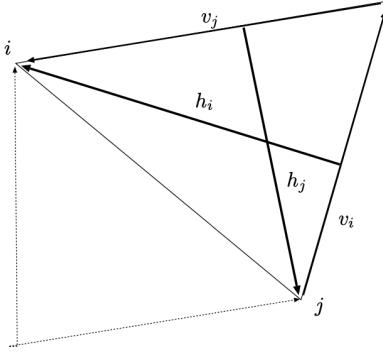


FIGURE 17.2 – Assemblage de la matrice élémentaire

17.4 Éléments finis et réseaux résistifs

Soit T_h une triangulation d'un domaine Ω , et A la matrice résultant de la discréétisation par éléments finis P^1 de la forme bilinéaire

$$a(u, v) = \int_{\Omega} \nabla u \cdot \nabla v.$$

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 discréétisé 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 chapitre ??), 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. Ils 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

$$p(i) = \frac{1}{C(i)} \sum_{j \sim i} c_{ij} p(j),$$

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 17.3).

Equation de conservation continue associée à la solution discrète

On peut associer à la solution discrète d'un problème de Laplace discrétilisé 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. Le 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 le chapitre ??)

$$\int_{\Omega} \nabla p \cdot \nabla q = 0 \quad \forall q \in V_h^0,$$

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

$$\mu(x) = \sum_{y \sim x} du(y) \delta_y, \quad du(y) = \sum_{x \sim y} u(y, x) = \sum_{y \sim x} c(y, x)(p(y) - p(x)).$$

On note G la mesure vectorielle associée aux flux discrets sur le maillage, selon la démarche décrite dans la section 2.8. On a alors, au sens des distributions, (voir proposition 2.33, page 58)

$$\nabla \cdot G = \mu.$$

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 positives⁷.

⁷. Un tel réseau serait irréalisable en pratique, qu'il s'agisse d'un circuit électrique, ou d'un réseaux de tuyaux au travers duquel s'écoule un fluide visqueux.

Chapitre 18

Espaces de Hilbert

Sommaire

18.1	Définitions, principales propriétés	358
18.2	Convergence faible	364
18.3	Somme Hilbertienne, bases Hilbertiennes	366
18.4	Décomposition spectrale des opérateur auto-adjoints compacts	367
18.5	Problèmes d'évolution	370
18.6	Minimisation de fonctionnelles convexes	371

18.1 Définitions, principales propriétés

Definition 18.1. (Produit scalaire)

Soit H un espace vectoriel sur \mathbb{R} . On appelle produit scalaire une forme bilinéaire $\langle u | v \rangle$ de $H \times H$ dans \mathbb{R} , symétrique, définie et positive :

$$\langle u | v \rangle = \langle v | u \rangle, \quad \langle u | u \rangle \geq 0 \quad \forall u \in H, \quad \text{et} \quad \langle u | u \rangle = 0 \iff u = 0.$$

Un produit scalaire définit sur H une structure d'espace vectoriel normé pour la norme

$$u \mapsto |u| = \langle u | u \rangle^{1/2}.$$

Definition 18.2. (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 18.1.1. 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,

$$\ell_\gamma^2 = \left\{ (u_n) \in \mathbb{R}^\mathbb{N}, \sum \gamma_n |u_n|^2 < +\infty \right\}.$$

Proposition 18.3. (Inégalité de Cauchy-Schwarz)

Tout produit scalaire vérifie l'inégalité de Cauchy-Schwarz

$$|\langle u | v \rangle| \leq |u| |v| \quad \forall u, v \in H.$$

Démonstration. On écrit

$$\langle u + tv | u + tv \rangle \geq 0 \quad \forall t \in \mathbb{R}.$$

Le minimum de cette quantité est atteint en $t = -\langle u | v \rangle / |v|^2$, on a donc en particulier

$$|u|^2 - 2 \frac{\langle u | v \rangle^2}{|v|^2} + \frac{\langle u | v \rangle^2}{|v|^2} \geq 0,$$

d'où $|\langle u | v \rangle| \leq |u| |v|$. □

Proposition 18.4. (Identité du parallélogramme)

Toute norme issue d'un produit scalaire vérifie l'identité du parallélogramme

$$\left| \frac{u+v}{2} \right|^2 + \left| \frac{u-v}{2} \right|^2 = \frac{1}{2}(|u|^2 + |v|^2).$$

Démonstration. Il suffit de développer le membre de gauche. □

Proposition 18.5. 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. □

Definition 18.6. (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 18.7. (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

$$|z - u| = \min_{v \in K} |z - v| = \text{dist}(z, K).$$

La projection u est caractérisée par la propriété

$$\begin{cases} u \in K \\ \langle z - u | v - u \rangle \leq 0 \quad \forall v \in K. \end{cases} \quad (18.1)$$

On notera $u = P_K z$.

Démonstration: On considère une suite minimisante (u_n)

$$u_n \in K, \quad |z - u_n| \longrightarrow d = \text{dist}(z, K).$$

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, $|u - z| = \text{dist}(z, K)$.

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 18.1. 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 18.2. 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 18.3. 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 18.8. Si K est un sous-espace affine fermé de H , alors la caractérisation (18.1) prend la forme

$$\begin{cases} u \in K \\ \langle z - u \mid v - u \rangle = 0 \quad \forall v \in K, \end{cases} \quad (18.2)$$

et si K est un sous-espace vectoriel de H , on a

$$\begin{cases} u \in K \\ \langle z - u \mid v \rangle = 0 \quad \forall v \in K. \end{cases} \quad (18.3)$$

Remarque 18.9. 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 18.10. 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 (18.1) :

$$\begin{aligned} \langle f - P_K f \mid P_K g - P_K f \rangle &\leq 0, \\ \langle g - P_K g \mid P_K f - P_K g \rangle &\leq 0. \end{aligned}$$

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 18.11. 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 18.4. 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 18.12. 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 . \square

Proposition 18.13. (Caractérisation de la densité)

Soit H un espace de Hilbert et K un sous-espace de H tel que l'implication suivante soit vérifiée :

$$\langle h | w \rangle = 0 \quad \forall w \in K \implies h = 0.$$

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 $\langle h | w \rangle = 0$ pour tout $w \in K$, et $h \neq 0$ car $u \notin \overline{K}$. \square

Théorème 18.14. (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 $h \in H$ et $\alpha \in \mathbb{R}$ tels que

$$\langle h | x \rangle \leq \alpha < \langle h | z \rangle \quad \forall x \in K.$$

Démonstration: On introduit la projection $u = P_K z$ de z sur K , et l'on prend $h = z - u$ et $\alpha = \langle h | u \rangle$. Pour tout $x \in K$, on a

$$\langle h | x \rangle - \alpha = \langle h | x \rangle - \langle h | u \rangle = \langle z - u | x - u \rangle \leq 0.$$

et on a par ailleurs $\langle h | z \rangle - \alpha = \langle h | z \rangle - \langle h | u \rangle = |z - u|^2 > 0$. \square

Exercice 18.5. Soient u, u_1, \dots, u_n , des éléments d'un espace de Hilbert H . Montrer l'équivalence suivante

$$\left(\bigcap u_i^\perp \right) \subset u^\perp \iff \exists \lambda_1, \dots, \lambda_n, u = \sum \lambda_i u_i.$$

Definition 18.15. (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 18.16. 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 18.17. (Riesz-Fréchet)

Soit $\varphi \in H'$ (dual topologique de H). Il existe $f \in H$ unique tel que

$$\langle \varphi, u \rangle = \langle f | u \rangle \quad \forall u \in H. \tag{18.4}$$

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 cela on considère $z \notin K$. D'après la caractérisation (18.3), 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),

$$\langle \varphi, v \rangle = \langle \varphi, h \rangle \langle v | h \rangle$$

d'où l'identité (18.4) avec $f = \langle \varphi, h \rangle h$. L'unicité d'un tel f est immédiate. \square

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 18.13 à un sous-espace du dual :

Proposition 18.18. (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 :

$$\langle \varphi, h \rangle = 0 \quad \forall \varphi \in K \implies h = 0.$$

Alors K est dense dans H' .

Proposition 18.19. (Continuité d'une forme bilinéaire)

Soit $a : H \times H \rightarrow \mathbb{R}$ 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

$$\left| a\left(r \frac{u}{|u|}, r \frac{v}{|v|}\right) \right| \leq 1 \implies |a(u, v)| \leq \frac{1}{r^2} |u| |v|.$$

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 $(u, v) \in H \times H$. \square

Definition 18.20. (Coercivité d'une forme bilinéaire)

Soit $a : H \times H \rightarrow \mathbb{R}$ une forme bilinéaire. On dit que a est coercive s'il existe $\alpha > 0$ tel que

$$a(u, u) \geq \alpha |u|^2 \quad \forall u \in H.$$

Remarque 18.21. 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 18.6. Soit $\alpha = (\alpha_n)$ une suite bornée de réels, et

$$a : (u, v) \in \ell^2 \times \ell^2 \mapsto \sum_{n=0}^{+\infty} \alpha_n u_n v_n.$$

A quelle condition sur α la forme bilinéaire $a(\cdot, \cdot)$ est-elle coercive ?

Remarque 18.22. On verra qu'il existe une définition plus générale de la coercivité (pour des fonctionnelles quelconques, voir théorème 18.53), équivalente à la définition ci-dessus dans le cas particulier des formes bilinéaires.

Proposition 18.23. 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 $u \mapsto Au$ 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

$$|Au| = \sup_{|v|=1} (Au, v) = \sup_{|v|=1} a(u, v) \leq C|u|,$$

où $\|a\|$ est la constante de continuité de a .

Si a est coercive, on a $(Au, u) = a(u, u) \geq \alpha|u|^2$, et donc $|Au| \geq \alpha|u|$ 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 $g \in H$,

$$(g, Au) = 0 \quad \forall u \in H \implies (g, Ag) = a(g, g) = 0$$

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. \square

Remarque 18.24. 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 18.25. (Lax-Milgram)

Soit H un espace de Hilbert, et a une forme bilinéaire continue et coercive sur $H \times H$. Pour tout $\tilde{f} \in H'$, il existe un $u \in H$ unique tel que

$$a(u, v) = \langle \tilde{f}, v \rangle \quad \forall v \in H. \tag{18.5}$$

Si a est symétrique, u est l'unique élément de H qui réalise le minimum de la fonctionnelle

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

Démonstration. D'après le théorème de représentation de Riesz-Fréchet, il existe un unique $f \in H$ tel que

$$\langle f | v \rangle = \langle \tilde{f}, v \rangle \quad \forall v \in H.$$

On introduit l'opérateur A associé à $a(\cdot, \cdot)$, qui est bijectif (voir proposition 18.23). 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 (18.6). Pour tout $h \in H$, l'application

$$t \mapsto \psi(t) = J(u + th) - J(u)$$

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 (18.6). \square

Corollaire 18.26. 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 $\tilde{f} \in H'$, il existe un $u \in K$ unique tel que

$$a(u, v) = \langle \tilde{f}, v \rangle \quad \forall v \in K^0. \quad (18.6)$$

Si a est symétrique, u est l'unique élément de K qui réalise le minimum de la fonctionnelle

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

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

$$a(\tilde{u}, v) = \langle \tilde{f}, v \rangle - a(U, v) \quad \forall v \in K^0,$$

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 18.27. (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 $\varepsilon : H \rightarrow H$ est telle que $\varepsilon(h) \rightarrow 0$ quand $h \rightarrow 0$. 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 $u \mapsto J'(u)$ est continue.

18.2 Convergence faible

Comme précédemment H désigne un espace de Hilbert réel muni du produit scalaire $(., .)$ et de la norme $|.|$.

Definition 18.28. (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

$$\langle u_n | v \rangle \rightarrow \langle u | v \rangle \quad \forall v \in H,$$

ou de façon équivalente, si

$$\langle \varphi, u_n \rangle \rightarrow \langle \varphi, u \rangle \quad \forall \varphi \in H'.$$

Proposition 18.29. 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 ?? au théorème de Banach-Steinhaus. □

Proposition 18.30. Si $u_n \rightharpoonup u$ et $|u_n| \rightarrow |u|$, alors la suite u_n converge fortement vers u .

Démonstration: On écrit

$$|u_n - u|^2 = |u_n|^2 - 2\langle u_n | u \rangle + |u|^2.$$

On a $(u_n, u) \rightarrow |u|^2$ d'où $|u_n - u|^2 \rightarrow 0$. □

Proposition 18.31. Soient E et F deux espaces de Hilbert, et $T \in \mathcal{L}(E, F)$. Alors

$$u_n \rightharpoonup u \implies Tu_n \rightharpoonup Tu.$$

Démonstration: On écrit simplement que, pour tout $z \in F$,

$$\langle Tu_n | z \rangle = \langle u_n | T^* z \rangle \longrightarrow \langle u | T^* z \rangle = \langle Tu | z \rangle,$$

qui exprime la convergence faible de Tu_n vers Tu . □

Le résultat fondamental de cette section est le suivant.

Théorème 18.32. Soit (u_n) une suite **bornée** dans un espace de Hilbert H . Alors on peut extraire une sous-suite convergente 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 $\langle u_n | x_1 \rangle$ est bornée dans \mathbb{R} on peut extraire une suite $u_{j_1(n)}$ telle que $\langle u_{j_1(n)} | x_1 \rangle$ converge.
2. Comme $\langle u_{j_1(n)} | x_2 \rangle$ est bornée dans \mathbb{R} on peut extraire de $u_{j_1(n)}$ une suite $u_{j_1 \circ j_2(n)}$ telle que $\langle u_{j_1 \circ j_2(n)} | x_2 \rangle$ converge.
3. Par récurrence, on construit une suite de sous-suites emboîtées $u_{j_1 \circ j_2 \circ \dots \circ j_k(n)}$ telle que $(u_{j_1 \circ j_2 \circ \dots \circ j_k(n)}, x_k)$ converge, pour tout k .
4. On utilise à présent le procédé d'extraction diagonale : on pose $\varphi(k) = j_1 \circ j_2 \circ \dots \circ j_k(k)$ (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 $(u_{j_1 \circ j_2 \circ \dots \circ j_k(n)})$, de telle sorte que $\langle u_{\varphi(n)} | x_k \rangle$ converge lorsque $n \rightarrow +\infty$.
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 $\langle u_{\varphi(n)} | x_k \rangle$ est de Cauchy, il existe un N au-delà duquel $|\langle u_{\varphi(p)} | x_k \rangle - \langle u_{\varphi(q)} | x_k \rangle| < \varepsilon$. Pour tous p, q supérieurs à N , on a donc

$$\begin{aligned} |\langle u_{\varphi(p)} | x \rangle - \langle u_{\varphi(q)} | x \rangle| &\leq |\langle u_{\varphi(p)} | x \rangle - \langle u_{\varphi(p)} | x_k \rangle| + |\langle u_{\varphi(p)} | x_k \rangle - \langle u_{\varphi(q)} | x_k \rangle| \\ &\quad + |\langle u_{\varphi(q)} | x_k \rangle - \langle u_{\varphi(q)} | x \rangle| \\ &\leq M\varepsilon + \varepsilon + M\varepsilon = (1 + 2M)\varepsilon, \end{aligned}$$

où M est un majorant de $|u_n|$.

On a donc démontré que, pour tout $x \in H$, $\langle u_{\varphi(n)} | x \rangle$ 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 $x \in H$

$$|h(x)| = \lim_{n \rightarrow \infty} |\langle u_{\varphi(n)} | x \rangle| \leq M|x|,$$

d'où h continue¹ 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 .

1. Remarquer qu'il n'est pas nécessaire ici d'utiliser le théorème de Banach-Steinhaus, du fait de l'hypothèse (u_n) bornée.

18.3 Somme Hilbertienne, bases Hilbertiennes

Definition 18.33. (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 Hilbertienne des E_n si

- (i) Les E_n sont deux à deux orthogonaux, c'est-à-dire

$$\langle u, v \rangle = 0 \quad \forall u \in E_n, \forall v \in E_m \quad \forall m, n \in \mathbb{N}, m \neq n.$$

- (ii) L'espace vectoriel engendré par les E_n est dense dans H .

Théorème 18.34. On suppose que H est 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 $S_k u$ vérifie (les E_n sont orthogonaux deux à deux)

$$|S_k u|^2 = \sum_{n=1}^k |u_n|^2.$$

D'autre part on a, pour tout n

$$\langle u | u_n \rangle = |u_n|^2,$$

d'où, en sommant de 1 à k ,

$$\langle u | S_k u \rangle = |S_k u|^2.$$

On a donc $|S_k u| \leq |u|$. 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 $S_k v = v$, et ainsi

$$|S_k u - u| \leq |S_k(u - v)| + |v - u| \leq 2\varepsilon.$$

on a donc bien convergence de $S_k u$ vers u .

D'autre part l'égalité, pour tout k

$$|S_k u|^2 = \sum_{n=1}^k |u_n|^2,$$

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. \square

Le théorème précédent permet d'introduire la notion de base Hilbertienne :

Definition 18.35. (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 18.36. Tout espace de Hilbert séparable admet une base Hilbertienne.

Démonstration. Soit H un espace de Hilbert séparable². 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. \square

18.4 Décomposition spectrale des opérateur auto-adjoints compacts

Le résultat principal de cette section est le théorème de décomposition spectrale des opérateurs auto-adjoints compacts positifs.

Lemme 18.37. Soit V un espace de Hilbert et $T \in L(V)$ un opérateur auto-adjoint compact et positif, i.e. $\langle Tv | v \rangle \geq 0$ pour tout $v \in V$. On note

$$M = \sup_{v \in V \setminus \{0\}} \frac{\langle Tv | v \rangle}{|v|^2}.$$

On a alors $M = \|T\|$, et M est valeur propre de T , i.e. il existe w tel que $Aw = Mw$.

Démonstration. On a

$$|\langle Tv | v \rangle| \leq \|T\| |v|^2,$$

d'où $M \leq \|T\|$.

Par ailleurs, pour tous u et v dans V , on a

$$\begin{aligned} 4\langle Tu | v \rangle &= \langle T(u+v) | (u+v) \rangle - \langle T(u-v) | (u-v) \rangle \leq \langle T(u+v) | (u+v) \rangle \\ &\leq M|u+v|^2 \leq 2M(|u|^2 + |v|^2), \end{aligned}$$

et ainsi

$$\|T\| = \sup_{|u|=1, |v|=1} \langle Tu | v \rangle \leq M.$$

On a donc $\|T\| = M$. Considérons maintenant une suite maximisante (u_n) , avec $|u_n| = 1$, et

$$\langle Tu_n | u_n \rangle \rightarrow M \quad n \rightarrow +\infty.$$

L'opérateur T étant compact, on peut extraire une sous-suite (notée toujours u_n pour simplifier) telle que

$$Tu_n \rightarrow w.$$

On a

$$|Tu_n - Mu_n|^2 = |Tu_n|^2 - 2M\langle Tu_n | u_n \rangle + M^2 \leq M^2 - 2M\langle u_n | u_n \rangle + M^2 \rightarrow 0 \text{ qd } n \rightarrow +\infty.$$

On a donc convergence forte de u_n vers w/M , d'où, par continuité de T , convergence de Tu_n vers Tw/M . On a donc finalement $Tw = Mv$.

\square

2. C'est à dire qu'il existe un ensemble dénombrable et dense. C'est le cas pour l'essentiel des espaces de Hilbert que l'on rencontre dans la "nature", en particulier pour les espaces fonctionnels de type $L^2(\Omega)$ ou $H^m(\Omega)$.

Proposition 18.38. Soit $T \in \mathcal{L}(V)$ un opérateur auto-adjoint. Deux vecteurs propres associés à des valeurs propres distinctes sont orthogonaux entre eux.

Démonstration. On a

$$Tu_1 = \lambda_1 u_1, \quad Tu_2 = \lambda_2 u_2 \implies \langle Tu_1, u_2 \rangle = \lambda_1 \langle u_1 | u_2 \rangle = \langle Tu_2 | u_1 \rangle = \lambda_2 \langle u_2 | u_1 \rangle,$$

d'où $(\lambda_2 - \lambda_1) \langle u_1 | u_2 \rangle = 0$, d'où la conclusion. \square

Lemme 18.39. Soit V un espace de Hilbert et $T \in \mathcal{L}(V)$ un opérateur auto-adjoint compact. Pour tout $\delta > 0$, il n'existe qu'un nombre fini de valeurs propres (comptées avec leur multiplicité) en dehors de l'intervalle $]-\delta, \delta[$.

Démonstration. Supposons qu'il existe une infinité d'éléments propres en dehors de l'intervalle considéré, on peut alors construire une suite (w_k, λ_k) , avec $Tw_k = \lambda_k w_k$, et les w_k de norme 1. Par hypothèse, (w_k/λ_k) est bornée et, comme T est compact, on peut donc une sous-suite telle que $Tw_{k'}/\lambda_{k'} = w_{k'}$ converge. Or, comme les w_k sont orthogonaux deux à deux, on a

$$|w_p - w_q|^2 = 2 \quad \forall p \neq q,$$

ce qui est en contradiction avec le critère de Cauchy pour la suite extraite convergente. \square

Théorème 18.40. Soit V un espace de Hilbert séparable de dimension infinie et $T \in L(V)$ un opérateur auto-adjoint compact défini positif, i.e. tel que $\langle Tu | u \rangle > 0$ pour tout $u \neq 0$. Alors T admet une suite infinie de valeurs propres (μ_k) strictement positives (numérotées dans l'ordre décroissant) qui décroissent vers 0. Les vecteurs propres associés engendrent un espace vectoriel dense dans V . Plus précisément, la suite des vecteurs propres normalisés est une base hilbertienne de V , c'est à dire que tout élément $v \in V$ admet la décomposition suivante :

$$v = \sum_{n=1}^{+\infty} \alpha_n u_n, \quad \alpha_n = \langle v | u_n \rangle.$$

Démonstration. D'après le lemme 18.37, l'ensemble des valeurs propres de T n'est pas vide. D'après le lemme 18.39, c'est un ensemble soit fini, soit infini dénombrable avec 0 comme seul point d'accumulation, 0 lui-même n'étant pas valeur propre. En construisant une base orthonormale de chacun des sous-espace propres (qui sont de dimension finie), et en utilisant la proposition 18.38, on peut construire une suite (u_k) de vecteurs propres unitaires associés aux valeurs propres λ_k (avec possible redondance). On note W l'adhérence dans V de l'espace vectoriel engendré par les u_k . Cet espace est stable par T , ainsi que son orthogonal W^\perp . La restriction de T à W^\perp est toujours un opérateur linéaire continu auto-adjoint compact, qui admet donc (si ça n'est pas l'opérateur nul), une valeur propre strictement positive, avec un vecteur propre associé, qui est de fait vecteur propre pour l'opérateur initial. C'est en contradiction avec le fait que W contenait tous les vecteurs propres de T , la restriction à W^\perp est donc nécessairement nulle, donc (T est défini positif) W^\perp est réduit au vecteur nul. \square

Théorème 18.41. (Valeurs propres d'un problème variationnel)

Soient V et H deux espaces de Hilbert, de dimension infinie, avec injection $V \subset H$ compacte et dense. Soit $a(\cdot, \cdot)$ une forme bilinéaire symétrique continue et coercive sur $V \times V$. Le problème de recherche d'un couple $(u, \lambda) \in H \times \mathbb{R}$ tel que

$$a(u, v) = \lambda \langle u | v \rangle \quad \forall v \in V,$$

admet une infinité de solutions. Les λ solutions, appelées valeurs propres de a , forment une suite

$$0 < \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_k \dots$$

qui tend vers l'infini. La famille (u_k) des vecteurs propres associés est une base hilbertienne de H , et $(u_k / \sqrt{\lambda_k})$ est une base hilbertienne de V pour le produit scalaire associé à $a(\cdot, \cdot)$.

Démonstration. Pour tout $f \in H$, le problème

$$a(u, v) = \langle f | v \rangle \quad \forall v \in V,$$

admet une solution unique u d'après le théorème de Lax-Milgram, on note Tf cette solution. L'opérateur T est linéaire de H dans V , et l'on a (on prend $v \in Tf$ dans la formulation variationnelle) :

$$a(Tf, Tf) = \langle f | Tf \rangle,$$

d'où

$$\alpha |Tf|_V^2 \leq a(Tf, Tf) \leq \|f\|_H \|Tf\|_H \leq C \|f\|_H |Tf|_V,$$

et ainsi

$$|Tf|_V \leq C \|f\|_H.$$

Cet opérateur, en tant qu'élément de $L(H)$, est donc compact par injection continue de V dans H . On a par ailleurs

$$\langle Tf | f \rangle = a(Tf, Tf)$$

qui est strictement positif dès que f est non nul. On a enfin

$$\langle f | Tg \rangle = a(Tf, Tg) = a(Tg, Tf) = \langle g | Tf \rangle,$$

ce qui assure le caractère auto-adjoint. Le théorème assure donc l'existence d'une suite de valeurs propres pour T , décroissante vers 0, avec une base Hilbertienne de vecteurs propres associés.

Retournons maintenant au problème de départ, qui s'écrit

$$a(u, v) = \lambda \langle u | v \rangle = \lambda a(Tu, v) \quad \forall V \in H,$$

qui est équivalent à

$$\lambda Tu = u.$$

Ce problème admet donc une suite de valeurs propres μ_i qui sont les inverses des valeurs propres de T , pour les mêmes vecteurs propres (u_k) . Les fonctions propres (u_k) associées, normalisées à 1 pour H , forment une base Hilbertienne de H .

Cette famille est aussi orthogonale pour le produit scalaire défini par $a(\cdot, \cdot)$ (d'après la proposition 18.38), et l'on a

$$a(u_k, u_k) = \lambda_k \|u_k\|^2 = \lambda_k.$$

La famille $(u_k / \sqrt{\lambda_k})$ est donc une base Hilbertienne sur V , pour le produit scalaire associé à la forme bilinéaire $a(\cdot, \cdot)$. \square

Remarque 18.42. On peut affaiblir l'hypothèse de coercivité de $a(\cdot, \cdot)$ dans le théorème précédent, en supposant seulement qu'il existe $\eta > 0$ et $\alpha > 0$ tels que

$$a(v, v) + \eta \|v\|^2 \geq \alpha |v|_V^2.$$

Le théorème reste inchangé, sauf que les valeurs propres ne sont pas forcément positives : il peut y avoir un premier paquet (fini) de valeurs propres négatives ou nulles. Cette remarque permet d'appliquer notamment le théorème au Laplacien avec conditions de Neuman sur le bord du domaine.

Dans le contexte du théorème précédent, on définit pour tout $v \in V$ le quotient de Rayleigh par

$$R(v) = \frac{a(v, v)}{\|v\|_H^2}.$$

Théorème 18.43. (Courant-Fisher)

On se place dans les hypothèses du théorème 18.41. On note E_k l'ensemble des sous-espaces vectoriels de V de dimension k . On a

$$\lambda_k = \min_{W \in E_k} \max_{w \in W \setminus \{0\}} R(w) = \max_{W \in E_{k-1}} \min_{w \in W^\perp \setminus \{0\}} R(w).$$

Remarque 18.44. La démonstration du théorème 18.40 permet de donner une définition simple de chaque valeur propre à partir des sous-espaces propres des valeurs propres précédentes. Si l'on note $F_k = \text{vec}(w_1, \dots, w_k)$, on a

$$\lambda_{k+1} = \min_{v \in F_k^\perp} \frac{a(v, v)}{\|v\|_H^2}.$$

18.5 Problèmes d'évolution

Théorème 18.45. Soient V et H deux espaces de Hilbert, de dimension infinie, avec injection $V \subset H$ compacte et dense. Soit $a(\cdot, \cdot)$ une forme bilinéaire symétrique continue et coercive sur $V \times V$, et $f \in L^2([0, T], H)$ un terme source. Le problème

$$\frac{d}{dt} \langle u(t) | v \rangle + a(u(t), v) = \langle f(t) | v \rangle \quad \forall v \in V, \quad 0 < t < T \quad (18.7)$$

avec condition initiale $u(0) = u_0 \in H$, a une unique solution $u \in L^2([0, T], V) \cap L^\infty([0, T], H)$. Cette solution s'exprime

$$u(t) = \sum_{k=1}^{+\infty} u_k^0 e^{-\lambda_k t} w_k + \sum_{k=1}^{+\infty} \left(\int_0^t e^{-\lambda_k(t-s)} \langle f(s), w_k \rangle ds \right) w_k,$$

où λ_k est la suite des valeurs propres du problème variationnel associé à $a(\cdot, \cdot)$, et (w_k) la base Hilbertienne associée (voir théorème 18.41), et (u_k^0) correspond à la décomposition de la donnée initiale dans cette base :

$$u^0 = \sum_{k=1}^{+\infty} u_k^0 w_k, \quad u_k^0 = \langle u^0 | w_k \rangle,$$

Démonstration. On raisonne dans un premier temps par condition nécessaire pour construire une solution Si le problème admet une solution régulière $u(x, t)$, on peut, pour tout t , décomposer $u(t)$ sur la base Hilbertienne w_k :

$$u(t) = \sum_{k=1}^{+\infty} u_k(t) w_k.$$

On injecte cette expression dans la formulation variationnelle, et on prend $v = w_k$:

$$\dot{u}_k(t) + \lambda_k u_k(t) = f_k(t).$$

La solution de cette équation différentielle ordinaire s'écrit

$$u_k(t) = u_k^0 e^{-\lambda_k t} + \int_0^t e^{-\lambda_k(t-s)} \langle f(s) | w_k \rangle ds.$$

On termine la démonstration en vérifiant que la série ainsi définie est bien de Cauchy dans les espaces fonctionnels $L^2([0, T], V)$ et $L^\infty([0, T], H)$. (voir détails dans Allaire³, section 8.2). □

Théorème 18.46. Soient V et H deux espaces de Hilbert, de dimension infinie, avec injection $V \subset H$ compacte et dense. Soit $a(\cdot, \cdot)$ une forme bilinéaire symétrique continue et coercive sur V , et $f \in L^2([0, T], H)$ un terme source. On se donne une donnée initiale $(u^0, u^1) \in V \times H$. Le problème

$$\frac{d^2}{dt^2} \langle u(t) | v \rangle + a(u(t), v) = \langle f(t) | v \rangle \quad \forall v \in V, \quad t \in [0, T],$$

avec conditions initiales $u(0) = u^0$, $du/dt(0) = u^1$, a une unique solution $u \in C([0, T], V) \cap C^1([0, T], H)$ qui s'écrit

$$u(t) = \sum_{k=1}^{+\infty} \left(u_k^0 \cos(\omega_k t) + \frac{u_k^1}{\omega_k} \sin(\omega_k t) \right) w_k + \sum_{k=1}^{+\infty} \left(\int_0^t e^{-\lambda_k(t-s)} \langle f(s) | w_k \rangle ds \right) w_k,$$

avec

$$u_k^0 = \langle u^0 | w_k \rangle, \quad u_k^1 = \langle u^1 | w_k \rangle, \quad \omega_k = \sqrt{\lambda_k},$$

où (λ_k) est la suite des valeurs propres du problème variationnel associé à $a(\cdot, \cdot)$, et (w_k) la base Hilbertienne associée (voir théorème 18.41),

Démonstration. Voir cours polycopié de G. Allaire précédemment cité. □

3. Grégoire Allaire, Analyse Numérique et Optimisation, cours de l'École Polytechnique,
https://www.editions.polytechnique.fr/files/pdf/EXT_1255_8.pdf

18.6 Minimisation de fonctionnelles convexes

Commençons par définir un certain nombre de notions générales afférentes aux applications à valeurs dans $\mathbb{R} \cup \{+\infty\}$.

Definition 18.47. (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 18.48. (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 18.49. (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)

$$\text{epi } J = \{(x, \lambda) \in E \times \mathbb{R}, J(x) \leq \lambda\},$$

est convexe.

On dit que J est strictement convexe si

$$J(\theta x + (1 - \theta)y) < \theta J(x) + (1 - \theta)J(y) \quad \forall x, y \in E \quad \forall \theta \in]0, 1[.$$

Théorème 18.50. (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 $x_n \rightarrow 0$. 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

$$|(y_1, x_{\varphi(2)})| = |(y_1, y_2)| \leq \frac{1}{2}.$$

3. Par récurrence, on construit à partir des termes déjà construits y_1, y_2, \dots, y_{n-1} , le n -ième terme y_n tel que

$$|(y_i, y_n)| \leq \frac{1}{n} \quad \forall i = 1, 2, \dots, n-1.$$

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

$$\begin{aligned} |\sigma_n|^2 &\leq \frac{1}{n^2} \left(\sum_{i=1}^n |y_i|^2 + 2 \sum_{k=1}^n \sum_{\ell=1}^{k-1} |(y_\ell, y_k)| \right) \leq \frac{1}{n^2} \left(nM^2 + 2 \sum_{k=1}^n \frac{k-1}{k} \right) \\ &\leq \frac{1}{n^2} (nM^2 + 2n) = \frac{M^2 + 2}{n}, \end{aligned}$$

et donc $\sigma_n \rightarrow 0$. \square

Ce théorème a plusieurs conséquences importantes, dont la première est le

Théorème 18.51. 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 18.50. \square

Exercice 18.7. 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 18.50 est le

Théorème 18.52. Soit $J : H \rightarrow \mathbb{R}$ une fonction convexe 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

$$J(y_n) \rightarrow L,$$

et telle que

$$\sigma_n = \frac{1}{n} \sum_{i=1}^n y_n \rightarrow x.$$

par semi-continuité inférieure de J , on a $J(x) \leq \liminf J(\sigma_n)$. D'autre part, J étant convexe

$$J(\sigma_n) \leq \frac{1}{n} \sum_{i=1}^n J(y_n) \rightarrow L.$$

On a donc bien $J(x) \leq L$. \square

Ce théorème va nous permettre d'établir le résultat principal de minimisation :

Théorème 18.53. Soit $J : H \rightarrow \mathbb{R}$ une fonction convexe s.c.i., $J \not\equiv +\infty$. On suppose que J est coercive, c'est-à-dire que

$$\lim_{|x| \rightarrow +\infty} J(x) = +\infty.$$

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

$$J(x_n) \rightarrow M := \inf_K J.$$

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 $J(x) \geq M$ 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

$$\lim_{|x| \rightarrow +\infty, x \in K} J(x) = +\infty.$$

Definition 18.54. (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

$$\partial\Psi(u) = \{w \in H, \Psi(u) + \langle w | h \rangle \leq \Psi(u+h) \quad \forall h \in H\}.$$

Chapitre 19

Compléments

19.1 Inégalités

Proposition 19.1. (Inégalité arithmético-géométrique)

Soient x_1, \dots, x_n des réels positifs ou nuls, et $(\alpha_n) \in [0, +\infty[^n$ une famille de poids. On a

$$(x_1^{\alpha_1} \dots x_n^{\alpha_n})^{1/\alpha} \leq \frac{1}{\alpha} \sum_{i=1}^n \alpha_i x_i,$$

avec $\alpha = \sum \alpha_i$.

Démonstration. Par concavité de la fonction logarithme, on a

$$\frac{1}{\alpha} \sum \alpha_n \log x_n \leq \log \left(\frac{1}{\alpha} \sum \alpha_n x_n \right),$$

d'où l'inégalité en prenant l'exponentielle. \square

Proposition 19.2. (Inégalité de Young)

Soient a et b deux réels positifs où nuls, et p, q deux réels > 0 conjugués, i.e. tels que $\frac{1}{p} + \frac{1}{q} = 1$. On a alors

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}.$$

Démonstration. C'est une conséquence de l'inégalité arithmético-géométrique (proposition 19.1), avec $\alpha_1 = 1/p$, $\alpha_2 = 1/q$, $x_1 = a^p$, et $x_2 = b^q$. \square

Proposition 19.3. (Inégalité de Hölder)

Soient p et q deux réels positifs conjugués, i.e. tels que $1/p + 1/q = 1$, et $\theta_i \in [0, +\infty[^n$. Pour tous $x = (x_i)$, $y = (y_i) \in \mathbb{R}^n$, on a

$$\sum_{i=1}^n |\theta_i x_i y_i| \leq \left(\sum_{i=1}^n \theta_i |x_i|^p \right)^{1/p} \left(\sum_{i=1}^n \theta_i |y_i|^q \right)^{1/q}.$$

Proposition 19.4. (Inégalité de Minkowski)

Soit $p \in [1, +\infty]$, et $\theta_i \in [0, +\infty[^n$. Pour tous $x = (x_i)$, $y = (y_i) \in \mathbb{R}^n$, on a

$$\left(\sum_{i=1}^n \theta_i |x_i + y_i|^p \right)^{1/p} \leq \left(\sum_{i=1}^n \theta_i |x_i|^p \right)^{1/p} + \left(\sum_{i=1}^n \theta_i |y_i|^p \right)^{1/p}.$$

Démonstration. On applique l'inégalité de Hölder avec les indices p et $p/(p-1)$.

□

19.2 Théorème d'Arzela Ascoli

Théorème 19.5. Soit X un espace métrique compact, et F une famille de fonctions de K dans \mathbb{R} , supposée bornée (il existe M tel que $|f(x)| \leq M$ pour tout $f \in F$, tout $x \in X$), et *uniformément équicontinue*, i.e.¹

$$\forall \varepsilon > 0, \exists \eta > 0 \text{ t.q. } \forall x, y \in K, d(x, y) < \eta \implies |f(x) - f(y)| < \varepsilon \quad \forall f \in F.$$

Alors la famille F est relativement compacte, i.e. de toute suite de fonctions de F on peut extraire une sous-suite uniformément convergente.

Proposition 19.6. Soit K un compact de \mathbb{R}^d , et F une famille bornée (norme L^∞) de fonction L -Lipschitziennes. Alors F est relativement compacte dans $C(K)$.

19.3 Théorèmes de point fixe

Théorème 19.7. (Théorème de 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 unique point fixe x . Comme $T^p(T(x)) = T \circ T^p(x) = T(x)$, l'élément $T(x)$ est aussi point fixe de T^p , il s'identifie donc à x par unicité : on a $T(x) = x$. □

Théorème 19.8. (Théorème de point fixe de Brouwer)

Soit K un convexe fermé de l'espace euclidien \mathbb{R}^d , et T une application continue de K dans lui-même. Alors T admet un point fixe : il existe $x \in K$ tel que $T(x) = x$.

19.4 Théorème de Krein-Milman

Definition 19.9. (Points extrémaux)

Soit F une partie convexe d'un espace affine E . On dit que $x \in F$ est *point extrémal* de F si

$$x = \frac{x_0 + x_1}{2}, \quad x_0, x_1 \in F \implies x_0 = x_1 = x.$$

1. On prendra garde au fait que le η est le même pour tous les (x, y) et tous les f_n de la famille.

Théorème 19.10. (Krein-Milman)

Soit F une partie convexe compacte d'un espace affine de dimension finie. Alors F l'enveloppe convexe de ses points extrémaux.

Ce théorème se généralise à la dimension infinie, plus précisément aux espaces localement convexes séparés, dans lesquels tout convexe compact est enveloppe convexe fermée de ses points extrémaux.

19.5 Théorèmes des fonctions implicites et d'inversion locale

Théorème 19.11. Soit f une fonction définie sur un ouvert W de $\mathbb{R}^n \times \mathbb{R}^m$, à valeurs dans \mathbb{R}^m . On suppose f continûment différentiable sur W , et l'on suppose que la différentielle partielle de f par rapport à y , notée $\partial_y f(x, y)$, est inversible en tout point² de W . On considère un point $(x_0, y_0) \in W$ qui annule f :

$$f(x_0, y_0) = 0.$$

On peut alors exprimer y comme fonction de x au voisinage de (x_0, y_0) . Plus précisément : il existe des voisinages ouverts $U \in \mathbb{R}^n$ et $V \in \mathbb{R}^m$ de x_0 et y_0 , respectivement, et une fonction Ψ de U dans V , tels que

$$(x, y) \in U \times V, \quad f(x, y) = 0 \iff y = \Psi(x).$$

La fonction Ψ est continûment différentiable sur U , et sa différentielle s'exprime

$$d\Psi(x) = -(\partial_y f(x, y))^{-1} \circ \partial_x f(x, y), \quad \text{avec } y = \Psi(x).$$

Démonstration. La démarche, de nature constructive, est basée sur un processus itératif construit selon les principes suivants. On considère x proche de x_0 (dans un sens précisé plus loin), et l'on cherche y tel que $f(x, y) = 0$. On suppose que l'on dispose d'une première approximation y_k du y recherché, et on cherche un y_{k+1} qui en soit une meilleure approximation. On a

$$f(x, y_{k+1}) = f(x, y_k + (y_{k+1} - y_k)) \approx f(x, y_k) + \partial_y f(x, y_k) \cdot (y_{k+1} - y_k).$$

On souhaite annuler cette quantité, ce qui suggère de définir y_{k+1} comme

$$y_{k+1} = y_k - (\partial_y f(x, y_k))^{-1} \cdot f(x, y_k).$$

Il s'agit de la méthode dite de *Newton* pour trouver le zéro d'une fonction. Nous allons considérer ici une version modifiée de cette méthode, en remplaçant la différentielle partielle en y par sa valeur au point (x_0, y_0) . Partant de y_0 (en fait, on peut partir d'une valeur initiale différente de y_0 , mais nous le fixons comme point de départ pour simplifier), on construit donc la suite (y_k) par récurrence, selon la formule

$$y_{k+1} = y_k - Q^{-1} \cdot f(x, y_k), \quad \text{avec } Q = \partial_y f(x_0, y_0).$$

N.B. : On prendra garde au fait que, pour (x, y) donné, $\partial_y f(x, y)$ est une application linéaire de \mathbb{R}^m dans \mathbb{R}^m . Cette application dépend du point $(x, y) \in \mathbb{R}^n \times \mathbb{R}^m$ où elle est prise, mais sans que la différentielle soit prise par rapport à la variable x . Cette différentielle partielle est définie par le développement limité suivant, où l'on ne perturbe que la variable y : pour $h \in \mathbb{R}^m$,

$$f(x, y + h) = f(x, y) + \partial_y f(x, y) \cdot h + o(h).$$

Il s'agit donc d'un champ d'applications linéaires, auquel on peut associer un champ de matrices carrées $m \times m$ (leurs représentations dans la base canonique de \mathbb{R}^m), qui vit sur un espace de dimension $n \times m$. L'application Q est simplement la valeur particulière de ce champ au point (x_0, y_0) .

Cette récurrence peut s'écrire $y_{k+1} = \Phi_x(y_k)$, où la fonction Φ_x est définie par

$$y \mapsto \Phi_x(y) = y - Q^{-1} \cdot f(x, y),$$

2. Comme précisé dans la remarque 19.13 ci-après, il suffit de vérifier que la différentielle soit inversible en (x_0, y_0) pour qu'elle le soit dans un voisinage de ce point.

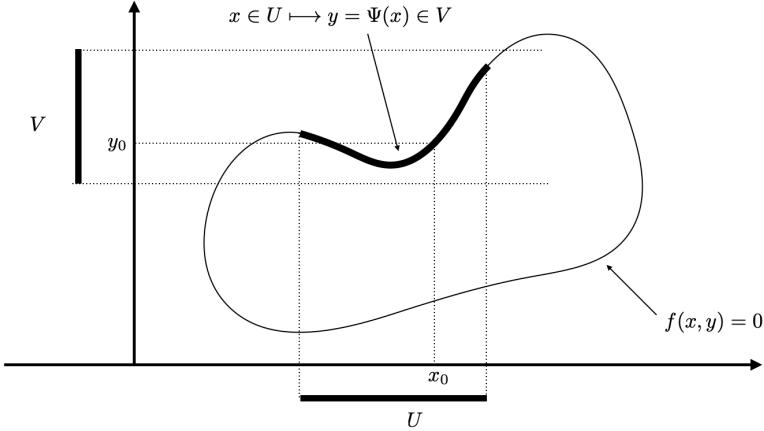


FIGURE 19.1 – Théorème des fonctions implicites

pour tout y tel que $(x, y) \in W$. Noter que y est point fixe de Φ_x si et seulement si $f(x, y) = 0$. Nous allons montrer que cette fonction admet bien un unique point fixe sur un voisinage de y_0 . Cette fonction est différentiable sur son domaine de définition, de différentielle

$$d\Phi_x(y) = I - Q^{-1} \circ \partial_y f(x, y).$$

En écrivant $I = Q^{-1}Q$ on obtient

$$\|d\Phi_x(y)\| = \|Q^{-1}(\partial_y f(x_0, y_0) - \partial_y f(x, y))\| \leq \|Q^{-1}\| \|\partial_y f(x_0, y_0) - \partial_y f(x, y)\|$$

Fixons $\kappa = 1/2$. La différentielle étant continue, il existe un $r > 0$ tel que, pour tout point $x \in \overline{B}(x_0, r)$, tout $y \in \overline{B}(y_0, r)$ (on prend r suffisamment petit pour que $\overline{B}(x_0, r) \times \overline{B}(y_0, r) \subset W$),

$$\|\partial_y f(x_0, y_0) - \partial_y f(x, y)\| \leq \kappa \|Q^{-1}\|^{-1},$$

de telle sorte que

$$\forall x \in \overline{B}(x_0, r), \quad y \in \overline{B}(y_0, r), \quad \|d\Phi_x(y)\| \leq \kappa.$$

On a donc, pour tous y, y' dans $\overline{B}(y_0, r)$,

$$\|\Phi_x(y) - \Phi_x(y')\| \leq \kappa \|y - y'\|$$

d'après le théorème des accroissements finis (théorème ??, page ??), avec $\kappa = 1/2$. L'application Φ_x est donc contractante sur $\overline{B}(y_0, r)$. Montrons qu'elle laisse stable une boule autour de y_0 . Comme l'application

$$x \mapsto \Phi_x(y_0) = y_0 - Q^{-1} \cdot f(x, y_0)$$

est continue en x_0 , il existe un $r' < r$ tel que, pour tout $x \in \overline{B}(x_0, r')$, on ait

$$\|\Phi_x(y_0) - \Phi_{x_0}(y_0)\| \leq (1 - \kappa)r,$$

avec $\Phi_{x_0}(y_0) = y_0$ car $f(x_0, y_0) = 0$. On a alors, pour tout $x \in \overline{B}(x_0, r')$, tout $y \in \overline{B}(y_0, r)$,

$$\|\Phi_x(y) - y_0\| = \|\Phi_x(y) - \Phi_{x_0}(y_0)\| \leq \underbrace{\|\Phi_x(y) - \Phi_x(y_0)\|}_{\leq \kappa \|y - y_0\|} + \underbrace{\|\Phi_x(y_0) - y_0\|}_{\leq (1 - \kappa)r} \leq \kappa r + (1 - \kappa)r = r.$$

Pour tout $x \in \overline{B}(x_0, r')$, l'application Φ_x est donc bien définie de $\overline{B}(y_0, r)$ dans lui-même, et cet ensemble est complet comme fermé dans le complet \mathbb{R}^m . Elle par ailleurs contractante comme montré précédemment. D'après le théorème 19.7, elle admet donc un unique point fixe sur $\overline{B}(y_0, r)$, c'est-à-dire qu'il existe un unique $y \in \overline{B}(y_0, r)$ tel que $f(x, y) = 0$. On note Ψ l'application qui à x associe cette unique solution en y de $f(x, y) = 0$.

Montrons maintenant la continuité de Ψ , et précisons le choix des voisinages U et V . Soient x_1 et x_2 deux points de $\overline{B}(x_0, r')$, et $y_1 = \Psi(x_1)$, $y_2 = \Psi(x_2)$. On a

$$\|y_2 - y_1\| = \|\Phi_{x_2}(y_2) - \Phi_{x_1}(y_1)\| \leq \|\Phi_{x_2}(y_2) - \Phi_{x_2}(y_1)\| + \|\Phi_{x_2}(y_1) - \Phi_{x_1}(y_1)\|.$$

Comme Φ_{x_2} est κ -contractante sur $\overline{B}(y_0, r)$, on a $\|\Phi_{x_2}(y_2) - \Phi_{x_2}(y_1)\| \leq \kappa \|y_2 - y_1\|$, d'où

$$\begin{aligned} \|y_2 - y_1\| &\leq \frac{1}{1 - \kappa} \|\Phi_{x_2}(y_1) - \Phi_{x_1}(y_1)\| = \frac{1}{1 - \kappa} \|Q^{-1} \cdot (f(x_2, y_1) - f(x_1, y_1))\| \\ &\leq \frac{1}{1 - \kappa} \|Q^{-1}\| \max_{\overline{B}(x_0, r') \times \overline{B}(y_0, r)} \|\partial_x f\| \|x_2 - x_1\| \end{aligned}$$

d'après le théorème des accroissements finis ?? (f étant continûment différentiable sur le compact $\overline{B}(x_0, r') \times \overline{B}(y_0, r)$, sa différentielle partielle par rapport à x est bornée). Cette quantité tend en particulier vers 0 quand x_2 tend vers x_1 . L'application Ψ est donc continue sur $\overline{B}(x_0, r')$ à valeurs dans $\overline{B}(y_0, r)$, et même lipschitzienne : il existe $C > 0$ tel que

$$\|\Psi(x_2) - \Psi(x_1)\| \leq C \|x_2 - x_1\|.$$

Soit V voisinage ouvert de y_0 inclus dans $\overline{B}(y_0, r)$. Comme Ψ est continue, il existe un voisinage ouvert de x_0 , $U \subset \overline{B}(x_0, r')$, tel que $\Psi(U) \subset V$.

Il reste à montrer que Ψ est différentiable sur U . Soit $x \in U$, $y = \Psi(x) \in V$. On considère une variation h de x telle que $x + h \in U$. Il existe un unique g tel que $y + g \in V$ vérifie

$$f(x + h, y + g) = 0.$$

D'après ce qui précède il existe $C > 0$ tel que $\|g\| \leq C \|h\|$. La différentiabilité de f en (x, y) s'exprime

$$\underbrace{f(x + h, y + g)}_{=0} = \underbrace{f(x, y)}_{=0} + \partial_x f(x, y) \cdot h + \partial_y f(x, y) \cdot g + o(h, g).$$

On a donc

$$g = -\left((\partial_y f(x, y))^{-1} \circ \partial_x f(x, y)\right) \cdot h + o(h),$$

(le $o(h, g)$ s'est bien transformé en $o(h)$ du fait que la norme de h domine celle de g , comme indiqué précédemment). Ce g est, par construction, $\Psi(x + h) - \Psi(x)$, on a donc

$$\Psi(x + h) = \Psi(x) - \left((\partial_y f(x, y))^{-1} \circ \partial_x f(x, y)\right) \cdot h + o(h).$$

ce qui exprime que l'application $x \mapsto \Psi(x)$ est différentiable sur U , de différentielle

$$d\Psi(x) = (\partial_y f(x, \Psi(x)))^{-1} \circ \partial_x f(x, \Psi(x)).$$

Comme f est continûment différentiable, et que Ψ est continue, $x \mapsto d\Psi(x)$ est continue. \square

Remarque 19.12. On notera que, par construction, Ψ est bien définie sur tout U mais, comme illustré par la figure 19.1, elle n'est pas nécessairement surjective (cette remarque sera importante pour la démonstration du théorème des fonctions implicites, dans lequel il s'agira de construire deux ouverts en bijection). Par ailleurs, pour $x \in U$, il peut exister plusieurs y tels que $(f(x, y) = 0$, mais un seul qui soit dans V .

Remarque 19.13. Pour vérifier l'applicabilité du théorème précédent en un point (x_0, y_0) qui annule f , et au voisinage duquel f est définie, il suffit de vérifier que la différentielle de f par rapport à y est inversible en (x_0, y_0) . En effet, si c'est le cas, l'application $(x, y) \mapsto \partial_y f(x, y)$ étant continue, et le déterminant étant une fonction continue, la différentielle reste inversible sur un ouvert de (x_0, y_0) , qui peut jouer le rôle du W dans les hypothèses du théorème précédent. On dira que le théorème des fonctions implicites s'applique *en* (x_0, y_0) , ou *au voisinage de* (x_0, y_0) .

Definition 19.14. Soit φ une application d'un ouvert $U \subset \mathbb{R}^n$ dans un ouvert $V = \varphi(U)$ dans \mathbb{R}^n . On dit que φ est un C^1 – difféomorphisme de U vers V si φ est bijective, et si φ et sa réciproque φ^{-1} sont continûment différentiables.

Proposition 19.15. On se place dans les hypothèses de la définition précédente. La différentielle de φ est inversible en tout point de U , et son inverse est la différentielle de l'application réciproque φ^{-1} : pour tout $x \in U$, $y = \varphi(x) \in V$,

$$d\varphi^{-1}(y) = (d\varphi(x))^{-1}.$$

Démonstration. On a, pour tout $y \in V$,

$$\varphi \circ \varphi^{-1}(y) = y.$$

La règle de différentiation en chaîne implique donc (avec $x = \varphi^{-1}(y)$)

$$d\varphi(x) \circ d\varphi^{-1}(y) = \text{Id},$$

qui conclut la preuve. \square

Théorème 19.16. (Inversion locale)

Soit φ une application continûment différentiable d'un ouvert $W \subset \mathbb{R}^n$ dans \mathbb{R}^n . On suppose que $d\varphi(x)$ est inversible pour tout $x \in W$. Alors φ est un C^1 -difféomorphisme local : pour tout $x_0 \in W$, il existe un voisinage ouvert $U \subset W$ de x_0 et un voisinage ouvert V de $y_0 = \varphi(x_0)$ tel que $\varphi|_U$ soit un C^1 -difféomorphisme de U vers V .

Démonstration. On considère l'application (noter que l'on écrit (y, x) du fait qu'il va s'agir, contrairement à l'usage, d'exprimer x en fonction de y) :

$$g : (y, x) \in \mathbb{R}^n \times W \mapsto g(y, x) = \varphi(x) - y.$$

Cette application est différentiable sur $\mathbb{R}^n \times W$, de différentielle partielle par rapport à x

$$\partial_x g(y, x) = d\varphi(x).$$

Cette différentielle est inversible sur W par hypothèse. Soit $x_0 \in W$, et $y_0 = \varphi(x_0)$, d'où $g(y_0, x_0) = 0$. D'après le théorème des fonctions implicites, il existe un voisinage V de y_0 , un voisinage $\tilde{U} \subset W$ de x_0 , et Ψ une application continûment différentiable de V dans \tilde{U} , tels que

$$(y, x) \in V \times \tilde{U}, \quad g(y, x) = 0, \quad \text{i.e. } y = \varphi(x) \iff x = \Psi(y).$$

L'application Ψ est donc la réciproque de φ . Il reste à préciser les voisinages ouverts de x_0 et y_0 qui sont en bijection. Il faut prendre garde à une difficulté (annoncée dans la remarque 19.12) : Ψ , qui est bien définie sur tout V , (cet ouvert V est noté U dans le théorème des fonctions implicites, du fait du renversement des rôles de x et y que nous avons effectué ici), n'est pas nécessairement *surjective* de V dans \tilde{U} . Pour garantir que les deux ouverts soient en bijection, on réduit l'ouvert \tilde{U} en introduisant

$$U = \tilde{U} \cap \Psi(V).$$

Comme, pour tout $y \in V$, l'équation $y = \varphi(x)$ n'a qu'une solution en $x \in \tilde{U}$, cet ensemble s'écrit aussi $U = \tilde{U} \cap \varphi^{-1}(V)$. Il s'agit donc bien d'un ouvert par continuité de V . \square

19.6 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 canoniquement à Ψ une application (linéaire et continue) de F dans E' , le dual topologique de E (espace des formes linéaires continues) :

$$y \in F \mapsto Ty \in E', \quad \langle Ty, x \rangle = \Psi(x, y) \quad \forall x \in E. \tag{19.1}$$

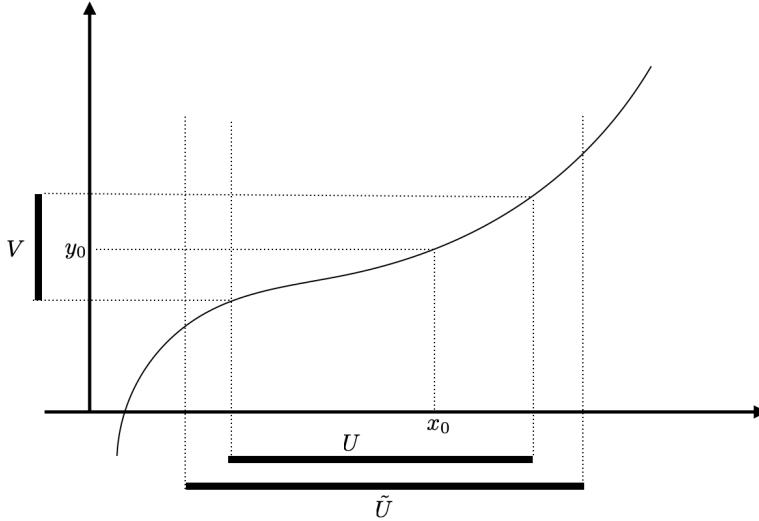


FIGURE 19.2 – Théorème d'inversion locale

Proposition 19.17. Soient E et F deux espaces vectoriels normés. Si E est séparable³, alors de toute suite (y_n) bornée dans F on peut extraire une suite $(y_{n'})$ qui converge au sens suivant :

$$\exists \varphi \in E', \quad T y_{n'} \xrightarrow{*} \varphi,$$

où T est définie par (19.1). Autrement dit, il existe $\varphi \in E'$ telle que

$$\psi(x, y_{n'}) \longrightarrow \langle \varphi, x \rangle \quad \forall x \in E.$$

Démonstration. Il existe une famille dénombrable $\{x_k\}_{k \in \mathbb{N}}$ dense dans E . On se propose de suivre le procédé d'extraction diagonale de Cantor.

1. Comme $\Psi(x_1, y_n)$ est bornée dans \mathbb{R} on peut extraire une suite $y_{j_1(n)}$ telle que $\Psi(x_1, y_{j_1(n)})$ converge.
2. Comme $\Psi(x_2, y_{j_1(n)})$ est bornée dans \mathbb{R} on peut extraire de $y_{j_1(n)}$ une suite $y_{j_1 \circ j_2(n)}$ telle que $\Psi(x_2, y_{j_1 \circ j_2(n)})$ converge.
3. Par récurrence, on construit une suite de sous-suites emboîtées $y_{j_1 \circ j_2 \circ \dots \circ j_k(n)}$ telle que $\Psi(x_k, y_{j_1 \circ j_2 \circ \dots \circ j_k(n)})$ converge, pour tout k .
4. On utilise à présent le procédé d'extraction diagonale : on pose $j(k) = j_1 \circ j_2 \circ \dots \circ j_k(k)$ (de telle sorte que j est strictement croissante), et on considère $y_{j(n)}$. Pour tout k , on remarque que $y_{j(n)}$, à partir du rang k , est aussi une suite extraite de $(y_{j_1 \circ j_2 \circ \dots \circ j_k(n)})$, de telle sorte que $\Psi(x_k, y_{j(n)})$ converge lorsque $n \rightarrow +\infty$.
5. On utilise pour finir la densité des x_k pour montrer que, pour tout $x \in H$, $\Psi(x, y_{j(n)})$ est une suite de Cauchy. Soit $\varepsilon > 0$, il existe (x_k) tel que $|x - x_k| < \varepsilon$. Comme $\Psi(x_k, y_{j(n)})$ est de Cauchy, il existe un N au-delà duquel $|\Psi(x_k, y_{j(p)}) - \Psi(x_k, y_{j(q)})| < \varepsilon$. Pour tous p, q supérieurs à N , on a donc

$$\begin{aligned} & |\Psi(x, y_{j(p)}) - \Psi(x, y_{j(q)})| \\ & \leq |\Psi(x, y_{j(p)}) - \Psi(x_k, y_{j(p)})| + |\Psi(x_k, y_{j(p)}) - \Psi(x_k, y_{j(q)})| + |\Psi(x_k, y_{j(q)}) - \Psi(x, y_{j(q)})| \\ & \leq (1 + 2C \|\Psi\|) \varepsilon, \end{aligned}$$

où $\|\Psi\|$ est la constante de continuité de Ψ (telle que $|\Psi(x, y)| \leq \|\Psi\| \|x\| \|y\|$), et C un majorant de $\|y_n\|$.

3. Il admet une famille dénombrable dense.

La suite $(y_{j(n)})$ est donc telle que $\Psi(x, y_{j(n)})$ converge, pour tout x , vers un réel noté $h(x)$. Cette limite est de façon linéaire par rapport à x , et de norme majorée par une constante fois la norme de x , il s'agit donc d'une forme linéaire continue sur F . \square

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 tels que la dualité est *séparante*, c'est à dire (on privilégie ici l'espace E) que

$$\Psi(x, y) = 0 \quad \forall x \implies y = 0.$$

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 :

$$\Psi(x, y) = 0 \quad \forall y \implies x = 0.$$

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) est à distance au moins 1 de $T(\ell^1)$.

Corollaire 19.18. 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 .

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 convergentes⁴.

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 19.19. Soit E un espace de Banach dont le dual est séparable. De toute suite bornée dans E on peut extraire une sous-suite qui converge⁵ 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 19.20. 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. \square

Espaces fonctionnels, mesures

On considère Ω un domaine de \mathbb{R}^d (qui peut être l'espace tout entier).

4. Autant dire qu'elle n'est pas commode à *utiliser* dans la vie de tous les jours.

5. Plus précisément son image par la surjection canonique de E dans E'' .

Le corollaire 19.19 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 $f_n = \mathbf{1}_{]n,n+1[}$ 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 identifié⁶. En particulier, le fait que l'on puisse (ou pas) extraire une sous-suite convergente de la suite définie précédemment n'est pas aisément 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 $f_n = \mathbf{1}_{]n,+\infty[}$. Si elle 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 converge (faible-*) 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 des fonctions 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 $f_n = n\mathbf{1}_{]0,1/n[}$ 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 $f_n = n\mathbf{1}_{]n,n+1/n[}$ converge vaguement ou faiblement vers 0, mais il n'en existe aucune sous-suite qui convergerait étroitement.

Exercice 19.1. 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-★ pour $L^1(\mathbb{R}^d)$ mis en dualité avec E . Que peut-on dire de la suite $f_n = n\mathbf{1}_{]n,n+1/n[}$ 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 $-\infty$.)

19.7 Espaces de Sobolev et traces : le point de vue de la modélisation

19.7.1 Interprétation des 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

6. 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 ??, page ??, donc indirectement de l'axiome du choix.

ses extrémités). On note (x_i) la configuration de référence⁷, avec $x_i = i/N$. La position de la masse i est notée $x_i + u_i$. L'énergie potentielle élastique du système est

$$E_N = \frac{1}{2} \sum_{i=0}^{N-1} k_N |x_{i+1} - x_i + u_{i+1} - u_i - \ell_N|^2.$$

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

$$E_N = \frac{1}{2} \sum_{i=0}^{N-1} k_N |u_{i+1} - u_i|^2,$$

que l'on peut aussi écrire

$$E_N = \frac{1}{2} \sum_{i=0}^{N-1} \ell_N (k_N \ell_N) \left| \frac{u_{i+1} - u_i}{\ell_N} \right|^2.$$

En choisissant $k_N = K/\ell_N$, on reconnaît 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 continûment différentiable $x \mapsto u(x)$), vers

$$\frac{K}{2} \int_0^1 |u'(x)|^2 dx,$$

ce qui permet d'interpréter le carré de la semi-norme H^1 comme l'énergie potentielle mécanique d'un système élastique 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

$$E_N^0 = \frac{1}{2} \sum_{i=1}^{N-1} k_N^0 |u_i|^2$$

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, pour obtenir des énergies ni infinie ni nulles à la limite, on doit faire tendre la raideur des ressorts "externes" vers 0, et celle des ressorts internes 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 noeuds d'un réseau cartésien couvrant le carré unité. 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 utilise en élasticité 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

7. Cette configuration minimise l'énergie potentielle dans le cas où la longueur au repos est inférieure à $1/\ell_N$.

ressorts de longueur au repos nulle, et de raideur k_N . Les masses sur le bord sont supposées 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

$$\begin{aligned} & \sum_{0 \leq i \leq N-1} \sum_{0 \leq j \leq N-1} \frac{1}{2} k_N \left(2\ell_N^2 + |u_{i+1,j} - u_{i,j}|^2 + |u_{i,j+1} - u_{i,j}|^2 \right) \\ &= K_N + \sum_{0 \leq i \leq N-1} \sum_{0 \leq j \leq N-1} \frac{1}{2} k_N \ell_N^2 \left(\left| \frac{u_{i+1,j} - u_{i,j}}{\ell_N} \right|^2 + \left| \frac{u_{i,j+1} - u_{i,j}}{\ell_N} \right|^2 \right) \end{aligned}$$

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 précédente correspond à l'énergie du réseau non déformé (qui est non nulle du fait que les longueurs au repos ont été prises égales à 0). On trouve donc ici une interprétation mécanique de la semi-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 le chapitre 2, page 37.

On peut (voir section 19.7.2 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 le cas d'un fluide).

19.7.2 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 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}^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 H/D 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 (espace des suites qui tendent vers 0) l'est dans ce cas : toute la richesse de l'espace est d'une certaine manière "au bord" (comportement en $n \mapsto +\infty$).

Considérons maintenant, pour $(\alpha_n) \in]0, +\infty[^{\mathbb{N}}$ donné, l'espace

$$H = \left\{ u = (u_n) \in \mathbb{R}^{\mathbb{N}}, u_0 = 0, \sum \alpha_n |u_{n+1} - u_n|^2 < +\infty \right\}, \quad (19.2)$$

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 19.21 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 $u_n = 1 + 1/2 + \dots + 1/n$, qui tend vers $+\infty$, est dans H . On peut construire aussi très simplement⁸ des suites qui tendent vers n'importe quelle valeur réelle en $+\infty$. Symétriquement, dans 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 parce qu'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 19.21. Soit H l'espace défini par (19.2), et H_0 l'adhérence de D (sous espace des suites nulles au delà d'un certain rang). On a

$$\sum \frac{1}{\alpha_n} < +\infty \implies H/H_0 \simeq \mathbb{R}, \quad \sum \frac{1}{\alpha_n} = +\infty \implies H/H_0 \simeq \{0\}.$$

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 $p < q$,

$$\begin{aligned} |u_q - u_p| &\leq \sum_{k=p}^{q-1} |u_{k+1} - u_k| = \sum_{k=p}^{q-1} \frac{1}{\sqrt{\alpha_n}} \sqrt{\alpha_n} |u_{k+1} - u_k| \\ &\leq \left(\sum_{k=p}^{q-1} \frac{1}{\alpha_n} \right)^{1/2} \left(\sum_{k=p}^{q-1} \alpha_n |u_{k+1} - u_k|^2 \right)^{1/2}, \end{aligned}$$

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

$$|u_n| = |u_n - u_{n-1} + u_{n-1} - \dots - u_0 + u_0| \leq \left(\sum \frac{1}{\sqrt{\alpha_n}} \right)^{1/2} \left(\sum \alpha_n |u_{n+1} - u_n|^2 \right)^{1/2} \leq \frac{1}{\alpha} \|u\|_H.$$

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

$$h_n = \sum_{k=1}^n (h_k - h_{k-1}) = q \sum_{k=1}^n \frac{1}{\alpha_{k-1}} \rightarrow \frac{q}{\alpha},$$

8. 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*.

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

$$\alpha_n(u_{n+1}^N - u_n^N) = q$$

constant pour n entre N et $M - 1$. On a donc

$$u_N = u_N^N = u_N^N - u_{N+1}^N + \cdots - u_{M-1}^N + u_M^N - u_M^N = q \sum_{n=N}^{M-1} \frac{1}{\alpha_n} = qr_{NM}.$$

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 petit, 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 . \square

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

$$H = \left\{ u = (u_n) \in \mathbb{R}^{\mathbb{N}}, u_0 = 0, \sum u_n^2 + \sum \alpha_n |u_{n+1} - u_n|^2 < +\infty \right\}. \quad (19.3)$$

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 certaine manière une trace (les suites de H sont de Cauchy d'après la partie différentielle de la norme), mais cette trace est nécessairement nulle du fait de la présence du terme ℓ^2 dans la norme.

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_2, \dots , mis bout à bout. On note $\alpha_n = 1/r_n$ la conductivité du fil n . Pour faciliter 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.

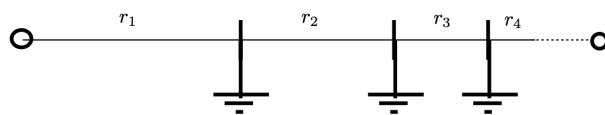


FIGURE 19.3 – Réseau linéaire semi-infini

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

$$\sum_{n=1}^N \alpha_n |u_n - u_{n-1}|^2 = \sum_{n=1}^N \frac{1}{r_n} |u_n - u_{n-1}|^2,$$

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

$$\sum_{n=1}^N \frac{1}{r_n} |u_n - u_{n-1}|^2 = \sum_{n=1}^N r_n |q_n|^2 = \underbrace{\sum_{n=1}^N r_n |q|^2}_{=R_N},$$

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 énergétiquement 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 (19.3). 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 (et donc nécessiterait de fournir une puissance infinie au système).

Remarque 19.22. 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 structurer 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 19.21, 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 (plus précisément un seul chemin vers l'infini, un seul *bout*), l'espace des traces ne peut donc être 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 électriques en forme d'échelle semie-infinie (voir figure 19.4), et en définissant l'espace de potentiels aux noeuds de ce réseaux qui correspondent à une puissance dissipée finie. On note $\alpha_n = 1/r_n$, et l'on définit

$$H = \left\{ u = (u_n^1, u_n^2), u_0^1 = u_0^2, \sum \alpha'_n |u_n^2 - u_n^1|^2 < +\infty, \sum \alpha_n |u_{n+1}^i - u_n^i|^2 < +\infty, i = 1, 2 \right\}$$

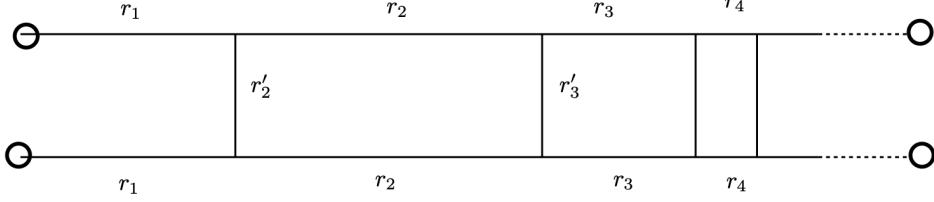


FIGURE 19.4 – Réseau semi-infini

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 transverses 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 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 pas 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.

Ces questions de traces peuvent également se poser pour des réseaux résistifs (voir chapitre ??, plus précisément la sous-section 2.10 pour le cadre des réseaux infinis). On peut par exemple identifier \mathbb{Z}^d à un réseau résistif infini (en considérant que tous les côtés ont la même résistance), et montrer que l'espace des traces est trivial pour $d = 1$ ou $d = 2$, et quasi-trivial (i.e. de dimension 1) pour les dimensions $d \geq 3$. La situation est plus riche dans le cas des arbres résistifs, la structure d'arbre elle-même assurant une certaine tolérance vis à vis des variations transverses, qui permet aux espaces de traces (si la rigidité longitudinale est suffisante pour assurer qu'il se passe quelque chose au bout) d'avoir plus de richesse que dans le cas du réseau cartésien. On se reportera à la section 3.7 pour une application de cette démarche au cas du poumon.

19.8 Introduction aux 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 Santambrogio⁹ ou Villani¹⁰ 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 ajouté 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 $\nu = T_\sharp^\varepsilon \rho$, 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 transformation :

$$T^\varepsilon \sharp \rho(x + \varepsilon w) = \frac{\rho(x)}{|i + \varepsilon \nabla w|}.$$

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

9. F. Santambrogio, *Optimal Transport for Applied Mathematicians*, Birkhäuser, NY, 2015.

10. C. Villani, "Topics in Optimal Transportation," Grad. Stud. Math., **58**, AMS, Providence, RI, 2003.

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 gradient¹¹ 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 $-\nabla \varphi$:

$$\partial_t \rho - \nabla \cdot (\rho \nabla \varphi) = 0.$$

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 Dirac¹²). 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 $x_1^0, \dots, x_N^0 \in \mathbb{R}^d$, 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)

$$\frac{dx_i}{dt} = -\nabla \varphi(x_i(t)), \quad x_i(0) = x_i^0.$$

Ce flot gradient est donc une généralisation macroscopique des flots gradients ponctuels dans l'espace euclidien.

Remarque 19.23. 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

11. On définit plus généralement la notion de sous-différentiel, qui correspond à l'ensemble des vecteurs v tels que

$$\Psi(\rho) + \varepsilon \int v \cdot w \rho(x) dx \leq \Psi(T^\varepsilon \sharp \rho) + o(\varepsilon),$$

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.

12. 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.

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

$$T^\varepsilon \sharp \rho(x + \varepsilon w) = \frac{\rho(x)}{|i + \varepsilon \nabla w|} = \rho(x)(1 - \varepsilon \nabla \cdot w + o(\varepsilon)).$$

On obtient donc

$$\begin{aligned}\Psi(T^\varepsilon \sharp \rho) &= \int \frac{\rho(x)(1 - \varepsilon \nabla \cdot w + o(\varepsilon))}{\rho(1 - \varepsilon \nabla \cdot w + o(\varepsilon))} \rho(x) dx \\ &= \int (\varphi(\rho) - \varepsilon \rho \nabla \cdot w \varphi'(\rho) + o(\varepsilon)) (1 + \varepsilon \nabla \cdot w + o(\varepsilon)) dx \\ &= \Psi(\rho) + \varepsilon \int (\varphi(\rho) - \rho \varphi'(\rho)) \nabla \cdot w dx + o(\varepsilon) \\ &= \Psi(\rho) + \varepsilon \int w \cdot \nabla (\rho \varphi'(\rho) - \varphi(\rho)) dx + o(\varepsilon) \\ &= \Psi(\rho) + \varepsilon \int w \cdot (\rho \nabla \varphi'(\rho) + \varphi'(\rho) \nabla \rho - \varphi'(\rho) \nabla \rho) + o(\varepsilon) \\ &= \Psi(\rho) + \varepsilon \int w \cdot \nabla \varphi'(\rho) \rho dx + o(\varepsilon),\end{aligned}$$

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.

19.9 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

$$\nabla \cdot u = \begin{pmatrix} \frac{\partial}{\partial x_1} \\ \frac{\partial}{\partial x_2} \end{pmatrix} \cdot \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \frac{\partial u_1}{\partial x_1} + \frac{\partial u_2}{\partial x_2}.$$

Soit $u = (u_1, u_2)^T$ un champ de vecteur, son gradient est la matrice

$$\nabla u = \begin{pmatrix} \frac{\partial u_1}{\partial x_1} & \frac{\partial u_1}{\partial x_2} \\ \frac{\partial u_2}{\partial x_1} & \frac{\partial u_2}{\partial x_2} \end{pmatrix}$$

Pour tout vecteur n , on a

$$\nabla u \cdot n = \begin{pmatrix} \frac{\partial u_1}{\partial x_1} & \frac{\partial u_1}{\partial x_2} \\ \frac{\partial u_2}{\partial x_1} & \frac{\partial u_2}{\partial x_2} \end{pmatrix} \begin{pmatrix} n_1 \\ n_2 \end{pmatrix} = \begin{pmatrix} \frac{\partial u_1}{\partial x_1} n_1 + \frac{\partial u_1}{\partial x_2} n_2 \\ \frac{\partial u_2}{\partial x_1} n_1 + \frac{\partial u_2}{\partial x_2} n_2 \end{pmatrix},$$

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 unitaire¹³, 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

$$A : B = \sum_{i,j} a_{ij} b_{ij}.$$

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 champs de vecteurs

$$\nabla u : \nabla v = \begin{pmatrix} \frac{\partial u_1}{\partial x_1} & \frac{\partial u_1}{\partial x_2} \\ \frac{\partial u_2}{\partial x_1} & \frac{\partial u_2}{\partial x_2} \end{pmatrix} : \begin{pmatrix} \frac{\partial v_1}{\partial x_1} & \frac{\partial v_1}{\partial x_2} \\ \frac{\partial v_2}{\partial x_1} & \frac{\partial v_2}{\partial x_2} \end{pmatrix} = \sum_{i=1}^2 \sum_{j=1}^2 \frac{\partial u_i}{\partial x_j} \frac{\partial v_i}{\partial x_j}.$$

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 divergence de la ligne de la matrice correspondant à cette composante.

$$\nabla \cdot \sigma = \nabla \cdot \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} = \begin{pmatrix} \frac{\partial \sigma_{11}}{\partial x_1} + \frac{\partial \sigma_{12}}{\partial x_2} \\ \frac{\partial \sigma_{21}}{\partial x_1} + \frac{\partial \sigma_{22}}{\partial x_2} \end{pmatrix}$$

Soit $u = (u_1, u_2)^T$ un champ de vecteur, on note $u \otimes u$ la matrice $(u_i u_j)_{i,j}$.

13. Cette hypothèse reflète le caractère assez peu naturel de cette notation. C'est un peu comme si, pour une fonction $x \mapsto f(x)$, 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 sortante en ce 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 . Cette notion n'a rien à voir avec la notation couramment utilisée pour désigner la dérivée normale.

Si $\nabla \cdot u = 0$, on a

$$\nabla \cdot (u \otimes u) = (u \cdot \nabla) u = \begin{pmatrix} u_1 \frac{\partial u_1}{\partial x_1} + u_2 \frac{\partial u_1}{\partial x_2} \\ u_1 \frac{\partial u_2}{\partial x_1} + u_2 \frac{\partial u_2}{\partial x_2} \end{pmatrix}$$

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

$$\nabla \cdot (\nabla u + {}^t \nabla u) = \nabla \cdot \nabla u = \Delta u.$$

Intégration par parties, formule de Green Ostrogradski

Soit v un champ de vecteurs. on a

$$\int_{\Omega} \nabla \cdot v = \int_{\Gamma} v \cdot n \quad (19.4)$$

Soit σ un champ de matrices. on a

$$\int_{\Omega} \nabla \cdot \sigma = \int_{\Gamma} \sigma \cdot n \quad (19.5)$$

Soit q un champ scalaire. On a

$$\int_{\Omega} \nabla q = \int_{\Gamma} q n. \quad (19.6)$$

Soit v un champ de vecteurs, et q un champ scalaire. On a

$$\int_{\Omega} q \nabla \cdot u + \int_{\Omega} u \cdot \nabla q = \int_{\Gamma} q u \cdot n. \quad (19.7)$$

Soient u et v des champs scalaires. On a

$$\int_{\Omega} v \Delta u = \int_{\Omega} \nabla u \cdot \nabla v + \int_{\Gamma} v \frac{\partial u}{\partial n}, \quad (19.8)$$

où n est la normale sortante au domaine.

Soit u un champ de vecteurs, et q un champ scalaire.

$$\int_{\Omega} q \nabla \cdot u + \int_{\Omega} u \cdot \nabla q = \int_{\Gamma} q u \cdot n. \quad (19.9)$$

Soient u et v des champs de vecteurs. On a

$$\int_{\Omega} \Delta u \cdot v + \int_{\Omega} \nabla u : \nabla v = \int_{\Gamma} v \cdot \frac{\partial u}{\partial n}. \quad (19.10)$$

Si en outre $\nabla \cdot u = 0$, on a

$$0 + \int_{\Omega} {}^t \nabla u : \nabla v = \int_{\Gamma} v \cdot ({}^t \nabla u \cdot n) \quad (19.11)$$

En conséquence, si $\nabla \cdot u = 0$, alors

$$\int_{\Omega} \Delta u \cdot v + \int_{\Omega} \nabla u : (\nabla v + {}^t \nabla v) = \int_{\Gamma} v \cdot (\nabla u + {}^t \nabla u) \cdot n. \quad (19.12)$$

Pour tous champs vectoriels u et v , on a

$$\int_{\Omega} \nabla u : {}^t \nabla v = \int_{\Omega} {}^t \nabla u : \nabla v, \quad (19.13)$$

de telle sorte que

$$\int_{\Omega} \nabla u : (\nabla v + {}^t \nabla v) = \frac{1}{2} \int_{\Omega} (\nabla u + {}^t \nabla u) : (\nabla v + {}^t \nabla v) \quad (19.14)$$

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

$$\frac{d}{dt} \int_{\omega(t)} F(x, t) = \int_{\omega(t)} \frac{\partial F}{\partial t}(x, t) + \int_{\partial\omega(t)} F(x, t) u \cdot n. \quad (19.15)$$

Proposition 19.24. Soient u et v deux champs de vecteurs réguliers définis sur Ω . On suppose que u est à divergence nulle. On a alors

$$0 = - \int_{\omega} {}^t \nabla u : \nabla v + \int_{\partial\omega} v \cdot ({}^t \nabla u \cdot n)$$

Démonstration. On écrit

$$\begin{aligned} \int_{\partial\omega} v \cdot ({}^t \nabla u \cdot n) &= \int_{\partial\omega} n \cdot (\nabla u \cdot v) = \int_{\omega} \nabla \cdot (\nabla u \cdot v) \\ &= \sum_i \partial_i \sum_j v_j \partial_j u_i = \sum_i \sum_j \partial_i v_j \partial_j u_i + \sum_j v_j \partial_j \sum_i \partial_i u_i. \end{aligned}$$

Le second terme ci-dessus est nul car u est à divergence nulle, d'où l'on déduit l'identité annoncée. \square

Proposition 19.25. 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

$$\begin{aligned} \int_{\Gamma} (\nabla u \cdot v) \cdot n &= \int_{\Omega} \nabla \cdot (\nabla u \cdot v) \\ &= \int_{\Omega} \sum_i \partial_i \sum_j v_j \partial_j u_i \\ &= \int_{\Omega} \sum_i \sum_j ((\partial_i \partial_j u_i) v_j + \partial_j u_i \partial_i v_j) \\ &= \int_{\Omega} v (\nabla \nabla \cdot u) + \int_{\Omega} \nabla u : {}^t \nabla v \\ &= \int_{\Omega} (\nabla \cdot u) v \cdot n - \int_{\Omega} (\nabla \cdot u) (\nabla \cdot v) + \int_{\Omega} \nabla u : {}^t \nabla v \end{aligned}$$

19.10 Cercles de Gershgorin

Definition 19.26. Une matrice $A = (a_{ij}) \in \mathcal{M}_n(\mathbb{C})$ est dite à diagonale strictement dominante si

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}| \quad \forall i = 1, \dots, n.$$

Proposition 19.27. (Gershgorin)

Soit $A = (a_{ij}) \in \mathcal{M}_n(\mathbb{C})$. Soit $\text{Sp}(A)$ l'ensemble des valeurs propres de A . On a

$$\text{Sp}(A) \subset \bigcup_{i=1}^n D(a_{ii}, r_i), \quad r_i = \sum_{j \neq i} |a_{ij}|,$$

où $D(a, r) \subset \mathbb{C}^2$ désigne le disque fermé de centre a et de rayon r .

19.11 Spectre du Laplacien discret

La matrice

$$A = \begin{pmatrix} 2 & -1 & 0 & \cdot & \cdot & 0 \\ -1 & 2 & -1 & 0 & \cdot & \cdot \\ 0 & -1 & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & & \cdot & 2 & -1 & \\ 0 & \cdot & \cdot & 0 & -1 & 2 \end{pmatrix} \in \mathcal{M}_{N-1}(\mathbb{R}) \quad (19.16)$$

possède $N - 1$ valeurs propres distinctes

$$\lambda_k = 4 \sin^2 \left(\frac{k\pi}{2N} \right), \quad k = 1, \dots, N - 1. \quad (19.17)$$

Le vecteur propre associé à la valeur propre λ_k s'écrit

$$u_k = {}^t \left(\sin \left(\frac{k\pi}{N} \right), \sin \left(\frac{2k\pi}{N} \right), \dots, \sin \left(\frac{(N-1)k\pi}{N} \right) \right).$$

19.12 Théorème spectral généralisé

Proposition 19.28. Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice symétrique, et $M \in \mathcal{M}_N(\mathbb{R})$ une matrice symétrique définie positive. Le problème aux valeurs propres généralisé

$$Au = \lambda Mu$$

admet N valeurs propres généralisées $(\lambda_j)_{1 \leq j \leq N}$ réelles, associées à N vecteurs propres généralisés $(w_j)_{1 \leq j \leq N}$ qui forment une base M -orthogonale, i.e. $\langle Mw_i | w_j \rangle = \delta_{ij}$ (symbole de Kronecker).

Démonstration. On introduit la racine carrée ¹⁴ de M , notée $M^{1/2}$, et l'on effectue le changement d'inconnue $v = M^{1/2}u$. Le problème en v s'écrit

$$AM^{-1/2}v = \lambda M^{1/2}v \iff M^{-1/2}AM^{-1/2}v = \lambda v.$$

On se ramène ainsi à un problèmes aux valeurs propres standard impliquant la matrice symétrique $M^{-1/2}AM^{-1/2}$, il existe donc une famille (λ_j) de réels, et une famille de vecteurs propres associés

(v_j) qui constitue une base orthogonale pour le produit scalaire euclidien standard. Les $w_j = M^{-1/2}$ forment donc une famille de vecteurs propres généralisés solution du problème de départ, et l'on a

$$\langle w_j | w_k \rangle_M = \langle Mw_j | w_k \rangle = \left\langle MM^{-1/2}v_j | M^{-1/2}v_k \right\rangle = \langle v_j | v_k \rangle = 0$$

dès que $j \neq k$. La famille est donc bien M -orthogonale, et peut être rendue M -orthonormale par normalisation. \square

19.13 Entiers p -adiques, espaces ultra-métriques

Cette section est consacrée à la construction du complété de \mathbb{N} pour une certaine métrique, appelée p -adique. Nous détaillons cette construction pour le cas $p = 2$, qui aboutit à l'espace dit des entiers 2-adiques, noté \mathbb{Z}_2 , dont on présente ci-dessous certaines propriétés.

Distance 2-adique sur \mathbb{N}

Definition 19.29. (Valuation et valeur absolue 2-adiques)

Tout $a \in \mathbb{Z}$ non nul peut s'écrire de façon unique $a'2^k$, où a' est impair. On appelle $v_2(a) = k$ la *valuation* 2-adique de a , et l'on définit la valeur absolue 2-adique

$$|a|_2 = 2^{-v_2(a)}.$$

On pose $v_2(0) = +\infty$, et $|0|_2 = 0$.

On se place maintenant sur $X = \mathbb{N}$ l'ensemble des entiers naturels, et l'on définit

$$(a, b) \in \mathbb{N}^2 \mapsto d(a, b) = |b - a|_2.$$

On peut donner une expression équivalente de cette définition, en considérant que tout entier naturel a s'écrit (écriture dyadique)

$$a = \sum_{k=0}^{+\infty} a_n 2^n, \quad a_n \in \{0, 1\} \quad \forall n, \quad a_n \text{ nul au delà d'un certain rang.} \quad (19.18)$$

Deux nombres a et b peuvent ainsi être écrits en base 2 sous la forme de suites finies (a_n) et (b_n) de 0 et de 1, et l'on a $d(a, b) = 2^{-n_{ab}}$, où

$$n_{ab} = \min \{n, a_n \neq b_n\}, \quad (19.19)$$

pris égal à $+\infty$ si les bits de a et b s'identifient (i.e. si $a = b$). En effet, par définition de n_{ab} , on a

$$a - b = 2^{n_{ab}} c,$$

où $c \in \mathbb{Z}$ est impair.

Proposition 19.30. L'application $d(\cdot, \cdot)$ est une distance sur \mathbb{N} , qui est *ultramétrique*, c'est-à-dire qu'elle vérifie l'inégalité triangulaire renforcée :

$$d(a, b) \leq \max(d(a, c), d(b, c)) \quad \forall a, b, c \in X.$$

14. La matrice M étant symétrique définie positive, elle s'écrit $M = UDU^{-1}$, où D est une matrice diagonale aux coefficients > 0 . On définit la racine carrée par $M^{1/2} = U D^{1/2} U^{-1}$, elle est par construction symétrique définie positive.

Démonstration. La séparation est immédiate par hypothèse, et l'on a bien $d(a, b) = d(b, a)$. Pour l'inégalité triangulaire renforcée, on utilise la formulation (19.19). Pour tous entiers a, b, c , on considère les écriture dyadiques associées. Les $n_{ab} - 1$ premiers bits de a et b s'identifient par définition, ainsi que les $n_{bc} - 1$ premiers bits de b et c . Les premiers bits de a et c s'identifient donc au moins sur les $\min(n_{ab}, n_{bc}) - 1$ indices, d'où

$$n_{ac} \geq \min(n_{ab}, n_{bc}),$$

et ainsi

$$d(a, c) = 2^{-n_{ac}} \leq \max(2^{-n_{ab}}, 2^{-n_{bc}}) = \max(d(a, b), d(b, c)),$$

qui est l'inégalité ultramétrique annoncée, qui entraîne l'inégalité triangulaire usuelle. \square

Cette distance munit \mathbb{N} d'une distance aux propriétés non standard, dont certaines sont énoncées dans la suite (voir proposition 19.34 et les suivantes).

Le diamètre de (\mathbb{N}, d) est 1, et ce diamètre est atteint en de multiples couples : tout entier impair est en particulier diamétralement opposé à 0, ainsi qu'à tout entier pair. La sphère unité (sphère de centre 0 et de rayon 1) est l'ensemble des nombres impairs. De façon générale, la sphère de centre 0 et de rayon 2^{-k} est l'ensemble des nombres du type $2^k b$, avec b impair. Ces sphères de centre 0 et de rayons 1, $1/2$, $1/4$, ..., et 0, réalisent une partition de \mathbb{N} (comme un emboîtement infini de poupées russes).

La suite (2^n) tend vers 0, ainsi que toute suite telle que l'exposant de 2 dans la décomposition en facteurs premiers des termes tend vers $+\infty$.

Nous terminons cette section par une propriété négative sur (\mathbb{N}, d) , qui justifie la démarche de complétion qui suit.

Proposition 19.31. L'espace ultramétrique (\mathbb{N}, d) n'est pas complet.

Démonstration. Considérons la suite $(2^n - 1)$. Cette suite est de Cauchy pour $d(\cdot, \cdot)$. Si elle converge vers $a \in \mathbb{N}$, alors 2^n converge vers $a + 1$, d'où $a + 1 = 0$, équation qui n'admet pas de solution dans \mathbb{N} . \square

Remarque 19.32. Noter que l'on aurait pu choisir comme espace de départ $X = \mathbb{Z}$ au lieu de \mathbb{N} , auquel cas le contre-exemple ci-dessus n'en est plus un, puisque $a + 1$ admet une solution dans \mathbb{Z} . Pour se convaincre qu'il manque plus à (\mathbb{N}, d) que les nombres négatifs pour être complet, on peut considérer par exemple la suite

$$a_N = \sum_{n=0}^{2N} a_n 2^n,$$

avec $a_n = n + 1 \bmod 2$ (alternance périodique de 1 et de 0). On a

$$a_N = \sum_{n=0}^N 2^{2n} = \frac{1 - 4^{N+1}}{1 - 4} = \frac{4^{N+1} - 1}{3}.$$

Cette suite est de Cauchy (comme toute série partielle de ce type). S'il existe $a \in \mathbb{N}$ tel que a_N converge vers a , alors on a

$$\left| \frac{4^{N+1} - 1}{3} - a \right|_2 \rightarrow 0 \Rightarrow |4^{N+1} - 1 - 3a|_2,$$

d'où, comme précédemment, $3a + 1 = 0$, équation qui n'admet pas de solution dans \mathbb{N} , ni dans \mathbb{Z} .

Complété de \mathbb{N}

Definition 19.33. Le complété de \mathbb{N} (voir théorème ??) pour la distance définie ci-dessus est appelé ensemble des entiers 2-adiques. Il est noté \mathbb{Z}_2 .

L'écriture (19.18), qui permet de représenter tout entier comme une suite finie de 0 ou de 1 (écriture en base 2), permet de se faire une meilleure idée de \mathbb{Z}_2 , et de préciser à quoi correspondent certains de ses éléments. Considérons une suite (a^k) dans \mathbb{N} , de Cauchy pour $d(\cdot, \cdot)$. On peut écrire les termes

$$a^k = (a_0^k, a_1^k, a_2^k, \dots)$$

Si une suite est de Cauchy alors, comme dans le cas des réels en écriture décimale (voir la démonstration de la proposition ??), chacune des suites $(a_n^k)_k$ finit par se stabiliser à une valeur a_n . On peut donc représenter la limite par une suite (infinie) de 0 et de 1. On écrira le nombre correspondant somme somme d'une série, ou en écriture flottante, avec la convention de placer les bits *avant* la virgule :

$$a = \sum_{n=0}^{+\infty} a_n 2^n, \quad \text{ou } a = \dots a_3 a_2 a_1 a_0, 0.$$

Considérons par exemple le nombre $\dots 111, 0$. On a

$$\dots 111, 0 = \lim_{N \rightarrow \infty} \left(\sum_{n=0}^N 2^n \right) = \lim_{N \rightarrow \infty} \left(\frac{1 - 2^{N+1}}{1 - 2} \right) = -1 - \lim_{N \rightarrow \infty} 2^{N+1}$$

qui tend donc selon les règles de calcul usuel vers un nombre qui n'est pas dans l'espace de départ, et qu'il est tentant de noter -1 .

La métrique induite sur \mathbb{Z}_2 est définie essentiellement de la même manière que sur \mathbb{N} . Pour deux éléments a et b de \mathbb{Z}_2 , si l'on note n_{ab} le plus petit indice pour lequel les bits diffèrent ($n_{ab} = +\infty$ si $a = b$), la distance entre a et b est simplement définie par $d(a, b) = 2^{-n_{ab}}$.

La distance sur \mathbb{Z}_2 définie ci-dessus hérite des propriétés ultramétriques de la distance de départ sur \mathbb{N} . L'espace \mathbb{Z}_2 possède donc des propriétés propres aux espaces ultramétriques, telles que celles énoncées ci-après.

Proposition 19.34. Soit (X, d) un espace ultramétrique. Tout point d'une boule (ouverte ou fermée) est centre de cette boule.

Démonstration. Soient $x \in X$ et $r \in \mathbb{R}_+$. On considère $x' \in \overline{B}(x, r)$, i.e. tel que $d(x, x') \leq r$. Pour tout $y \in \overline{B}(x, r)$, on a

$$d(y, x') \leq \max(d(y, x), d(x, x')) \leq r, \quad \text{d'où } y \in B(x', r).$$

On montre de la même manière que tout $y \in \overline{B}(x', r)$ est à distance de x inférieure ou égale à r , d'où l'identité de $\overline{B}(x', r)$. La démonstration est identique pour une boule ouverte. \square

Corollaire 19.35. Soit (X, d) un espace ultramétrique. L'intersection entre deux boules est soit vide soit l'une des deux boules. Si les rayons sont les mêmes, deux boules sont donc soit disjointes, soit identiques.

Démonstration: Soit $B(x, r)$ et $B(x', r')$. Si y appartient aux deux boules, il est aussi centre de ces deux boules d'après ce qui précède, la plus petite est donc incluse dans la plus grande, avec égalité si les rayons sont les mêmes.

Proposition 19.36. Soit (X, d) un espace ultramétrique. Toute boule ouverte est un fermé.

Démonstration. Si $r = 0$ (alors $B(x, r) = \emptyset$) ou si $B(x, r) = X$, c'est immédiat. Sinon, pour tout $y \in B(x, r)^c$, $B(x, r) \cap B(y, r) = \emptyset$ d'après la proposition précédente, d'où $B(y, r) \subset B(x, r)^c$. Le complémentaire de $B(x, r)$ est donc un ouvert, la boule ouverte elle-même est donc un fermé. \square

Proposition 19.37. Un espace ultramétrique est totalement discontinu, au sens où chaque point s'identifie à sa propre composante connexe.

Démonstration. Soient x et $y \neq x$ dans X , et $0 < r < d(x, y)$. La boule ouverte $B(x, r)$ et son complémentaire (qui est aussi un ouvert) réalisent une partition de X en 2 ouverts, x et y ne peuvent donc appartenir à la même composante connexe. \square

Proposition 19.38. Soit (X, d) un espace ultramétrique. Tout triangle dans X est isocèle (avec deux grands côtés et un petit côté).

Proposition 19.39. On considère trois points dans X , et l'on note ℓ_1, ℓ_2 , et ℓ_3 les longueurs des côtés. On a $\ell_1 \leq \max(\ell_2, \ell_3)$, et les mêmes propriétés obtenues en permutant les indices. Si les longueurs sont distinctes deux à deux, on a par exemple $\ell_1 < \ell_2 < \ell_3$, qui invalide $\ell_3 \leq \max(\ell_1, \ell_2)$. Deux des longueurs au moins sont donc identiques, par exemple $\ell_1 = \ell_2$. Le troisième côté est de longueur $\ell_3 \leq \max(\ell_1, \ell_2) = \ell_1$.

Proposition 19.40. Soit (X, d) un espace ultramétrique. Une suite (x_n) est de Cauchy si et seulement si

$$\lim_{n \rightarrow +\infty} d(x_{n+1}, x_n) = 0.$$

Addition sur \mathbb{Z}_2 . On peut définir une addition sur \mathbb{Z}_2 , qui est “l’addition de l’écolier” en partant de la droite dans l’écriture 2-adique ci-dessus. On considère deux entiers dyadiques a et b , et l’on cherche à construire la somme $c = a + b$, qui étende la somme sur \mathbb{N} . Si $a_0 + b_0 = 0$ ou 1 , on affecte à c_0 cette valeur, et on passe au rang suivant. Si la somme vaut 2 on pose $c_0 = 0$, et l’on garde une retenue de 1 pour le rang suivant. Au rang 1 on se retrouve dans la même situation, avec éventuellement une retenue de 1 en plus. La somme peut donc maintenant valoir 3 . Si c’est le cas on pose $c_1 = 1$ et on garde 1 de retenue pour le rang suivant, etc …

De façon assez frappante, le fait de prendre le complété de \mathbb{N} pour la distance choisie, qui est une démarche purement topologique, conduit à espace qui possède une structure algébrique que l’espace de départ n’avait pas.

Proposition 19.41. L’espace \mathbb{Z}_2 muni de l’addition définie ci-dessus est un groupe additif.

Proposition 19.42. L’espace \mathbb{Z}_2 n’est pas dénombrable.

Démonstration. Nous avons vu que tout élément de \mathbb{Z}_2 peut se représenter de façon unique par une suite infinie de 0 ou de 1 . On peut donc identifier (d’un point de vue ensembliste) \mathbb{Z}_2 à l’ensemble des parties de \mathbb{N} , en considérant la suite (a_n) comme la fonction indicatrice d’une partie de \mathbb{N} . L’ensemble \mathbb{Z}_2 n’est donc pas dénombrable. \square

Definition 19.43. (Ordre lexicographique)

On peut définir sur \mathbb{Z}_2 un ordre *lexicographique*, en considérant, pour deux éléments différents $\dots a_2 a_1 a_0, 0$ et $\dots b_2 b_1 b_0, 0$ le plus petit indice n pour lequel les deux diffèrent, et poser $a < b$ si $a_n < b_n$.

Noter que l’ordre défini ci-dessus est très différent de l’ordre usuel. Comme $\mathbb{Z} \subset \mathbb{Z}_2$, on peut comparer de ce point de vue deux éléments de \mathbb{Z} , on retrouve certaines propriétés usuelles du type $1 < 3, 1 < 5$, mais aussi des choses plus déroutantes, comme $5 < 3$ et, plus globalement, s

$$0 \leq a \leq -1 \quad \forall a \in \mathbb{Z},$$

ce qui permet d’écrire

$$\mathbb{Z} = [0, -1].$$

Proposition 19.44. L’espace \mathbb{Z}_2 est compact.

Démonstration. Considérons une suite (a^k) dans \mathbb{Z}_2 . On considère la suite $(a_0^k)_k$ de 0 et de 1 . Cette suite visite une infinité de fois 0 ou 1 (ou les deux). On prend a_0 égal à une valeur visitée une infinité de fois, et l’on extrait la sous-suite $(a^{\varphi_0(k)})$ correspondante. On procède de même avec $(a_1^{\varphi_0(k)})$, pour

extraire une suite $(a^{\varphi_0 \circ \varphi_1(k)})$. On extrait ainsi des sous-suites emboitées les unes dans les autres. On définit maintenant (selon le processus d'extraction diagonale dit de Cantor)

$$\varphi(k) = \varphi_0 \circ \varphi_1 \circ \cdots \circ \varphi_k(k).$$

La suite ainsi construite est telle que $a_n^{\varphi(k)}$ a une valeur constante $a_n \in \{0, 1\}$ au-delà d'un certain rang (pour tout $k \geq n$), on a donc convergence de cette suite extraite vers $a = \dots a_3 a_2 a_1 a_0, 0$. \square

Système projectif, limite projective

Un *système projectif* désigne une famille (X_n) d'ensemble muni d'une famille d'applications $(f_n^m)_{n \leq m}$, avec $f_n^m : X_m \rightarrow E_n$, qui vérifient les propriétés suivantes

- (1) L'application f_n^n est l'identité sur X_n
- (2) Pour tous $n \leq m \leq q$, on a $f_n^m \circ f_m^q = f_n^q$.

On appelle limite projective l'ensemble des éléments du produit infini $X_0 \times X_1 \times \dots$ dont les projections sont compatibles avec les f_i^j au sens suivant :

$$\varprojlim \left((X_n), (f_i^j) \right) = \{ x = (x_0, x_1, x_2, \dots) \in X_0 \times X_1 \times \dots, f_n^m(x_m) = x_n \quad \forall i \leq j \}.$$

Pour tous $n \leq m$, on peut définir canoniquement¹⁵ une surjection de $\mathbb{Z}/2^m\mathbb{Z}$ dans $\mathbb{Z}/2^n\mathbb{Z}$. On note f_n^m cette surjection. On peut identifier ainsi \mathbb{Z}_2 à la limite projective du système $((\mathbb{Z}/2^n\mathbb{Z}), (f_n^m))$.

Représentation des arbres dyadiques, application au poumon humain

Le poumon humain se présente comme un arbre constitué de bronches (appelée bronchioles pour les plus petites), structuré de façon dyadique : la trachée se divise en deux, chacune des branches filles se divise elle-même en deux, etc... Pour l'arbre respiratoire d'un adulte, le nombre de bifurcations est de l'ordre de 23, soit autour de $2^{23} \approx 8 \times 10^8$ bronchioles terminales, dont on appelle *feuilles* les extrémités libres. Pour diverses raisons (construction d'un modèle homogénéisé du parenchyme, construction d'un opérateur de la ventilation, qui à un champ de pression aux feuilles associe un champ de flux, etc ...), il peut être intéressant *d'extrapoler* cet arbre vers un nombre de générations infini. La notion de feuille est alors remplacée par celle de *bout*, on parle de l'*espace des bouts*, chacun de ces bouts correspondant à un chemin centrifuge s'éloignant de la racine (entrée de la trachée). Il est naturel de coder chacun de ces chemins par une suite infinie de 0 et de 1, par exemple selon la convention par exemple (on se figurera une représentation avec la racine en haut, comme sur la figure 19.5) : partant de la racine, si l'on part à gauche, on prend $a_0 = 0$, et $a_0 = 1$ si c'est à droite. On encode le "choix" à chaque étape par un nombre entre 0 et 1.

La figure 19.5 représente la numérotation obtenue à chaque génération, où l'on a représenté chaque mot $(a_N \dots, a_0)$ par sa représentation entière $a = \sum a_n$.

On notera que cette numérotation ne correspond pas du tout à l'énumération linéaire 0, 1, Cette numérotation présente un énorme avantage par rapport à l'énumération linéaire, en cela qu'elle respecte la structure de l'arbre. Plus précisément, considérons deux feuilles de la 4-ième génération, d'indices a et b . Leur proximité *vis à vis de l'arbre*, qui correspondrait à un degré de parenté s'il s'agissait d'un arbre généalogique, ne dépend que de $|a - b|$, contrairement à ce qui se passerait pour la numérotation linéaire, comme on peut s'en convaincre facilement. Par exemple deux feuilles dont la différence d'indices est impaire appartiennent nécessairement à des lobes différents. Si la différence est divisible par 2, et pas par 4, les deux point appartiennent nécessairement tous deux à l'un des 4 sous-arbres issus de la génération 2, etc...

¹⁵. On peut considérer la relation d'équivalence sur $\mathbb{Z}/2^m\mathbb{Z}$ définie par $z \mathcal{R} z' \iff z \equiv z' [2^n]$. L'espace quotient s'identifie à $\mathbb{Z}/2^n\mathbb{Z}$.

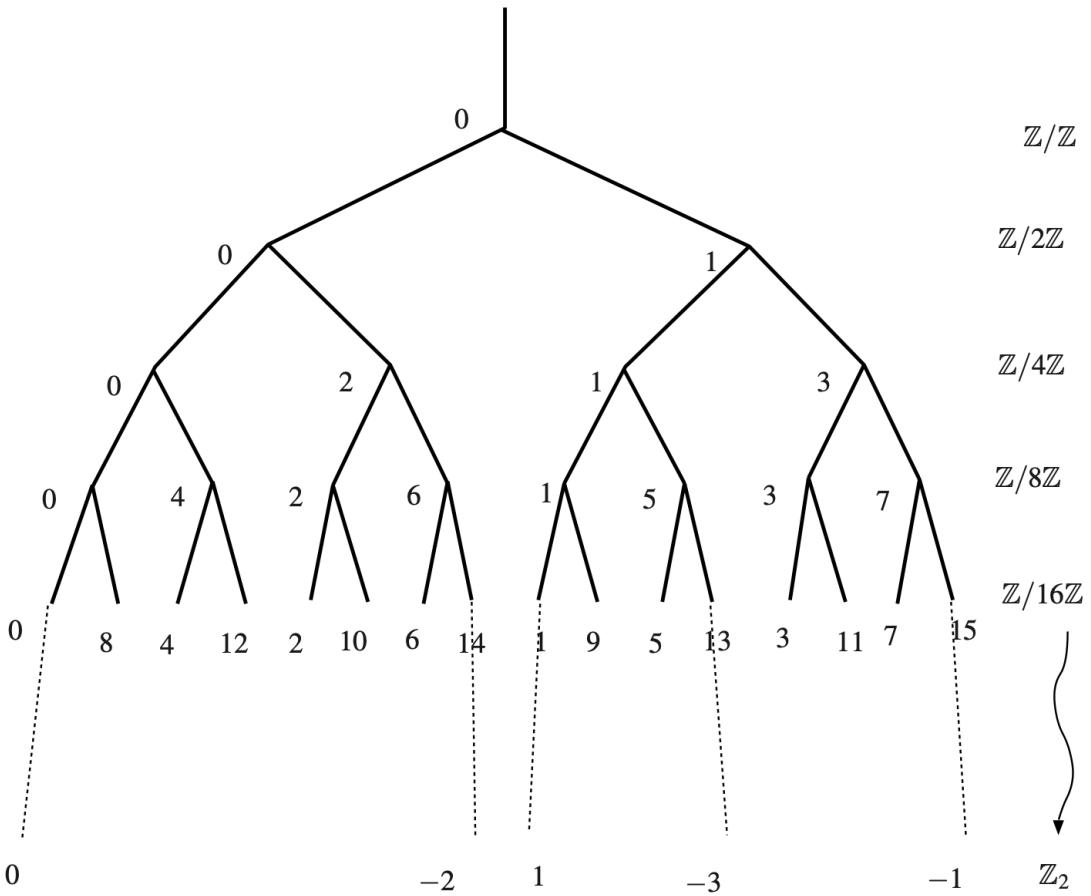


FIGURE 19.5 – Numérotation 2-adique

Plus précisément, si l'on considère l'arbre à 4 générations représenté sur la figure, qui présente $2^4 = 16$ feuilles, on peut identifier la métrique dyadique sur l'ensemble des feuilles (identifié à $\llbracket 0, 15 \rrbracket$ ou $\mathbb{Z}/2^4\mathbb{Z}$) avec la métrique du plus court chemin au travers de l'arbre considéré comme un graphe pondéré. Pour retrouver exactement la distance 2-adique, on peut vérifier qu'il suffit de considérer que les arêtes des deux dernières générations (3 et 4) ont pour longueur $1/16$, longueur $1/8$ pour la génération 2 (4 branches), et $1/4$ pour la génération 1.

Si l'on considère maintenant le poumon infini, on voit apparaître des indices déjà identifiés pour certains bouts. Par exemple le bout le plus à gauche est clairement 0. Pour le bout correspondant au chemin le plus à droite, on retrouve notre $\dots 1111, 0 = -1$. Noter que -1 est le plus “grand” élément de \mathbb{Z}_2 pour la relation d’ordre lexicographique (voir définition 19.43), et 0 le plus petit. De façon générale, la représentation de la figure 19.5 correspond à cet ordre (croissant de la gauche vers la droite).

On notera que, pour tout sous-arbre infini, le bout le plus à gauche est un entier positif, et le bout le plus à droite un entier négatif.

Nombres entiers p -adiques

On peut généraliser cette approche en remplaçant 2 par un nombre p premier quelconque : tout $a \in \mathbb{Z}$ non nul peut s'écrire de façon unique $a'p^k$, où $a' \in \mathbb{Z}$ n'est pas divisible par p . On appelle $v_p(a) = k$ la *valuation p -adique* de a , et l'on définit la valeur absolue p -adique

$$|a|_p = p^{-v_p(a)}.$$

On pose $v_p(0) = +\infty$, et $|0|_p = 0$. Cette valeur absolue vérifie $|ab|_p = |a|_p |b|_p$.

On peut définir de la même manière que précédemment une distance d_p sur \mathbb{N} , et considérer le complété de \mathbb{N} pour cette distance, que l'on note \mathbb{Z}_p .

L'identification des éléments de \mathbb{Z}_p peut se faire à partir de la décomposition d'un entier en base p :

$$a = \sum_{n=0}^{+\infty} a_n p^n, \quad a_n \in \llbracket 0, p-1 \rrbracket \quad \forall n, \quad a_n = 0 \text{ au-delà d'un certain rang.}$$

Un élément de \mathbb{Z}_p peut ainsi s'écrire comme une série infinie du type de celle ci-dessus, ou simplement codée par ses coefficients (on garde la convention d'une infinité de chiffres *avant* la virgule)

$$a = \dots a_2 a_1 a_0, 0.$$

Noter que cette complétion et l'identification à des nombres écrits en base p avec une infinité de chiffres avant la virgule ne nécessite pas que p soit premier. Si p n'est pas premier, on perd la propriété $|ab|_p = |a|_p |b|_p$, par exemple $|2 \times 5|_{10} = 10^{-1}$ alors que $|2|_{10} = |5|_{10} = 1$. Mais tant que l'on s'en tient à des aspects métriques, et que l'on se contente d'additionner les nombres entre eux, la construction est valide. On peut en particulier construire l'objet \mathbb{Z}_{10} des entiers 10-adiques, qui présente une forte analogie apparente (le codage semble le même, à symétrie près) avec l'ensemble des réels de l'intervalle $[0, 1]$, mais qui présente des propriétés très différentes. Pour l'anecdote, si l'on considère la relation d'ordre lexicographique sur \mathbb{Z}_{10} , son plus grand nombre est (on notera que l'infinité de 9 avant la virgule n'est pas ici pathologique, le codage de \mathbb{Z}_{10} est tout à fait injectif, contrairement au codage décimal des réels)

$$\dots 9999, 0 = \sum_{n=0}^{+\infty} 9 \times 10^n = 9 \frac{1}{1-10} = -1,$$

de telle sorte que l'on peut écrire $\mathbb{Z}_{10} = [0, -1]$ (*sic*).

On se restreint néanmoins en général aux nombres premiers, qui permettent des développements très féconds dans un cadre algébrique. Si l'on se restreint ainsi aux nombres premiers, on a une formule à la fois spectaculaire et très simple à établir, qui relie toutes les valeurs absolues p -adiques entre elles. Plus précisément, si l'on note $|\cdot|_\infty$ la valeur absolue usuelle sur \mathbb{Z} , on a

$$\forall a \in \mathbb{Z}, \quad |a|_\infty \prod_{p \text{ premier}} |a|_p = 1.$$

On notera que pour p assez grand (notamment plus grand que $|a|$), on a $|a|_p = 1$, le produit ci-dessus peut donc se ramener à un produit fini, dont la définition ne nécessite pas d'arguments topologiques.

Mesure sur \mathbb{Z}_2 , sur \mathbb{Z}_p

Nous décrivons succinctement dans cette section la démarche permettant de construire une mesure sur \mathbb{Z}_2 . Le rôle joué par les intervalles dans le cas réel est ici joué par les boules. Nous privilégierons les boules fermées $\overline{B}(a, 2^{-k})$, en gardant en tête que ce sont aussi des boules ouvertes. En effet les valeurs prises par la distance étant quantifiées (ce sont les 2^{-k}), on a $B(a, 2^{-k}) = \overline{B}(a, 2^{-(k-1)})$. On notera que la boule fermée de centre a et de rayon 2^{-k} s'écrit aussi

$$a + 2^k \mathbb{Z}_2 = \{a + 2^k z, z \in \mathbb{Z}_2\} = \{\dots a_{k-1} a_{k-2} \dots a_0\}$$

c'est-à-dire l'ensemble des éléments de \mathbb{Z}_2 dont l'écriture 2-adique commence comme celle de a (au moins jusqu'au rang $k-1$).

Nous noterons $\mathcal{B} = \mathcal{B}(\mathbb{Z}_2)$ la tribu des boréliens sur \mathbb{Z}_2 , i.e. la tribu engendrée par les ouverts de \mathbb{Z}_2 .

Proposition 19.45. La tribu \mathcal{B} des boréliens est engendrée par les $a + 2^k \mathbb{Z}_2$, avec a et k parcourant \mathbb{N} .

Démonstration. Montrons en premier lieu que tout ouvert est réunion dénombrable de boules (fermées!). Soit un ouvert U de \mathbb{Z}_2 , et $a \in U$. Il existe k tel que $a + 2^k \mathbb{Z}_2 \subset U$. Notons $\bar{a} \in \mathbb{N}$ la troncature de a au rang k , i.e.

$$\bar{a} = \sum_{n=0}^{k-1} a_n 2^n.$$

On a $a + 2^k \mathbb{Z}_2 = \bar{a} + 2^k \mathbb{Z}_2$. L'ouvert U est donc union de boules du type $a + 2^{-k} \mathbb{Z}_2$, il s'agit donc d'une union dénombrable¹⁶.¹⁷ Noter, même si ça n'est pas nécessaire dans la démonstration, qu'on peut ne garder qu'une seule boule par $a \in \mathbb{N}$ impliqué, du fait que deux boules sont soit disjointes soit concentriques (l'une dans l'autre). On peut donc "coder" un ouvert de \mathbb{Z}_2 par une suite $(a_i, k_i)_i$ dans \mathbb{N}^2 . Cette propriété implique que la tribu engendrée par les $a + 2^k \mathbb{Z}_2$, contient la tribu des boréliens, et donc s'identifie à elle du fait qu'il s'agit d'ouverts. \square

On cherche maintenant à définir une mesure μ sur \mathbb{Z}_2 , qui affecte une masse 1 à \mathbb{Z}_2 . Nous verrons qu'il est possible de définir une telle mesure sur la tribu borélienne. Du fait que \mathbb{Z}_2 est union disjointe de 2^k boules fermées de rayon 2^{-k} , on affecte le volume 2^{-k} à toute boule ouverte de rayon 2^{-k} . Il peut sembler étonnant d'affecter un volume égal au *rayon*, mais on se souviendra que dans ce contexte ultramétrique, le rayon est égal au diamètre.

On se propose maintenant de construire, à partir de cette définition du volume sur le $\pi-$ système des $a + 2^k \mathbb{Z}_2$, une mesure extérieure sur \mathbb{Z}_2 , en suivant la démarche décrite sur \mathbb{R} .

Proposition 19.46. Pour tout $A \subset \mathbb{Z}_2$, on note C_A l'ensemble des suites de boules fermées dont l'union recouvre A :

$$C_A = \left\{ (a_i + 2^{k_i} \mathbb{Z}_2)_{i \in \mathbb{N}} , A \subset \bigcup_{\mathbb{N}} (a_i + 2^{k_i} \mathbb{Z}_2) \right\}.$$

On autorise les rayons à être nul (i.e. $k = +\infty$), ce qui autorise à considérer des collections avec un nombre fini de boules de volume non nul. On définit alors $\mu^* : \mathcal{P}(\mathbb{Z}_2) \longrightarrow [0, +\infty]$ par

$$\lambda^*(A) = \inf_{C_A} \left(\sum_i 2^{-k_i} \right). \quad (19.20)$$

Cette application est une mesure extérieure, et elle attribue à toute boule fermée de rayon 2^{-k} la valeur 2^{-k} .

Démonstration. En premier lieu $\mu^*(\emptyset) = 0$. Considérons maintenant une collection (A_n) de parties de \mathbb{Z}_2 . On peut trouver pour chaque partie une collection de boules qui la recouvre, et telle que la somme des volumes approche $\mu^*(A_n)$ à $\varepsilon/2^n$ près. Cela permet d'établir que

$$\mu^* \left(\bigcup A_n \right) \leq \sum \mu^*(A_n) + 2\varepsilon,$$

et ce pour tout ε . On en déduit la sous-additivité.

Montrons maintenant que cette mesure extérieure affecte 2^{-k} aux boules fermées de rayon 2^{-k} . Soit $B = a + 2^k \mathbb{Z}_2$ une telle boule. En premier lieu, la boule est recouverte par elle-même, d'où $\mu^*(B) \leq 2^{-k}$. Maintenant considérons un recouvrement de B par des boules. Ces boules fermées étant des ouverts, et B étant compacte, on peut en extraire un recouvrement fini, on peut enlever les boules qui sont incluses dans une autre du recouvrement, pour obtenir (d'après le corollaire 19.35) une partition finie, telle que la somme des volumes est exactement 2^{-k} . \square

On considère maintenant \mathcal{A} la tribu des parties mesurables pour μ^* , et l'on note μ la mesure définie sur \mathcal{A} issue de μ^* . Il s'agit maintenant de montrer que la tribu \mathcal{A} contient la tribu des boréliens. Comme dans le cas réel, il suffit de vérifier que les boules fermées, qui engendrent \mathcal{A} , sont mesurables.

17. Elle s'applique d'ailleurs à toute construction d'une mesure extérieure par l'extérieur précisément, basée sur ce principe de recouvrement par des objets donc on a fixé la taille.

Proposition 19.47. La tribu \mathcal{A} des parties mesurables pour μ^* contient les boules fermées, donc les boréliens. La mesure μ introduite ci-dessus est ainsi définie sur la tribu des boréliens.

Démonstration. Soit $B = a + 2^k \mathbb{Z}_2$ une boule fermée. Il s'agit de montrer que pour toute partie $A \subset \mathbb{Z}_2$,

$$\mu^*(A) \geq \mu^*(A \cap B) + \mu^*(A \cap B^c).$$

On considère un recouvrement de A par des boules fermées $B_n = a_n + 2^{k_n} \mathbb{Z}_2$, avec

$$\sum 2^{-k_n} \leq \mu^*(A) + \varepsilon.$$

Il s'agit de distribuer au mieux chacune de ces boules entre B et B^c . Là encore, la preuve est plus simple que dans le cas réel. Soit une telle boule B_n . Si elle ne rencontre pas B , alors soit incluse dans B^c , auquel cas on l'affecte à B^c . Si elle rencontre B , alors soit elle est incluse dans A , auquel cas on l'affecte à A soit, du fait de l'utramétricité, elle contient B . Son rayon est 2^{-k_n} , avec $k_n < k$. La boule B_n peut alors s'écrire, comme toute boule fermée, comme réunion disjointe de 2^{k-k_n} boules fermées de rayon 2^{-k_n} (l'une d'elle étant B). On atomise donc la boule B_n en 2^{k-k_n} boules plus petites, on affecte la boule qui s'identifie à B à B , et l'on affecte les $2^{k-k_n} - 1$ autres à B^c , ce qui se fait sans augmenter la masse totale. On construit ainsi à partir du recouvrement de A deux recouvrements de $A \cap B$ et $A \cap B^c$, respectivement, sans changer la masse totale, d'où l'on déduit que

$$\mu^*(A \cap B) + \mu^*(A \cap B^c) \leq \sum 2^{-k_n} \leq \mu^*(A) + \varepsilon$$

pour tout ε , d'où la propriété. □

19.14 Distance de Gromov-Wasserstein

Definition 19.48. Soient (X, d, μ) et (X', d', μ') deux espaces métriques finis probabilisés (i.e. munis d'une mesure définie sur la tribu discrète, de masse totale 1). On dit que ces espaces sont isomorphes s'il existe une bijection de X vers X' qui préserve les structures de distance et de mesure, i.e. s'il existe une isométrie T telle que

$$T_\sharp \mu = \mu' \quad \text{i.e.} \quad \mu'(A') = \mu(T^{-1}(A')) \quad \forall A \in \mathcal{P}(X').$$

Soient (X, d, μ) et (X', d', μ') deux espaces métriques finis probabilisés, de cardinaux respectifs N et N' , munis de leurs tribus discrètes respectives. On note Π l'ensemble des plans de transport entre μ et μ' :

$$\Pi = \left\{ \gamma = (\gamma_{xx'}) \in \mathbb{R}_+^{N \times N'}, \sum_x \gamma_{xx'} = \mu'_{x'}, \sum_{x'} \gamma_{xx'} = \mu_x \right\}.$$

Definition 19.49. Pour $p \in [1, +\infty[$, on définit

$$d_{GWp}(X, X') = \inf_{\gamma \in \Pi} \left(\sum_{xx'} \sum_{yy'} |d(x, y) - d(x', y')|^p \gamma_{xx'} \gamma_{yy'} \right)^{1/p}.$$

Lemme 19.50. L'infimum de la définition précédente est atteint.

Démonstration. L'ensemble admissible Π est compact, et l'application

$$\gamma \mapsto \left(\sum_{xx'} \sum_{yy'} |d(x, y) - d(x', y')|^p \gamma_{xx'} \gamma_{yy'} \right)^{1/p}$$

est continue. □

Proposition 19.51. La quantité définie ci-dessus est une distance sur l'ensemble des espaces métriques probabilisés finis (quotienté par les isomorphismes au sens de la définition 19.48)

Démonstration. Si l'on a $d_{GWp}(X, X') = 0$ alors, pour tous x, x', y, y' tels que $\gamma_{xx'} \neq 0$ et $\gamma_{yy'} \neq 0$, on a $d(x, y) = d(x', y')$. Soit maintenant x, x', y', y'' , tels que x envoie de la masse à x' et y' . On a $d(x', y') = d(x, x) = 0$, d'où $x' = y'$. Pour chaque x il existe donc unique x' tel que $\gamma_{xx'} \neq 0$, et l'on a donc $\gamma_{xx'} = \mu_x$. On peut mener le même raisonnement dans l'autre sens : pour chaque x' il existe un unique x tel que $\gamma_{xx'} \neq 0$, et l'on a donc $\gamma_{xx'} = \mu_{x'}$. Le plan de transport γ correspond donc à une bijection T entre X et X' , et l'on a

$$\begin{aligned} 0 = d_{GWp}(X, X') &= \left(\sum_{xx'} \sum_{yy'} |d(x, y) - d(x', y')|^p \gamma_{xx'} \gamma_{yy'} \right)^{1/p} \\ &= \left(\sum_x \sum_y |d(x, y) - d(T(x), T(y))|^p \right)^{1/p} \end{aligned}$$

La symétrie est immédiate d'après la définition.

Pour l'inégalité triangulaire, on considère 3 espaces métriques probabilisés X, X' , et X'' , et des plans de transport $(\gamma_{xx'})$ et $(\gamma_{x'x''})$ qui réalisent les distances entre X et X' et entre X' et X'' , respectivement. On construit à partir de ces plans un plan entre X et X'' (non nécessairement optimal, mais qui suffira pour l'inégalité triangulaire) en “collant” dans un premier temps les plans, puis en condensant l'espace X' intermédiaire. Plus précisément, on introduit

$$\gamma_{xx'x''} = \frac{\gamma_{xx'} \gamma_{x'x''}}{\mu_{x'}},$$

et l'on définit

$$\gamma_{xx''} = \sum_{x' \in X'} \gamma_{xx'x''}.$$

On a

$$\sum_{xx''} \sum_{yy''} |d(x, y) - d(x'', y'')| \gamma_{xx''} \gamma_{yy''} = \sum_{xx''} \sum_{yy''} |d(x, y) - d(x'', y'')| \sum_{x'} \frac{\gamma_{xx'} \gamma_{x'x''}}{\mu_{x'}} \sum_{y'} \frac{\gamma_{yy'} \gamma_{y'y''}}{\mu_{y'}}.$$

On écrit $|d(x, y) - d(x'', y'')| \leq |d(x, y) - d(x', y')| + |d(x', y') - d(x'', y'')|$, ce qui permet de majorer la quantité de départ par deux termes. le premier s'écrit

$$\begin{aligned} &\sum_{xx''} \sum_{yy''} |d(x, y) - d(x', y')| \sum_{x'} \frac{\gamma_{xx'} \gamma_{x'x''}}{\mu_{x'}} \sum_{y'} \frac{\gamma_{yy'} \gamma_{y'y''}}{\mu_{y'}} \\ &= \sum_x \sum_y |d(x, y) - d(x', y')| \sum_{x'} \frac{\gamma_{xx'}}{\mu_{x'}} \underbrace{\sum_{x''} \gamma_{x'x''}}_{\mu_{x'}} \sum_{y'} \frac{\gamma_{yy'}}{\mu_{y'}} \underbrace{\sum_{y''} \gamma_{y'y''}}_{\mu_{y'}} \\ &= \sum_x \sum_y |d(x, y) - d(x', y')| \sum_{x'} \gamma_{xx'} \gamma_{yy'} = d_{GW}(X, X'). \end{aligned}$$

Le second terme s'identifie de la même manière à $d_{GW}(X', X'')$

□

19.15 Dendrogrammes

On décrit ici un procédé de construction d'un arbre à partir d'un espace métrique fini, qui induit une nouvelle métrique sur cet espace, de nature ultramétrique, et permet de visualiser d'une certaine manière la structure de l'espace. Il s'agit d'un procédé constructif et graphique, dont nous donnons ici une définition abstraite

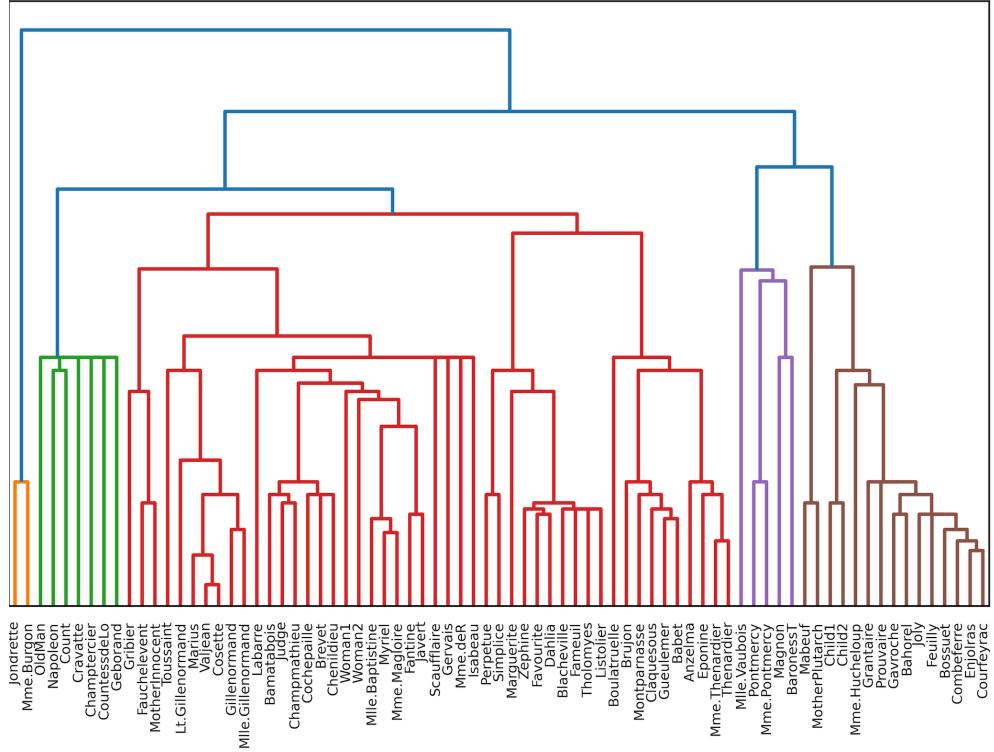


FIGURE 19.6 – Dendrogramme des Misérables

Definition 19.52. (Dendrogramme)

On considère un espace métrique fini (X, d) , de cardinal N . On appelle dendrogramme de X une suite finie $(X^k)_{0 \leq k < N}$ de partitions de X :

$$X^k = \{X_1^k, \dots, X_{N-k}^k\}, \quad X_i^k \in \mathcal{P}(X), \quad X_i^k \cap X_j^k = \emptyset \quad \forall i \neq j, \quad X = \bigcup_{j=1}^{N-k} X_j^k,$$

ainsi qu'une suite réelle $(D^k)_{1 \leq k \leq N}$, construits selon le processus d'agrégation suivant. La première partition X_0 est la plus fine, de cardinal N , et $D^0 = 0$. Supposons X^k connue, de cardinal $N - k$, avec $k \geq 0$. On construit la partition suivante en agrégeant 2 éléments de la partition, selon le principe suivant. On définit

$$D_{ij}^{k+1} = d(X_i^k, X_j^k) = \inf_{x \in X_i^k, y \in X_j^k} d(x, y),$$

et l'on choisit un couple (i, j) qui minimise cette quantité. On note D^{k+1} le minimiseur, et l'on définit la partition X^k en conservant les éléments autres que i et j , auxquels on adjoint la réunion de X_i^k et X_j^k . Le cardinal de la partition est donc $N - k - 1$. La suite D^k est croissante par construction, avec $D^1 > 0$, et X^{N-1} est la partition triviale ($X^{N-1} = \{X\}$).

Proposition 19.53. (Métrique ultramétrique associée à un dendrogramme)

On définit

$$\delta : (x, y) \in X \times X \mapsto \delta(x, y) \in \mathbb{R}_+$$

de la façon suivante : pour x et y dans X donné, on définit k_{xy} comme le plus petit entier tel que x et y sont dans une même sous-partie de X^k , et l'on fixe $\delta(x, y) = D^k$. L'application $\delta(\cdot, \cdot)$ ainsi définie est une distance, qui est ultra-métrique, c'est à dire qu'elle vérifie l'inégalité triangulaire renforcée

$$\delta(x, y) \leq \max(\delta(x, z), \delta(z, y)) \quad \forall x, y, z \in X$$

Démonstration. L'application δ prend bien des valeurs positives, on a $\delta(x, y) = 0$ si et seulement si $k_{xy} = 0$, si et seulement si $x = y$. Elle est symétrique par construction. Soient maintenant x, y, z , et les k_{xy} et k_{yz} associés. Pour $k \geq \max(k_{xy}, k_{yz})$, x et z sont dans la même composante, d'où

$$k_{yxz} \leq \max(k_{xy}, k_{yz}),$$

d'où l'inégalité ultramétrique par croissance de la suite (D_k) . □

À titre d'illustration, la figure 19.6 représente graphiquement le dendrogramme obtenu à partir de l'ensemble X des personnages des Misérables, muni d'une métrique qui prend en compte la proximité dans le roman des personnages (la distance est d'autant plus petite que les 2 personnages partagent un nombre important de scènes).

Chapitre 20

Problèmes

Sommaire

20.1	Conditions aux limites de Robin sur graphe	408
20.2	Propagation de la chaleur sur un réseau	411
20.3	Résistance de l'hypercube	413
20.4	Clivage	414
20.5	Charisme	416
20.6	Stabilité du poumon humain	418
20.7	Poumon non linéaire	420
20.8	Phénomène de limitation du débit expiratoire	423
20.9	Mobilité et équilibres de Wardrop	424
20.10	Optimisation d'une préparation de concours	426
20.11	Modèles de foules de type Nash	428
20.12	Mouvement de véhicules autonomes	431
20.13	Transport partiel	432
20.14	Transport sous contraintes	434
20.15	Décomposition polaire discrète	435
20.16	Entropie relative	436
20.17	Décroissance de l'entropie pour les schémas de différences finies	439
20.18	Flots de gradients discrets dans l'espace de Wasserstein, équation de la chaleur comme flot de gradient de l'entropie	440

20.1 Conditions aux limites de Robin sur graphe

On considère un réseau résistif $G = (V, E, r, \Gamma)$, où V est un ensemble fini de sommets, $E \subset V \times V$ symétrique, $r \in]0, +\infty[^E$ symétrique, et $\Gamma \subset V$ non vide. On supposera (sauf mention contraire dans certaines questions) le réseau *connexe*. On définit comme dans le cours les opérateurs ∂ et ∂^* :

$$\begin{aligned}\partial & : u \in \mathbb{R}^E \mapsto \partial u \in \mathbb{R}^V, \quad (\partial u)_x = \sum_{y \sim x} u_{yx}, \\ \partial^* & : p \in \mathbb{R}^V \mapsto \partial^* p \in \mathbb{R}^E, \quad (\partial^* p)_{xy} = p_y - p_x.\end{aligned}$$

Conditions de Dirichlet

On se donne un champ de pressions $P \in \mathbb{R}^\Gamma$ sur la frontière. On rappelle que le problème de Dirichlet consiste à trouver $p \in \mathbb{R}^{\mathring{V}}$ qui vérifie

$$\begin{cases} \partial c \partial^* p = 0 & \text{dans } \mathring{V}, \\ p = P & \text{sur } \Gamma. \end{cases} \quad (20.1)$$

L'équation de Laplace ci-dessus signifie que, pour tout $x \in \mathring{V}$,

$$\sum_{y \sim x} c_{xy} (p_x - p_y) = 0,$$

avec $c_{xy} = 1/r_{xy}$ pour tout $(x, y) \in E$.

N.B. : les questions 1 et 2 reprennent essentiellement les deux démarches proposées dans le cours pour établir le caractère bien posé du problème de Dirichlet.

1) On note $p = (\mathring{p}, P) \in \mathbb{R}^{\mathring{V}} \times \mathbb{R}^\Gamma$ un champ vérifiant la condition de Dirichlet.

a) Montrer que le problème (20.1) peut se mettre sous la forme

$$\mathring{A}\mathring{p} = b$$

où \mathring{A} est une matrice carrée de $\mathbb{R}^{\mathring{V} \times \mathring{V}}$, et b un vecteur de $\mathbb{R}^{\mathring{V}}$ que l'on précisera.

b) Montrer que la matrice \mathring{A} est injective, et en déduire l'existence et l'unicité d'une solution au problème.

2) (Approche variationnelle)

Pour $q = (\mathring{q}, P)$, on définit

$$F : \mathring{q} \longmapsto F(\mathring{q}) = \frac{1}{2} \sum_{(x,y) \in E} c_{xy} (q_x - q_y)^2, \quad (20.2)$$

(avec $q_y = P_y$ quand $y \in \Gamma$, et $q_y = \mathring{q}_y$ quand $y \in \mathring{V}$). On rappelle que dans toutes sommes sur les arêtes, on ne compte qu'une fois chaque arête ((x, y) et (y, x) sont réunies en un même terme).

Montrer que l'équation $\partial c \partial^* p = 0$ sur \mathring{V} (avec $p = (\mathring{p}, P)$) exprime l'annulation du gradient de F en \mathring{p} . Rappeler succinctement comment cette identification permet d'assurer l'existence et l'unicité d'une solution au problème de Dirichlet (20.1) (on pourra admettre ici la stricte convexité et la coercivité de F , des démonstrations analogues étant demandées plus loin, dans un contexte voisin).

3) a) Que peut-on dire de la solution p si Γ est un singleton ?

b) Que peut-on dire de la solution p si P est uniforme sur Γ ?

4) Que peut-on dire si le réseau n'est pas connexe ? (On précisera les hypothèses permettant d'avoir malgré tout un résultat d'existence et d'unicité d'une solution).

5) On suppose dans cette question que G est un arbre dyadique à N générations (comme dans le cas du poumon humain).

a) Décrire aussi précisément que possible la solution du problème de Dirichlet dans le cas où Γ est constitué de seulement 2 sommets parmi les feuilles de l'arbre (i.e. points de la génération N), en lesquels on impose les valeurs de pression 0 et 1. (*On pourra s'aider d'un dessin.*)

b) Comme à la question précédente, décrire la solution dans le cas où Γ est constitué de 2 sommets qu'on ne suppose plus appartenir à la dernière génération.

Conditions de Robin

On s'intéresse maintenant à un autre type de conditions aux limites, appelées conditions de Robin. On se donne un champ scalaire strictement positif sur Γ , noté $\rho \in]0, +\infty[^{\Gamma}$ (on comprendra plus loin que les ρ_x , même si elles sont afférentes à des sommets, sont homogènes à des résistances). On se donne également, comme pour les conditions de Dirichlet, un champ de pression sur la frontière : $P \in \mathbb{R}^{\Gamma}$. La condition aux limites s'écrit, en chaque point x de Γ ,

$$-\rho_x(\partial u)_x + p_x = P_x,$$

avec comme toujours $u = -c\partial^*p$. Le problème global s'écrit ainsi

$$\begin{cases} \partial c\partial^*p = 0 & \text{dans } \mathring{V}, \\ -\rho\partial u + p = P & \text{sur } \Gamma, \end{cases} \quad (20.3)$$

(N.B. : conformément à l'usage suivi en cours, $\rho\partial u$ est le champ scalaire défini sur Γ , qui prend la valeur $\rho_x(\partial u)_x$ au point $x \in \Gamma$.)

On définit la fonctionnelle

$$q \in \mathbb{R}^V \longmapsto \Phi(q) = \frac{1}{2} \sum_{(x,y) \in E} c_{xy}(q_y - q_x)^2 + \frac{1}{2} \sum_{x \in \Gamma} c_x(q_x - P_x)^2,$$

avec $c_x = 1/\rho_x$ pour tout x de Γ .

6) Montrer que p est solution du système (20.3) si et seulement si p annule le gradient de Φ .

7) Montrer que Φ est coercive.

(Indication : pour tout point x de \mathring{V} , on pourra considérer un chemin simple de $z \in \Gamma$ vers x , écrire p_x en fonction de P_z , p_z et des p_y du chemin, et en déduire une majoration de $|p_x|$ en fonction de $J(p)$). Pour le terme de bord, on pourra écrire $p_z = P_z + (p_z - P_z)$, et majorer $|p_z - P_z|$ par une expression qui fait intervenir un des termes de la fonctionnelle Φ .

8)a) Montrer que, si p et p' sont deux champs de \mathbb{R}^V , avec $p \neq p'$, alors au moins l'une des deux assertions suivantes est vérifiée :

(i) il existe $(x, y) \in E$ tel que $p_x - p_y \neq p'_x - p'_y$;

(ii) il existe $x \in \Gamma$ tel que $p_x \neq p'_x$.

b) Montrer que Φ est strictement convexe.

9) Montrer à l'aide des questions précédentes l'existence et l'unicité d'une solution au problème (20.3).

(N.B. : La démonstration peut tenir en quelques lignes, mais on prendra garde à bien préciser le raisonnement, et les hypothèses utilisées à chaque étape de ce raisonnement.)

10) Que peut-on dire si le réseau n'est pas connexe ? (on précisera les hypothèses permettant d'avoir un résultat d'existence et d'unicité d'une solution).

Les deux questions suivantes proposent une démarche alternative à ce qui précède, pour démontrer le caractère bien posé du problème de Robin, dans le cas d'un réseau connexe. A partir du réseau $G = (V, E, r, \Gamma)$, on construit un nouveau réseau en rajoutant, pour tout x de Γ , un sommet x' relié à x , et seulement à x . Si N est le cardinal de Γ , on a donc rajouté exactement N sommets et N arêtes. On note Γ' l'ensemble des N sommets rajoutés. On affecte à chaque arête (x, x') la résistance ρ_x . On note $G' = (V', E', r', \Gamma')$ le nouveau réseau, avec $V' = V \cup \Gamma'$, E' est l'union de E avec l'ensemble des N nouvelles arêtes, r' est le nouveau champ de résistances ($r'_{xy} = r_{xy}$ pour $(x, y) \in E$, et $r'_{xx'} = \rho_x$ pour chaque nouvelle arête (x, x')), et Γ' est la frontière du nouveau réseau. Noter que les points de Γ sont devenus des points intérieurs au nouveau réseau G' , ce que l'on peut écrire $\mathring{V}' = V$.

On note $P' \in \mathbb{R}^{\Gamma'}$ le champ P “recopié” sur Γ' , c'est à dire tel que $P'_{x'} = P_x$ pour tout $x' \in \Gamma'$, x' connecté à x .

- 11) Montrer que $p \in \mathbb{R}^V$ est solution du problème (20.3) avec conditions de Robin sur G si et seulement si $(p, P') \in \mathbb{R}^V \times \mathbb{R}^{\Gamma'} = \mathbb{R}^{V'}$ est solution du problème de Dirichlet sur G' , avec condition au bord P' sur la nouvelle frontière Γ' .
- 12) En déduire une démonstration alternative de l'existence et l'unicité d'une solution au problème (20.3).
- 13) Proposer une interprétation de la condition de Robin, en termes de modélisation, au vu des considérations précédentes.

De Robin vers Dirichlet

On considère le problème (20.3) sur réseau connexe, en supposant que ρ_x prend une même valeur $\beta > 0$ pour tout $x \in \Gamma$. On note p^β la solution associée à la valeur β , qui est, d'après ce qui précède, l'unique minimiseur de la fonctionnelle

$$q \in \mathbb{R}^V \longmapsto \Phi_\beta(q) = \frac{1}{2} \sum_{(x,y) \in E} c_{xy}(q_y - q_x)^2 + \frac{1}{2\beta} \sum_{x \in \Gamma} (q_x - P_x)^2.$$

- 14) On souhaite montrer que, quand β tend vers 0, p^β tend p^0 , défini comme l'unique minimiseur de la fonctionnelle F (définie par (20.2)), sur l'ensemble des champs de pressions qui vérifient la condition de Dirichlet $p_x = P_x$ pour tout x sur Γ .

- a) Montrer que la suite p^β est bornée, et en déduire que l'on peut extraire une sous-suite qui converge vers $\bar{p} \in \mathbb{R}$. On notera toujours p^β cette sous-suite pour simplifier l'écriture.
- b) Montrer que l'on a, pour tout $\beta > 0$,

$$F(p^\beta) \leq \Phi_\beta(p^\beta) \leq \Phi_\beta(p^0) = F(p^0).$$

- c) Déduire de la question précédente que $F(\bar{p}) \leq F(p^0)$, et que \bar{p} vérifie les conditions de Dirichlet $\bar{p}_x = P_x$ pour tout $x \in \Gamma$.

- d) Conclure.

20.2 Propagation de la chaleur sur un réseau

Exercice 20.1. (Propagation de la chaleur sur un réseau)

On considère un système de N corps, qui échangent de la chaleur entre eux. On note V (pour *vertices*, il s'agit des sommets d'un graphe que nous allons définir) l'ensemble des corps, u_x^t la température au temps t du corps x , et C_x la capacité thermique de x , de telle sorte que l'énergie thermique de x est $C_x u_x$. On représente par $E \subset V \times V$ symétrique (E pour *Edges*, il s'agit des arêtes du graphe) la connectivité des relations entre corps : $(x, y) \in E \iff (y, x) \in E$ si x et y échangent de la chaleur. On suppose le graphe connexe : on peut passer d'un point quelconque à tout autre point par un chemin constitué d'arêtes. On suppose que le flux de chaleur compté positivement de y vers x est proportionnel à la différence de température : il s'écrit

$$Q_{y \rightarrow x} = c_{xy}(u_y - u_x)$$

de telle sorte que l'évolution des températures suit la collection d'équations

$$C_x \frac{du_x^t}{dt} = - \sum_{y \sim x} c_{xy}(u_x - u_y) \quad \forall x \in V. \tag{20.4}$$

1) Donner l'expression de l'énergie thermique totale du système (supposé isolé du monde extérieur), et montrer qu'elle se conserve au cours du temps.

2) On explore dans cette question la possibilité que la propagation de la chaleur de y vers x soit régie par une relation plus générale

$$Q_{y \rightarrow x} = \varphi_{xy}(u_y - u_x),$$

où $\varphi_{xy} = \varphi_{yx}$ est une fonction de \mathbb{R} dans \mathbb{R} qui caractérise la propagation de chaleur de y vers x . Donner une condition suffisante sur la collection des fonctions φ_{xy} pour que le premier principe, c'est-à-dire la conservation de l'énergie totale, soit vérifiée.

3) On revient au problème initial. Montrer que l'on peut définir une notion de température moyenne \bar{u} du système qui se conserve au cours du temps. Donner l'expression de la capacité thermique du système global.

4) Montrer que l'écart quadratique moyen (variance) des températures vis-à-vis de cette température moyenne \bar{u} diminue au cours du temps, et que cette diminution est stricte tant que l'équilibre thermique n'est pas atteint.

5) Montrer que, si la distribution de température n'est pas uniforme au départ, elle ne l'est pour aucun temps.

6) On revient au système linéaire 20.4. On se propose d'établir une certaine forme de *principe du maximum*, c'est à dire que si les valeurs initiales (u_x^0) sont dans un intervalle $[m, M]$, les valeurs u_x^t restent dans ce même intervalle pour tout temps. On établit dans un premier temps la préservation de la positivité. On suppose que les valeurs initiales sont toutes dans $[0, +\infty[$, qu'elles ne sont pas toutes égales, et l'on considère la solution du système $t \mapsto (u_x^t)_V$ pour $t \in [0, T]$.

a) Montrer que le minimum de u_x^t est atteint sur $V \times [0, T]$, en un point (x_0, t_0) .

On cherche à montrer dans les questions suivantes que $t_0 = 0$.

b) Montrer que l'on peut supposer sans perte de généralité que x_0 , qui réalise le minimum à t_0 , est voisin d'au moins un point en lequel la température est strictement supérieure au minimum. Montrer que l'on a alors

$$-\sum_{y \sim x} c_{x_0 y} (u_{x_0} - u_y) > 0.$$

c) Montrer que, si $t_0 > 0$, alors $d u_{x_0}^t / dt \leq 0$ en $t = t_0$.

d) En conclure que $t_0 = 0$, et que toutes les températures sont positives.

e) Déduire de ce qui précède que, si les valeurs initiales (u_x^0) sont dans un intervalle $[m, M]$, les valeurs u_x^t restent dans ce même intervalle pour tout temps.

On suppose que toutes les températures initiales sont strictement positives, et l'on fixe les valeurs initiales des entropies de tous les corps à 0.

7) Montrer que les températures restent > 0 pour tout temps.

8) écrire le système d'équations vérifié par la collection des entropies s_x^t .

9) Montrer que l'entropie (physique) totale du système croît, et que cette croissance est stricte tant que l'état d'équilibre (température uniforme) n'est pas atteint.

10) Dans l'esprit de la question 2, on explore ici la possibilité que la propagation de la chaleur de y vers x soit régie par une relation plus générale

$$Q_{y \rightarrow x} = \varphi_{xy}(u_y - u_x),$$

où $\varphi_{xy} = \varphi_{yx}$ est une fonction de \mathbb{R} dans \mathbb{R} qui caractérise la propagation de chaleur de y vers x . Donner une condition suffisante sur la collection des fonctions φ_{xy} pour que à la fois le premier et le second principes de la thermodynamique soient respectés, et interprétez physiquement cette condition.

20.3 Résistance de l'hypercube

On considère le graphe non orienté $G = (V, E)$ défini de la façon suivante : $V = \{0, 1\}^n$ est l'ensemble des n -uplets de 0 ou 1, ou “mots” de n bits, et l'on dit que $(x, y) \in E$ si les mots x et y ne diffèrent que d'un bit, i.e. si $x = a_1a_2\dots a_n$, et $y = b_1b_2\dots b_n$,

$$(x, y) \in E \iff (y, x) \in E \iff \exists ! i \text{ tel que } a_i \neq b_i$$

On rappelle que la longueur d'un chemin est définie comme le nombre d'arêtes qui le constituent, et la distance $d(x, y)$ entre deux sommets x et y comme la longueur du plus court chemin entre ces deux sommets. Pour $x = a_1a_2\dots a_n \in V$, on définit w_x comme le nombre de 1 dans l'écriture de x , i.e.

$$w_x = \sum_{i=1}^n a_i.$$

- 1) a) Préciser le nombre de sommets, le degré des sommets, et le nombre d'arêtes du graphe (en comptant pour une seul arête (x, y) et (y, x)).
- b) On appelle G l'hypercube de dimension n , ou simplement n -cube. Justifier cette appellation.
- c) Pour $n = 3$, donner des exemples de cycles de longueurs 4 et 6.
- d) Pour n quelconque, montrer qu'il n'existe aucun cycle de longueur impaire.
- 2) a) Préciser le diamètre du graphe (plus grande distance entre deux sommets).
- b) Soit $x = a_1a_2\dots a_n$ (avec $a_i = 0$ ou 1). Montrer qu'il existe un unique point \tilde{x} “diamétralement opposé” à x , c'est à dire dont la distance à x est égale au diamètre, et préciser son expression $\tilde{x} = \tilde{a}_1\tilde{a}_2\dots \tilde{a}_n$ en fonction de celle de x .
- c) Combien y a-t-il de paires de points diamétralement opposés ?
- d) Combien y a-t-il de plus courts chemins entre deux sommets x et \tilde{x} diamétralement opposés ?
- d) Pour $x \in V$, k entier entre 0 et n , on introduit la sphère centrée en x et de rayon k :

$$S(x, k) = \{y \in V, d(x, y) = k\} .$$

Donner le cardinal de $S(x, k)$ pour tout k , et vérifier que le cardinal de V s'écrit bien comme la somme des cardinaux des $S(x, k)$, pour k allant de 0 à n .

- e) On se place dans le cas $x = o = 00\dots 0$. Montrer que

$$S(o, k) = \{y \in V, w_y = k\} ,$$

où w_y est le nombre de 1 dans l'écriture de y (comme défini en préambule).

- 3) Soit φ un élément du groupe symétrique, i.e. une bijection de l'ensemble $\{1, \dots, n\}$ dans lui-même. On associe à φ l'application T_φ de V dans lui-même définie par

$$x = a_1a_2\dots a_n \mapsto T_\varphi(x) = a_{\varphi(1)}a_{\varphi(2)}\dots a_{\varphi(n)}$$

Montrer que le champ (w_x) est invariant par T_φ , pour toute bijection φ , c'est à dire que

$$w_{T_\varphi(x)} = w_x \quad \forall x \in V.$$

On associe à G un réseau résistif en considérant que la résistance de chaque arête est égale à 1. On définit la racine comme $o = 000\dots 0$, et le point γ (unique point de la frontière hors o) comme $\gamma = 11\dots 1$. On cherche à estimer la résistance entre o et γ (autrement dit entre o et $\Gamma = \{\gamma\}$, pour reprendre les notations du cours). On note p la solution du problème de Dirichlet

$$\begin{cases} \sum_{y \sim x} (p_x - p_y) = 0 & \forall x \in \mathring{V}, \\ p_o = 0, \\ p_\gamma = 1. \end{cases} \quad (20.5)$$

4) a) Rappeler brièvement les arguments qui permettent d'assurer qu'il existe une solution unique à ce problème.

- b) Montrer que, pour toute bijection φ , le champ $p \circ T_\varphi$ est aussi solution du problème de Dirichlet.
- c) En déduire que p est nécessairement invariant par T_φ , pour tout φ .
- d) En déduire que p est constant sur chaque sphère $S(o, k)$ (ensemble des sommets x tel que $w_x = k$).
- e) Pour $0 \leq k \leq n - 1$, quel est le nombre d'arêtes entre un sommet donné de $S(o, k)$ et $S(o, k + 1)$?
- f) En déduire le nombre total d'arêtes reliant n'importe quel sommet de $S(o, k)$ à n'importe quel sommet de $S(o, k + 1)$?

5) Montrer que l'on peut estimer la résistance équivalente en remplaçant chaque "étage" de l'hypercube, i.e. chaque $S(o, k)$, par un sommet unique noté s_k , de telle sorte que le réseau résultant soit linéaire

$$o \longleftrightarrow s_1 \longleftrightarrow \dots \longleftrightarrow s_n = \gamma.$$

Préciser le nombre d'arêtes de résistance 1 reliant s_k à s_{k+1} , et en déduire la résistance équivalente du n -hypercube entre les deux points diamétralement opposés o et γ .

6) On s'est intéressé dans ce qui précède à la résistance entre deux sommets diamétralement opposés, on cherche ici à définir et estimer la résistance entre deux *hyperfaces*.

- a) Proposer une définition de la notion de résistance entre deux sous-parties de sommets disjointes (sans supposer que l'une est réduite à un singleton comme on l'a fait dans le cours).
- b) On définit, pour $\alpha = 0$ ou $\alpha = 1$, les hyperfaces

$$\Gamma_\alpha = \{x = (a_1, a_2, \dots, a_{n-1}, \alpha), a_i \in \{0, 1\} \quad \forall i = 1, \dots, n-1\}.$$

Estimer la résistance équivalente du réseau entre Γ_0 et Γ_1 .

20.4 Clivage

On représente un réseau social par un graphe orienté $G = (V, E)$. On considère ici un modèle discret pour lequel les influences sont susceptibles de varier au cours du temps. Plus précisément, on cherche à intégrer au modèle le fait que, si une personne a une opinion trop éloignée de la notre, l'influence qu'elle exerce sur nous est susceptible de diminuer. On se donne une collection $u^0 = (u_x^0) \in \mathbb{R}^V$ d'opinions initiales et une collection $(K_{xy}^0) \in \mathbb{R}^E$ d'influences initiales, avec

$$\sum_{x \rightarrow y} K_{xy}^0 = 1 \quad \forall x \in V, \quad E = \text{supp } K^0 = \{(x, y), K_{xy}^0 > 0\}.$$

On propose le problème d'évolution discret en temps sur le couple (u, K) défini par

$$\begin{cases} u_x^{n+1} = \sum_{y \leftarrow x} K_{xy}^n u_y^n & \forall x \in V \\ K_{xy}^{n+1} = \frac{K_{xy}^n \exp(-\beta |u_x^n - u_y^n|)}{\sum_{y' \leftarrow x} K_{xy'}^n \exp(-\beta |u_x^n - u_{y'}^n|)} & \forall (x, y) \in E, \end{cases} \quad (20.6)$$

avec $\beta > 0$. On écrira de façon plus concise

$$(K^{n+1}, u^{n+1}) = \Psi(K^n, u^n).$$

1) a) Montrer que la matrice K^n est stochastique pour tout $n \geq 0$, c'est à dire que, pour tout n ,

$$K_{xy}^n \geq 0 \quad \forall (x, y) \in E, \quad \sum_{x \rightarrow y} K_{xy}^n = 1 \quad \forall x \in V,$$

et que le support de K reste égal à E au cours des itérations, c'est à dire que, pour tout $(x, y) \in V \times V$, et tout $n \in \mathbb{N}$, on a $K_{xy}^n > 0$ si et seulement si $(x, y) \in E$.

b) Expliquer pourquoi ce système est de nature à réduire l'influence sur tout point x des personnes qui ont une opinion éloignée de celle de x , par rapport à celles qui ont une opinion voisine. Préciser le rôle du paramètre $\beta > 0$ dans le modèle, et décrire le comportement du modèle lorsque β tend vers 0, et lorsque β tend vers $+\infty$.

2) Montrer que l'opinion (u^n) vérifie le principe du maximum suivant :

$$u_x^n \in [\min u^0, \max u^0].$$

3) On suppose ici que le graphe est “bouclé” (i.e. $(x, x) \in E$ pour tout $x \in V$).

Pour toute partition

$$V_1 \cup \dots \cup V_p = V \quad (\text{union disjointe})$$

de l'ensemble des sommets, toute collection d'opinions U_1, \dots, U_p , montrer qu'il existe un champ K d'influences tel que (u, K) est point d'équilibre¹, c'est-à-dire que $(K, u) = \Psi(K, u)$, où u est le champ d'opinion uniforme (valeur U_j) sur chaque sous-communauté V_j .

Que peut-on dire dans le cas où l'on ne suppose plus que le graphe contient toutes les boucles ?

4) On cherche à évaluer la stabilité des situations clivées évoquées dans la question précédente, dans un cadre simplifié. On considère un graphe à 2 sommets indexés par x et y , et E est le graphe complet.

a) On considère la situation “éclatée” $u = (1, -1)$, $K_{xx} = K_{yy} = 1$. Vérifier que ce couple (u, K) est point d'équilibre du système.

On considère une perturbation du point d'équilibre ci-dessus, en considérant que chacun des 2 individus, qui ne s'écoutait que lui-même, se met brusquement à écouter un peu l'autre : $u^0 = (1, -1)$, $K_{xx}^0 = K_{yy}^0 = 1 - \varepsilon$.

b) Montrer que la situation reste symétrique au cours des itérations, c'est-à-dire que, pour tout n

$$u^n = (u_x^n, -u_x^n), \quad K_{xx}^n = K_{yy}^n.$$

c) Montrer que le système tend au départ à revenir spontanément à une situation clivée, plus précisément que $K_{xx}^1 > 1 - \varepsilon$.

On cherche maintenant à montrer que, au delà de cette tendance immédiate à revenir à une situation clivée, le système évolue effectivement, si la perturbation n'est pas trop grande, vers un nouveau clivage, potentiellement moins marqué.

d) On note pour simplifier $K^n = K_{xx}^n$.

Montrer que, tant que u_x^n reste supérieur à $1/2$, et pour $\varepsilon \in [0, 1/2]$, on a

$$K^n \geq 1 - 2\varepsilon\alpha^n, \quad \text{avec } \alpha = \exp(-\beta).$$

1. On admettra ici que K_{xy} puisse être nul pour certains $(x, y) \in E$.

(Indication : on pourra chercher à établir une majoration de $P^n - 1$, avec $P^n = 1/K^n$.)

e) En déduire que, si ε est assez petit, le système évolue vers une autre situation clivée, du type

$$u^\infty = (u_x^\infty, -u_x^\infty), \quad K_{xx}^\infty = K_{yy}^\infty = 1,$$

avec $u_x^\infty \in]0, 1[$.

5) On cherche maintenant à montrer l'instabilité des situations de compromis “tendu”, avec des agents complètement influençables, qui sont influencés par des personnes qui ont des opinions différentes. On cherche plus précisément à montrer que, à la moindre perturbation, l'opinion du suiveur va basculer vers l'une ou l'autre des opinions extrêmes. On considère une situation simplifiée à 3 agents : x est l'influencé, et y et z sont les influenceurs. On considère la topologie suivante

$$x \rightarrow y, \quad x \rightarrow z, \quad y \rightarrow y, \quad z \rightarrow z,$$

avec $K_{xy} = K_{xz} = 1/2$. On considère la situation de référence $u_y = -1, u_z = 1, u_x = 0$.

a) Montrer la situation ci-dessus correspond à un point d'équilibre

b) Vérifier l'instabilité de cette configuration en décrivant le plus précisément possible l'évolution du système à partir de la condition initiale perturbée $K_{xy} = 1/2 - \varepsilon, K_{xz} = 1/2 + \varepsilon$.

20.5 Charisme

On représente un réseau social par un graphe orienté pondéré $G = (V, E, K)$, où $K = (K_{xy}) \in \mathbb{R}_+^{V \times V}$ est une matrice stochastique, i.e. telle que

$$\sum_{y \leftarrow x} K_{xy} = 1 \quad \forall x \in V.$$

Étant donnée une collection $u^0 = (u_x^0) \in \mathbb{R}^V$ d'opinions initiales, on considère le problème d'évolution discret en temps

$$u_x^{k+1} = \sum_{x \rightarrow y} K_{xy} u_y^k.$$

On ne considèrera dans la suite (en particulier pour les exemples demandés) que des graphes *connexes*.

1) a) On se donne une mesure de probabilité $m = (m_x) \in]0, +\infty[^V$ sur V , avec $\sum m_x = 1$. Montrer que l'on a, pour tout k ,

$$\bar{u}^k = \sum_{x \in V} m_x u_x^k \in [\min(u^0), \max(u^0)].$$

b) Montrer, en donnant un exemple (on précisera le réseau, l'opinion initiale, et la collection de poids), que la quantité \bar{u}^k peut converger vers une certaine valeur quand k tend vers $+\infty$ sans que la suite des u^k ne soit convergente.

2) a) Donner un exemple de graphe pour lequel il n'existe aucune mesure de probabilité $m = (m_x) \in]0, +\infty[^V$ (attention, on demande que les masses soient toutes strictement positives) telles que la quantité

$$\sum_{x \in V} m_x u_x^k$$

se conserve au cours des itérations pour toute condition initiale.

b) Montrer que, si l'on n'impose pas la stricte positivité des masses, on peut avoir plusieurs mesures pour lesquelles la moyenne associée est préservée.

On suppose maintenant le réseau de type charismatique, c'est à dire qu'il existe une collection de poids $m = (m_x) \in]0, +\infty[^V$ telle que

$$m_x K_{xy} = m_y K_{yx} = C_{xy} \quad \forall (x, y) \in E,$$

(la matrice $C = (C_{xy})$ est donc symétrique par construction). On supposera que la somme des charismes m_x vaut 1, de telle sorte que $m = (m_x)$ est une mesure de probabilité sur V . On suppose de plus le réseau *connexe*.

3) Expliquer pourquoi il est licite de remplacer, dans les sommations sur les voisins, ' $y \leftarrow x$ ' par ' $y \sim x$ ', et montrer que (m_x) et (C_{xy}) sont liés par la relation

$$m_x = \sum_{y \sim x} C_{xy} \quad \forall x \in V.$$

4) Rappeler pourquoi la moyenne pondérée des opinions

$$\bar{u}^k = \sum_{x \in V} m_x u_x^k$$

se conserve exactement au cours des itérations. On notera \bar{u} cette moyenne indépendante de k .

On revient maintenant au cas d'un réseau charismatique (avec une matrice M diagonale), comme précédemment.

5) Montrer que l'opinion d'un individu ne peut peser plus que pour la moitié des opinions dans l'opinion moyenne, plus précisément que, pour tout $x \in V$, on a

$$m_x \leq \sum_{y \sim x} m_y,$$

et en déduire que $m_x \leq 1/2$ (on rappelle que les C_{xy} ont été normalisés de façon à ce que la somme des m_x soit égale à 1).

On se place dans la position d'un individu x qui souhaite augmenter le poids de sa propre opinion dans l'opinion moyenne \bar{u} (relativement au champ des charismes). Partant d'un réseau charismatique défini par sa matrice $C = (C_{xy})$, avec

$$m_x = \sum_{y \sim x} C_{xy}, \quad K_{xy} = \frac{C_{xy}}{m_x},$$

on considère que x rajoute au graphe un point \tilde{x} (sorte de *jumeau virtuel* de x qu'il se crée lui-même), qui n'est relié qu'à lui, et qui calque son opinion initiale sur la sienne, i.e. $u_{\tilde{x}}^0 = u_x^0$. Cette nouvelle arête est dotée d'une "conductance" $C_{x\tilde{x}} = \beta > 0$. On crée ainsi un nouveau graphe non orienté pondéré sur $V \cup \{\tilde{x}\}$, qui définit un nouveau réseau charismatique.

6)a) On considère la situation initiale : x pense 1 (et donc son jumeau \tilde{x} aussi), et tous les autres pensent 0. Montrer que la valeur moyenne pondérée \bar{u} tend vers 1 lorsque β tend vers $+\infty$.

b) Expliquer en quoi ce comportement n'est pas en contradiction avec le fait qu'un individu ne peut compter au maximum que pour moitié dans l'opinion moyenne (question 5).

7) Décrire ce qui se passe si plusieurs individus (mais pas tous!), dont on désigne l'ensemble par $\Gamma \subset V$, décident chacun de se créer un jumeau virtuel, sur la base du même paramètre $\beta > 0$, destiné à tendre vers $+\infty$. On précisera en particulier vers quoi tend la valeur moyenne pondérée associée \bar{u} , et l'on décrira aussi précisément que possible l'évolution du système aux premiers instants.

On considère maintenant un modèle d'évolution continu en temps, et non linéaire :

$$\frac{du_x}{dt} = -\frac{1}{\eta} \sum_{y \sim x} K_{xy} \varphi(u_x - u_y), \quad (20.7)$$

où φ est une fonction continue de \mathbb{R} dans \mathbb{R} , impaire, et positive sur \mathbb{R}_+ . On fait comme précédemment une hypothèse de structure charismatique pour ce problème non linéaire : il existe m_1, \dots, m_N strictement positifs tel que

$$m_x K_{xy} = m_y K_{yx} = C_{xy}.$$

8) a) Commenter les hypothèses faites sur φ en termes de modélisation, et donner un exemple d'une telle fonction (qui ne soit pas simplement $\varphi(v) = v$, qui correspondrait alors au modèle linéaire) qui vous paraîtrait pertinente (on pourra proposer plusieurs fonctions possibles, en précisant les choix de modélisation qu'elles expriment).

b) Montrer que l'équation d'évolution est de type flot de gradient pour la métrique associée à $M = \text{diag}(m_x)$, c'est à dire qu'il s'écrit

$$\frac{du_x}{dt} = -\frac{1}{\eta} \nabla^M \Phi(u) = -\frac{1}{\eta} M^{-1} \nabla \Phi(u), \quad M = \text{diag}(m_x),$$

pour une certaine fonctionnelle Φ que l'on explicitera.

(On pourra rappeler la fonctionnelle dans le cas où φ est l'identité, et généraliser ce cas linéaire en introduisant une primitive ψ de φ .)

9) Montrer que l'opinion moyenne (relativement à la collection de charismes), c'est à dire la quantité

$$\bar{u} = \sum_x m_x u_x$$

se conserve au cours de l'évolution.

10) On se place dans un cas où l'opinion $u = (u_x)$ converge vers un consensus, i.e. toutes les opinions convergent vers une même valeur α . Exprimer α en fonction de la donnée initiale $u^0 = (u_x^0)$.

11) Montrer la moyenne quadratique des opinions relativement à m décroît, c'est à dire que

$$\frac{d}{dt} \left(\sum_x m_x u_x^2 \right) \leq 0.$$

20.6 Stabilité du poumon humain

Ce problème porte sur une version abstraite du poumon humain. Comme la structure d'arbre ne joue aucun rôle ici, nous nous plaçons dans le cadre d'un réseau de topologie quelconque. Précisons que, malgré le caractère très abstrait et général de l'approche proposée ici, l'étude de stabilité du modèle couplé arbre-alvéoles correspond à une problématique réelle, qui est l'une des premières causes non viabilité des nouveaux-nés prématurés.

On considère un réseau résistif enraciné (V, E, r, o, Γ) , connexe, avec V fini. La racine o est un point quelconque du réseau, et Γ une partie non vide de V , appelée frontière, ne contenant pas la racine o . On rappelle les notations suivantes : pour tout $u \in \mathbb{R}^E$ antisymétrique, $(\partial u)_x = \sum_{y \sim x} u_{yx}$ (divergence discrète), et pour tout $p \in \mathbb{R}^V$, $(\partial^* p)_{xy} = p_y - p_x$ (gradient discret). Les points intérieurs au réseau forment un ensemble (supposé également non vide) noté $\mathring{V} = V \setminus \Gamma$. Un champ $p_\Gamma \in \mathbb{R}^\Gamma$ étant donné sur la frontière, on considère le problème aux limites discret

$$\begin{cases} \partial c \partial^* p(x) &= 0 \quad \forall x \in \mathring{V}, \\ p(o) &= 0 \\ p(x) &= p_\Gamma(x) \quad \forall x \in \Gamma, \end{cases} \quad (20.8)$$

où $c = 1/r \in]0, +\infty[^E$ est le champ de conductance, et

$$\partial c \partial^* p(x) = \sum_{y \sim x} c(x, y)(p(x) - p(y)).$$

On rappelle que ce problème est bien posé, et l'on note $p \in \mathbb{R}^V$ la solution. On note $u = -c\partial^* p$ le champ de flux associé (loi de Poiseuille), et l'on appelle S l'application qui à $p_\Gamma \in \mathbb{R}^\Gamma$ associe le champ des flux rentrants sur Γ , champ qui prend la valeur $-\partial u(x)$ en tout point x de Γ . L'espace \mathbb{R}^Γ est identifié à l'espace euclidien standard \mathbb{R}^n muni du produit scalaire canonique (où n est le cardinal de Γ), mais nous conservons la notation \mathbb{R}^Γ car il n'y a pas lieu de numérotter les sommets.

1) Montrer que S est un opérateur linéaire de \mathbb{R}^Γ dans lui-même, symétrique défini positif, c'est à dire que $S p_\Gamma \cdot q_\Gamma = p_\Gamma \cdot S q_\Gamma$ pour tous $p_\Gamma, q_\Gamma \in \mathbb{R}^\Gamma$, et que $p_\Gamma \neq 0 \implies S p_\Gamma \cdot p_\Gamma > 0$.

On considère maintenant que chaque sortie de Γ débouche sur un petit ballon sphérique (que nous appellerons alvéole) de volume w_x , de façon étanche et incompressible, c'est-à-dire que toute diminution (resp. augmentation) de volume est exactement compensée par une entrée (resp. sortie) du même volume dans le réseau au travers de x . Toutes ces alvéoles sont plongées dans un milieu extérieur porté à une pression uniforme $P \in \mathbb{R}$, susceptible de varier au cours du temps.

On prend en compte le phénomène dit de tension surfacique en considérant que le saut de pression entre l'intérieur et l'extérieur est proportionnel à l'inverse du rayon de courbure de la sphère, i.e. proportionnel à $1/w_x^{1/3}$. La pression à l'intérieur de l'alvéole attachée en x s'écrit donc $\kappa/w_x^{1/3} + P$, où κ est une constante strictement positive. On considère que les pressions et flux sur Γ sont reliés par l'opérateur S introduit précédemment. Le système s'écrit donc

$$\frac{dw}{dt} = -S \left(\frac{\kappa}{w^{1/3}} + P(t) \right) \quad (20.9)$$

où $w = (w_x)_{x \in \Gamma}$ est le vecteur des volumes, et où, par commodité, on a noté $\kappa/w^{1/3} + P(t)$ le vecteur de \mathbb{R}^Γ qui prend la valeur $\kappa/w_x^{1/3} + P(t)$ en x .

- 2) a) Énoncer un résultat d'existence et d'unicité de solution portant sur la collection des volumes $w \in U =]0, +\infty[^{\Gamma}$, en précisant bien les hypothèses faites. Les solutions sont elles en général globales (i.e. définies sur l'intervalle de temps $[0, +\infty[$) ?
- b) Proposer une modification du modèle qui permettrait de garantir l'existence de solutions définies globalement (on ne cherchera pas à justifier le choix en termes de modélisation).
- c) Montrer que, si l'on fixe $P(t)$ à une valeur constante strictement négative, l'équation admet un unique point d'équilibre. Montrer que ce point d'équilibre est *instable*.

3) (Modèle nez bouché)

On cherche ici à montrer que cette instabilité native n'est pas due à la seule présence d'un sommet (la racine) du réseau ouvert sur le monde extérieur. Adapter la démarche proposée depuis le début de ce problème au cas "sans racine o", c'est à dire que les seuls points échangeant du fluide avec l'extérieur du réseau sont les points de Γ . On s'attachera en particulier à préciser les propriétés de l'opérateur S défini dans ce nouveau cadre, et l'on montrera que les points d'équilibre du système couplé avec les alvéoles sont toujours instables.

[Ce qui précède semble indiquer que le système respiratoire est intrinsèquement instable, et de fait il le serait sans un ingrédient supplémentaire appelé surfactant, qui a pour effet de modifier le coefficient de tension surfacique lorsque le volume varie. Ce point fait l'objet de la question suivante.]

- 4) On se replace dans le cas d'un réseau avec racine o . On suppose maintenant que le coefficient κ est susceptible de dépendre du volume, de telle sorte que la pression dans l'alvéole reliée à $x \in \Gamma$ s'écrit $\kappa(w_x)/w_x^{1/3} + P$. On considère une situation où tous les volumes sont les mêmes, i.e. $w_x = W > 0$, et l'on fixe la pression à une valeur constante de façon à ce qu'il s'agisse d'un point d'équilibre ($P = -\kappa(W)/W^{1/3}$).

Montrer que, sous certaines conditions sur la fonction $\kappa(\cdot)$ au voisinage de W (conditions que l'on précisera), le point d'équilibre est asymptotiquement stable.

On montre dans la question suivante qu'il est optimal en termes énergétiques d'exercer une pression

uniforme sur Γ pour faire passer un flux donné au travers du réseau. Cette question porte sur le réseau seul, non couplé aux alvéoles.

5) Soit $\Phi \in \mathbb{R}$ une valeur de flux fixée. On note H_Φ l'ensemble des $p_\Gamma \in \mathbb{R}^\Gamma$ tels que, si l'on note $p \in \mathbb{R}^V$ la solution du problème (20.8) associée, le flux $du(o)$ sortant de o est égal à Φ . On note

$$J(p_\Gamma) = \frac{1}{2} \sum_E c(x, y)(p(x) - p(y))^2$$

la puissance dissipée au sein du réseau (au facteur 1/2 près).

a) Montrer que J admet un minimiseur unique p_Γ sur H_Φ .

b) Montrer que (sans aucune hypothèse de symétrie sur le réseau) le minimiseur de la question précédente est uniforme sur Γ .

6) (Mesure harmonique)

On considère le problème de Dirichlet (20.8), avec un champ p_Γ supposé uniforme : $p_\Gamma(x)$ est pris égal à $P_\Gamma \in \mathbb{R}$, valeur commune à tous les $x \in \Gamma$. On note u le champ de flux associé.

a) Montrer qu'il existe une unique valeur de P tel que $-du(o)$ (flux rentrant en o) soit égal à 1.

b) On note $p \in \mathbb{R}^V$ le champ de pression correspondant à la valeur P de la question précédente. Montrer que $(du(x))_{x \in \Gamma}$ est une loi de probabilité sur Γ , qui charge tous les points de Γ (i.e. $du(x) > 0$ pour tout x de Γ).

c)(*) On considère une marche aléatoire sur le réseau, issue de o , avec des probabilités de transition $x \rightarrow y$ données par $c(x, y)/C(x)$ pour tout y relié à x , où $C(x)$ est une constante de normalisation. Montrer que, pour tout $x \in \Gamma$, la probabilité que le point de rencontre de cette marche avec Γ soit x est égale à $\partial u(x)$.

20.7 Poumon non linéaire

On considère un réseau résistif enraciné (V, E, r, o, Γ) , connexe, avec V fini. Les résistances $r(e)$ associées aux arêtes sont supposées appartenir à $]0, +\infty[$, il en est donc de même pour les conductances $c(e) = 1/r(e)$.

On rappelle que la résistance effective de ce réseau est définie par $\bar{R} = 1/\partial u(o)$, où u est le champ de flux $u = -c\partial^* p$, avec p défini comme solution du problème de Dirichlet

$$\begin{cases} \partial c\partial^* p(x) = 0 & \forall x \in \mathring{V}, \\ p(o) = 0, \\ p(x) = 1 & \forall x \in \Gamma. \end{cases}$$

1) On cherche à montrer que cette résistance peut être définie comme minimum de l'énergie dissipée pour les champs de flux conservatifs qui réalisent un flux unitaire au travers du réseau. Plus précisément, on définit²

$$V_1 = \left\{ u \in \mathbb{R}^E, \partial u(o) = \sum_{y \sim o} u(y, o) = 1, \partial u(x) = 0 \quad \forall x \in \mathring{V} \right\}$$

et

$$\mathcal{P}(u) = \sum_E r(e)u(e)^2.$$

2. On rappelle que, comme dans le cours, les flux $u \in \mathbb{R}^E$ sont supposés antisymétriques, i.e. $u(y, x) = -u(x, y)$, avec la convention usuelle de sommation sur l'ensemble des arêtes (on ne compte qu'une fois chaque arête dans les sommes sur E).

a) Montrer que la fonctionnelle $\mathcal{P}(u)/2$ admet un minimiseur unique sur V_1 , et écrire les conditions d'optimalité sous contrainte à l'aide de multiplicateurs de Lagrange $(p(x))_{x \in \mathring{V} \cup \{o\}}$ associés aux contraintes (on pourra considérer p comme un élément de \mathbb{R}^V , avec $p_x = 0$ pour tout $x \in \Gamma$).

b) Montrer que l'on a

$$\bar{R} = \min_{V_1} \mathcal{P}(u) = \min_{V_1} \sum_E r(e)u(e)^2,$$

2) a) Montrer que la conductance effective $\bar{C} = 1/\bar{R}$ peut être définie comme minimum de

$$\sum_E c(e)(p(x) - p(y))^2,$$

sur

$$K_1 = \{p \in \mathbb{R}^V, p(o) = 0, p(x) = 1 \quad \forall x \in \Gamma\}.$$

b) En déduire que la conductance effective est une fonction concave de la collection des conductances $c \in]0, +\infty[^E$. Est-elle strictement concave ?

3) On s'intéresse à la manière dont la conductance globale \bar{C} dépend des conductances locales $c(e)$, que l'on suppose toujours strictement positives. On écrit ainsi

$$\bar{C}(c) = J(c, p_c), \text{ avec } J(c, p) = \sum_E c(e)(p(x) - p(y))^2,$$

où p_c est la solution du problème de minimisation associé à c (question 2a).

a) Montrer que l'application $c \mapsto p_c$ est différentiable. On notera $D_c p_c \in \mathcal{L}(\mathbb{R}^E, \mathbb{R}^V)$ sa différentielle.

b) Montrer que le gradient de \bar{C} (comme fonction de c) s'écrit

$$\nabla \bar{C} = \nabla_c J + (D_c p_c)^\star \nabla_p J.$$

c) En déduire que, pour tout e , la dérivée partielle de \bar{C} par rapport à $c(e)$ est égale à

$$(p_c(x) - p_c(y))^2.$$

d) Donner un exemple de réseau pour lequel il existe une arête e telle que la dérivée de \bar{C} par rapport à $c(e)$ soit nulle.

On s'intéresse maintenant au cas où la loi de Poiseuille (ou d'Ohm pour le cas d'un réseau électrique) est remplacée par une relation non linéaire entre le flux au travers d'une arête et le saut de pression entre ses extrémités. Plus précisément on suppose que

$$u(x, y) = c(x, y)\varphi(p(x) - p(y)) = c(x, y)\varphi(-\partial^\star p(x, y)),$$

où φ est une fonction continûment différentiable sur \mathbb{R} , impaire, et de dérivée strictement positive partout.

On suppose toujours que les flux vérifient la loi des noeuds sur les points de \mathring{V} (points intérieurs). Pour tout champ P donné sur Γ ($P \in \mathbb{R}^\Gamma$), on appelle problème de Dirichlet le problème consistant à rechercher $p \in \mathbb{R}^V$, tel que

$$\left\{ \begin{array}{lcl} \sum_{y \sim x} c(x, y)\varphi(p(x) - p(y)) & = & 0 \quad \forall x \in \mathring{V}, \\ p(o) & = & 0, \\ p(x) & = & P(x) \quad \forall x \in \Gamma. \end{array} \right. \tag{20.10}$$

3) (*Principe du max pour le problème non linéaire*)

Montrer que si un champ $p \in \mathbb{R}^V$ est solution de (20.10), alors le maximum et le minimum de p sont atteints sur le bord $\Gamma \cup \{o\}$.

4) Montrer que le problème de Dirichlet (20.10) exprime les conditions nécessaires d'optimalité pour un problème de minimisation d'une certaine fonctionnelle sur un certain ensemble (on précisera la fonctionnelle et l'ensemble).

5) Montrer que le problème de minimisation de la question précédente admet un unique minimiseur, et en déduire (en détaillant soigneusement le raisonnement suivi) l'existence et l'unicité d'une solution au problème (20.10).

6) Proposer (sans forcément rentrer dans les détails ni dans l'analyse) une (ou plusieurs) stratégie(s) qui permettrai(en)t de calculer effectivement une approximation du problème (20.10).

7) On considère le problème de Dirichlet associé à une donnée constante sur Γ . On note toujours P (qui désigne donc maintenant un réel) cette valeur constante, $p \in \mathbb{R}^V$ la solution associée, et $u = c\varphi(-\partial^* p) \in \mathbb{R}^E$ le flux correspondant. On note

$$F(P) = \partial u(o) = \partial c\varphi(-\partial^* p)(o)$$

le flux sortant (i.e. compté positivement quand le fluide sort effectivement par la racine).

a) On suppose $P > 0$. Montrer qu'alors le champ p associé vérifie $p(x) \in]0, P[$ pour tout $x \in \mathring{V}$, et en déduire que $F(P) > 0$.

b)(*) Montrer que la fonction $P \mapsto F(P)$ est strictement croissante.

(*Indication : on pourra admettre que la fonction est différentiable et montrer, en utilisant la stratégie du problème adjoint, que sa dérivée est strictement positive pour toute valeur de P . Autre approche possible : utiliser le théorème des fonctions implicites.*)

c) On définit la conductance équivalente du réseau (qui dépend maintenant de P du fait du caractère non linéaire du système) par

$$\overline{C} = \overline{C}(P) = \frac{F(P)}{P}.$$

Montrer que cette conductance peut s'exprimer

$$\overline{C}(P) = \frac{1}{P^2} \sum_{e=(x,y) \in E} c(x, y)(p(x) - p(y))\varphi(p(x) - p(y)),$$

où $p \in \mathbb{R}^V$ est la solution du problème (20.10) associé à la condition aux limites $p \equiv P$ sur Γ . Comment peut-on interpréter la quantité en facteur de $1/P^2$ dans l'expression ci-dessus ?

d) Que peut-on conjecturer sur la valeur $\overline{C}(0)$ de la conductance au voisinage de 0 ? (une justification rigoureuse n'est pas demandée, mais les plus courageux pourront chercher à démontrer le résultat conjecturé).

8) (*) Faites le bilan des endroits, dans l'établissement des propriétés établies précédemment, auxquels les différentes hypothèses sur φ ont été utilisées, et discuter de la possibilité d'affaiblir ou supprimer ces hypothèses.

Correction à mettre en forme :

20.8 Phénomène de limitation du débit expiratoire

Contexte : on s'intéresse ici, dans le cadre des modèles de ventilation de type EDO, à la prise en compte du fait que la résistance du “tuyau” est susceptible de varier. On s'intéresse en particulier au phénomène appelé Expiratory Flow Limitation (EFL dans la suite) : dans certaines situations, le débit d'air à l'expiration que l'on peut produire en exerçant une forte pression positive peut culminer à une valeur maximale pour une certaine valeur de pression, de telle sorte qu'une augmentation de pression au-delà de cette valeur soit contre-productive.

- 1) On considère dans un premier temps un système ballon + tuyau plongé dans un milieu à pression uniforme $P > 0$. On néglige pour l'instant le caractère élastique du ballon, considérant que la pression à l'intérieur du ballon est P . On considère que l'écoulement dans le tuyau obéit à une loi de type Poiseuille mais que, du fait de la déformabilité du tuyau (qui représente une bronche typique), la résistance est une fonction de la pression. La loi de Poiseuille s'écrit donc

$$P - 0 = R(P)Q.$$

On suppose la fonction $P \mapsto R(P)$ continument différentiable, croissante, et strictement convexe sur $[0, +\infty[$, avec $R(0) = R_0 > 0$.

- a) Expliquer pourquoi l'hypothèse de croissance est naturelle dans le contexte considéré.
- b) Décrire aussi précisément que possible les conditions sur $P \mapsto R(P)$ susceptibles de conduire à un effet EFL. (On pourra étayer la réponse par des illustrations et des exemples de fonctions pour lesquelles on a, ou on n'a pas, l'effet EFL)
- c) Expliquer pourquoi la présence ou non de l'effet EFL ne dépend pas de la viscosité du fluide impliqué dans l'écoulement.

On s'intéresse maintenant au modèle complet, avec ballon élastique, en considérant que les propriétés élastiques du ballon et du tuyau sont analogues, de telle sorte qu'ils se déforment de façon identique, et que l'on peut ainsi écrire la résistance comme fonction du volume. On négligera par ailleurs l'incidence sur le modèle du fait que le volume d'air dans le tuyau est susceptible de varier au cours du temps. On considère donc le modèle

$$R(V)\dot{V} + E(V - V_0) = -P(t), \quad V(0) = V_{init}, \quad (\star)$$

avec une élastance $E > 0$, et $V_0 > 0$ un volume à l'équilibre donné. On suppose la fonction

$$V \in]0, +\infty[\mapsto R(V) \in]0, +\infty[$$

continûment différentiable, et décroissante. On suppose par ailleurs que $t \mapsto P(t)$ est une fonction continue de $[0, +\infty[$ dans $[P_{min}, +\infty[$ (on s'intéresse à des pressions d'expiration potentiellement importantes, mais des pressions d'inspiration limitées).

- 2) Montrer que, si $V_{init} \leq V_0 - P_{min}/E$, alors toute solution régulière de l'équation vérifie cette même inégalité sur son intervalle de définition.
- 3) a) On suppose que $R(V) = 1/V^\gamma$, avec $\gamma > 1$, et l'on suppose toujours $V_{init} \leq V_0 - P_{min}/E$. Montrer que le problème de Cauchy (\star) admet une unique solution maximale $t \mapsto V(t)$ à valeurs dans $]0, +\infty[$, et que cette solution est globale (i.e. définie pour $t \in [0, +\infty[$).
- b) Que peut-on dire dans le cas où $\gamma \in]0, 1[$?
- 4) Enoncer et démontrer une propriété analogue sur le modèle avec inertie

$$I\ddot{V} + R(V)\dot{V} + E(V - V_0) = -P(t), \quad V(0) = V_{init}, \quad V'(0) = W_{init},$$

où $I > 0$ est un paramètre (appelé inertance) qui quantifie les aspects inertIELS.

20.9 Mobilité et équilibres de Wardrop

On considère une population de personnes dont l'effectif (supposé grand) est représenté par un nombre réel $\mu_0 > 0$, qui représente la “masse” des agents. Ces agents doivent se rendre tous les matins, sur la même plage horaire, d'un point x à un point y (même origine et même destination pour l'ensemble des agents). Ils ont pour cela le choix entre N parcours routiers, indexés par $j = 1, \dots, N$. On note L_j la longueur du trajet j , et l'on suppose que la vitesse v_j à laquelle on parcourt la voie j (supposée constante) dépend du nombre de personnes μ_j qui l'empruntent (ralentissement résultant de la congestion). Plus précisément on suppose

$$v_j = v_j(\mu_j) = V_j \left(1 - \frac{\mu_j}{\bar{\mu}_j}\right)$$

où $\bar{\mu}_j$ est un seuil de saturation (tout se fige quand μ_j atteint cette valeur, qui ne peut être dépassée), et V_j correspond à la vitesse maximale autorisée. On notera $T_j = L_j/V_j$ le temps de parcours à vide, et t_j le temps de parcours effectif, qui s'écrit donc

$$t_j = t_j(\mu_j) = T_j \left(1 - \frac{\mu_j}{\bar{\mu}_j}\right)^{-1}.$$

On note

$$K = \left\{ \mu = (\mu_1, \dots, \mu_N) \in \mathbb{R}_+^N, \mu_j < \bar{\mu}_j \quad \forall j = 1, \dots, N, \sum_{j=1}^N \mu_j = \mu_0 \right\}$$

l'ensemble des scénarios possibles en termes de distribution des flux sur les différents trajets. Pour $\mu \in K$ donné, on note I_μ le support de μ , i.e. l'ensemble des indices des routes utilisées effectivement pour la distribution μ :

$$I_\mu = \text{supp}(\mu) = \{j, \mu_j > 0\}.$$

0) Montrer que K est non vide si et seulement si

$$\mu_0 < \sum_{j=1}^N \bar{\mu}_j.$$

On fera cette hypothèse dans toute la suite du problème.

1) On introduit le temps moyen de parcours

$$\mu = (\mu_1, \dots, \mu_N) \in K \longmapsto \bar{T}(\mu) = \frac{1}{\mu_0} \sum_{j=1}^N \mu_j t_j = \frac{1}{\mu_0} \sum_{j=1}^N \mu_j T_j \left(1 - \frac{\mu_j}{\bar{\mu}_j}\right)^{-1} = \frac{1}{\mu_0} \sum_{j=1}^N \varphi_j(\mu_j).$$

a) Montrer qu'il existe un unique $\mu = (\mu_1, \dots, \mu_N) \in K$ qui minimise $\bar{T}(\mu)$ sur K .

b) Montrer que les dérivées des φ_j en μ_j sont égales pour les j dans I_μ , i.e.

$$\exists \beta > 0, \varphi'_j(\mu_j) = \beta \quad \forall j \in I_\mu.$$

c) Montrer que I_μ ne dépend de μ qu'au travers de β , plus précisément que

$$I_\beta = \{j, T_j < \beta\}.$$

(N.B. : on remarquera que $\varphi'_j(0) = T_j$.)

d) On suppose que $T_1 < T_2 < \dots < T_N$. Représenter sur un graphique, pour N petit (3 ou 4), les graphes de fonctions φ_i typiques, et représenter sur ce même graphe les minimiseurs associés à différentes valeurs de μ_0 (petite, intermédiaire, et grande).

2) On s'intéresse maintenant au temps de trajet maximum. :

$$\mu = (\mu_1, \dots, \mu_N) \in K \mapsto T^{max}(\mu) = \max_{j \in I_\mu} t_j(\mu_j).$$

a) Montrer que, si $\mu = (\mu_1, \dots, \mu_N) \in K$ minimise T^{max} , alors les temps de parcours associés aux routes utilisées sont les mêmes, i.e. qu'il existe $\lambda > 0$ tel que

$$t_j(\mu_j) = \lambda \quad \forall j \in I_\mu.$$

Montrer que l'on a pour ce minimiseur $I_\mu = \{j, T_j < \lambda\}$.

b) Montrer que si μ vérifie les deux propriétés ci-dessus, alors μ est minimiseur de T^{max} .

c) Montrer que, pour toute valeur de $\lambda > \min(T_j)$, il existe une unique distribution $\mu^\lambda \in \mathbb{R}_+^N$ (sans contrainte sur la masse totale) vérifiant les propriétés de la question (2a).

d) Montrer que la fonction

$$\lambda \in]\min(T_j), +\infty[\mapsto \mu_0^\lambda = \sum_{j=1}^N \mu_j^\lambda.$$

est strictement croissante.

e) On se donne maintenant $\mu_0 > 0$ (inférieur à la somme des $\bar{\mu}_j$, comme précisé en préambule), et l'on considère l'ensemble K associé. Montrer qu'il existe un unique minimiseur de T^{max} sur K .

f) On suppose que $T_1 < T_2 < \dots < T_N$. Comme pour la question (1d), représenter sur un graphique, pour N petit (3 ou 4), les graphes de fonctions φ_i typiques, et représenter sur ce même graphe les minimiseurs associés à différentes valeurs de μ_0 (petite, intermédiaire, et grande).

3) a) Montrer que

$$\min_K \bar{T} \leq \min_K T^{max}.$$

b) Donner un exemple de situation pour laquelle l'inégalité est stricte.

c) Expliquer pourquoi le minimiseur de T^{max} (question 2) correspond à une situation qui a des chances d'être observée en pratique, alors que le minimiseur de \bar{T} (question 1) n'a pas de raison de l'être.

d) Expliquer en quoi chercher à minimiser \bar{T} (par des incitations et / ou des contraintes) peut néanmoins avoir du sens.

4) On s'intéresse ici aux aspects environnementaux, considérant que la quantité de gaz à effet de serre émise est proportionnelle à l'énergie dépensée pour rouler. Si l'on a une route parcourue à vitesse v_j (supposée constante le long du parcours), on suppose que la force de trainée (résistance de l'air et friction mécanique) vaut v_j^2 (on a pris une constante de proportionnalité unitaire). La puissance s'écrit donc $v_i^2 \times v_i = v_i^3$, et l'énergie dépensée sur l'ensemble du trajet j est obtenue en multipliant la puissance par le temps

$$E_j = v_j^3 t_j = v_j^3 \frac{L_j}{v_j} = L_j v_j^2.$$

Sous ces hypothèses (contestable, voir question (4e)), l'énergie totale dépensée (et donc la quantité de CO₂ émis) est donc proportionnelle à

$$E = E(\mu) = \sum_{j=1}^N L_j V_j^2 \mu_j \left(1 - \frac{\mu_j}{\bar{\mu}_j}\right)^2.$$

a) Montrer l'existence d'un minimiseur de E sur \bar{K} (adhérence de K).

b) Dresser le tableau de variation complet³ (avec zones de convexité / concavité) de $x \mapsto x(1-x)^2$.

c) On se place dans le cas où l'on a seulement deux routes, qui sont identiques :

$$V_1 = V_2 = 1, \quad L_1 = L_2 = 1, \quad \bar{\mu}_1 = \bar{\mu}_2 = 1.$$

Décrire l'ensemble des minimiseurs associés à $\mu_0 \in [0, 2]$. (On pourra montrer en particulier qu'il existe 3 régimes distincts, associés chacun à un sous-intervalle de $[0, 2]$.)

d) Décrire aussi précisément que possible l'ensemble des minimiseurs dans des situations plus générales.

e) Critiquer l'approche précédente en terme de réalisme (on pourra en particulier se pencher sur l'estimation de l'énergie dépensée lorsque la densité est proche de la saturation, et au delà sur le fait que la minimisation brutale de E peut conduire à des scénarios complètement irréalistes), et proposer des pistes d'amélioration.

5) (*) On considère maintenant que la vitesse de la route j dépend de l'affluence selon la loi

$$v_j = V_j \left(1 - \frac{\mu_j}{\bar{\mu}_j}\right)^\eta, \quad \eta \in]0, 1].$$

a) Expliquer ce qu'exprime, en termes de modélisation, le fait de prendre η plus petit que 1. A quoi correspond le cas limite $\eta \rightarrow 0^+$?

b) Reprendre l'ensemble des questions précédentes pour ce nouveau modèle, avec $\eta \in]0, 1]$, puis dans le cas limite $\eta = 0^+$.

6) (*) proposer une formulation du problème correspondant à la situation où l'on aurait un continuum de trajets possibles, puis formuler (et si possible analyser) les versions continues des questions 1, 2, 4, et 5. On pourra représenter le continuum de parcours par une variable $\alpha \in [0, 1]$, qui joue le rôle de j dans ce qui précède.

7) (*) Proposer un ou des modèles de comportement effectif des agents au fil des jours, et étudier l'évolution des distributions correspondantes. On pourra se placer au départ dans une situation simplifiée (par exemple $N = 2$), et considérer que les agents, ou une partie d'entre eux, choisissent au jour $k + 1$ la stratégie qui s'est avérée la plus rapide au jour k .

8) (*) Proposer un cadre général permettant de formuler ce type de questions dans le cas où les agents sont distribués sur un ensemble X de sommets d'un graphe (qui représente un réseau de transport), et les destinations sur un ensemble Y de sommets (on peut avoir recouvrement total ou partiel entre X et Y). On pourra considérer que l'on se donne au départ un "plan origine-destination" $(\gamma_{xy}) \in \mathbb{R}_+^{X \times Y}$, où γ_{xy} représente le nombre de personnes (quantifié par un nombre réel) qui veulent se rendre de x à y en empruntant un chemin (il peut y en avoir plusieurs possibles) au travers du réseau.

20.10 Optimisation d'une préparation de concours

Cet énoncé se compose de deux questions précises (1 et 2), et de pistes de réflexion plus ouvertes (A, B, \dots), non ordonnées.

On se place dans la position d'un étudiant qui prépare un concours basé sur N épreuves, de coefficients $0 < C_1 < C_2 < \dots < C_N$. On se donne une fonction d'apprentissage commune à toutes les matières, notée φ , de \mathbb{R}^+ dans \mathbb{R}^+ . La valeur prise par $\varphi(\eta)$ correspond à la note espérée si l'on consacre un temps η à la matière en question. On suppose dans un premier temps φ continûment différentiable, strictement croissante, strictement concave, avec $\varphi(0) = 0$ et $\varphi(+\infty) = 1$ (la note maximale est normalisée à 1). Un étudiant consacrant t_i à la matière i aura selon ce modèle un nombre de points

3. Désolé de vous imposer cet exercice habituellement destiné aux lycéens, mais c'est important pour aborder la question suivante.

$C_i \varphi(t_i)$, de telle sorte que si sa stratégie de préparation est $t = (t_i) \in \mathbb{R}_+^N$, son total espéré est

$$U(t) = U(t_1, \dots, t_N) = \sum_{i=1}^N C_i \varphi(t_i)$$

On suppose que l'étudiant dispose d'un temps maximal T pour préparer son concours. Il s'intéresse donc au problème de maximisation de $U(t)$ (ou, de façon équivalente, de minimisation de $-U(t)$), sous la contrainte sur $t = (t_i)$ d'appartenance à

$$K = \left\{ t = (t_i) \in \mathbb{R}_+^N, S(t) = \sum_{i=1}^N t_i \leq T \right\}.$$

- 1) Montrer l'existence et l'unicité d'une stratégie optimale.
- 2) Déterminer, en justifiant avec soin⁴ la démarche, la stratégie optimale à suivre en fonction du temps total $T \in [0, +\infty[$ disponible.

Extensions

- A) Considérer la situation où la courbe d'apprentissage a une dérivée infinie en 0^+ (on pourra proposer un exemple réaliste de telle fonction, et faire l'analyse du problème d'optimisation sous contrainte).
- B) Le jury du concours se désespère que certains étudiants fassent l'impasse sur certaines épreuves. À la lumière de ce modèle (ou de ses extensions), explorer des pistes permettant de limiter ou supprimer ce phénomène (tout en maintenant autant que possible une différence de poids entre les épreuves), voire de s'assurer que chaque étudiant soit incité à passer un temps minimal t_{min} à la préparation de toutes les épreuves. (N.B. : on gardera en tête que la courbe d'apprentissage reflète à la fois les capacités de l'étudiant et la manière de structurer et de noter l'épreuve.)
- C) Proposer (et analyser) un ou des modèles correspondant à la situation où tous les coefficients sont les mêmes, mais les courbes d'apprentissage sont différentes.
- D) Justifier le fait qu'il puisse y avoir des situations où la courbe d'apprentissage n'est pas concave sur sa partie gauche, et explorer cette situation. (On pourra commencer par considérer le cas de 2 épreuves, dont une est non concave, et s'aider de dessins pour expliquer la démarche d'élaboration d'une stratégie optimale.)
- E) Dans l'esprit de D, justifier que les courbes d'apprentissage puissent prendre des valeurs strictement positives en 0, ou au contraire être identiquement nulles dans un voisinage de 0, et discuter de la manière dont ces situations modifient la démarche de recherche d'un optimum (à la fois d'un point de vue théorique, et de l'estimation effective).
- F) Proposer des stratégies pour approcher numériquement la stratégie optimale du problème initial, et / ou de certaines des extensions proposées ci-dessus.
- G) Concevoir et analyser autant que possible une formulation *continue* du problème, i.e. basée sur une infinité non dénombrable (un continuum) d'épreuves.
- H) Proposer d'autres cadres applicatifs conduisant à ce type de problème, et justifier dans ce nouveau cadre le sens des différentes hypothèses.
(Suggestion : on pourra par exemple penser à une personne qui dispose d'une somme totale à dépenser égale à T , ou à un gouvernement qui, disposant d'un budget total limité, cherche à mettre en place des mesures en vue de limiter les émissions de CO₂, mesures dont les courbes d'efficacité présentent des profils différents, en particulier non concaves pour certaines).

4. On prendra notamment garde au fait que \mathbb{R}_+^N n'est *pas un ouvert*.

20.11 Modèles de foules de type Nash

On cherche à modéliser le mouvement de N personnes en file indienne, identifiées à des disques rigides de rayon $r > 0$, repérés par les positions x_1, \dots, x_N de leurs centres, avec une condition de non chevauchement : on impose que $x = (x_1, \dots, x_N)$ appartienne à

$$K = \{x \in \mathbb{R}^N, x_{j+1} - x_j \geq 2r, j = 1, \dots, N-1\}.$$

Pour $x \in K$, on note C_x l'ensemble des vitesses admissibles (i.e. qui ne conduisent pas à un chevauchement entre disques) :

$$C_x = \{v \in \mathbb{R}^N, x_{j+1} - x_j = 2r \implies v_j - v_{j+1} \leq 0\}.$$

On se donne une collection de vitesses *souhaitées* $U = (U_j)$: la personne j souhaite aller à la vitesse U_j .

I
Modèle granulaire

Le premier modèle est basé sur l'hypothèse que, à chaque instant, la vitesse effective de la collection de personnes est la plus proche possible (au sens des moindres carrés) de la collection des vitesses souhaitées parmi les vitesses admissibles. On s'intéressera essentiellement (sauf dans les questions (6) et (11)) au problème “instantané” : pour $x \in K$ fixé, on cherche à minimiser la fonctionnelle

$$v \mapsto J(v) = \frac{1}{2} \sum |v_j - U_j|^2$$

sur C_x . On considèrera sauf avis contraire un “train” de personnes, i.e. x est tel que $x_{j+1} - x_j = 2r$ pour tout $j \leq N-1$ (il y a donc $N-1$ contacts au total).

1) Montrer que J admet un minimiseur unique sur C_x . On notera $u = P_{C_x} U$ ce minimiseur, qui est donc la projection de U sur C_x pour la norme euclidienne.

2)a) Écrire sous forme matricielle la contrainte d'appartenance à C_x , c'est-à-dire sous la forme $Bv \leq 0$ (inégalité coefficient par coefficient), et montrer l'existence d'un champ $p \in \mathbb{R}_+^{N-1}$, que l'on notera pour des raisons évidentes $p = (p_{j,j+1})_{1 \leq j \leq N-1}$, tel que

$$\begin{array}{rcl} u + B^* p & = & U \\ Bu & \leq & 0 \end{array}$$

avec $u_{j+1} > u_j \implies p_{j,j+1} = 0$.

b) Montrer qu'un tel p est *unique*.

c) De quel opérateur différentiel la matrice B est-elle la version discrète ?

d) Proposer une interprétation des valeurs $p_{j,j+1}$ en termes de modélisation.

3) a) Montrer que $S = \text{supp}(p)$ contient $\{(j, j+1), U_{j+1} < U_j\}$, c'est à dire que, si le couple de vitesses souhaitées pour j et $j+1$ viole la contrainte, alors le multiplicateur de Lagrange $p_{j,j+1}$ est strictement positif⁵.

b) Montrer sur un exemple que l'inclusion ci-dessus peut être stricte, c'est-à-dire que l'on peut avoir une pression non nulle entre deux personnes, alors que le couple de leurs vitesses spontanées ne violait pas a priori la contrainte. Décrire une situation de la vie courante (par exemple dans une file d'attente au cinéma) qui illustre cette propriété.

5. On prendra garde au fait qu'il ne s'agit pas d'une propriété universelle découlant d'un théorème général d'optimisation.

4)a) Montrer que, si tous les $p_{j,j+1}$ sont strictement positifs, alors p est solution de

$$BB^*p = BU.$$

Explicitier la matrice BB^* , et justifier que l'on puisse appeler cette matrice un *laplacien discret*.

b) Décrire un dispositif expérimental (dans le contexte des réseaux électriques résistifs, ou des écoulements de Poiseuille dans un réseau de tuyaux) qui conduirait à un problème analogue.

5)a) Montrer que la projection sur C_x vérifie le principe du maximum suivant :

$$u_j \in [\min U_k, \max U_k] \quad \forall j = 1, \dots, N.$$

b) Montrer que la quantité de mouvement est conservée par projection :

$$\sum_j u_j = \sum_j U_j.$$

c) Montrer que l'énergie cinétique décroît par projection, c'est-à-dire que

$$\sum_j |u_j|^2 \leq \sum_j |U_j|^2.$$

Donner un exemple de situation pour laquelle elle décroît strictement.

6) On s'intéresse dans cette question au problème instationnaire⁶. On se donne une configuration initiale $x^0 = (x_j^0)$ dans l'intérieur de C_x (aucun contact initialement entre les disques), et un champ de vitesses souhaitées $U = (U_j)$ constant (j veut tout le temps avancer à la vitesse U_j fixée au départ). On s'intéresse au système dynamique

$$\frac{dx}{dt} = u = P_{C_x}(U).$$

a) Quelle est la nature de ce modèle ? (lagrangienne ou eulérienne).

b) Expliquer aussi précisément que possible pourquoi ce modèle d'évolution *ne rentre pas* dans le cadre du théorème de Cauchy-Lipschitz.

c) Montrer que, sous certaines hypothèses (que l'on précisera) sur U , le modèle est en fait très simple puisqu'aucun contact ne se produit.

d) Lorsque l'hypothèse ci-dessus n'est pas vérifiée, et que des contacts se produisent, décrire, selon votre intuition, le comportement des solutions de ce problème, selon les vitesses souhaitées initiales. (*On pourra faire un dessin des trajectoires des centres dans l'espace-temps, avec les positions en abscisse, et le temps en ordonnée, en prenant éventuellement $r = 0$ pour simplifier le dessin.*)

7) On associe maintenant une masse $m_j > 0$ à l'individu j , et l'on projette le champ de vitesse souhaitée sur C_x selon la norme pondérée par les masses, ce qui revient à considérer maintenant la fonctionnelle

$$v = (v_1, \dots, v_N) \longmapsto J(v) = \frac{1}{2} \sum m_j |v_j - U_j|^2.$$

a) Quel sens peut on donner à m_j en termes de modélisation ?

b) (*) Reprendre une à une l'ensemble des questions précédentes, et les adapter à ce nouveau contexte.

c) On suppose toutes les masses égales à 1, sauf la j -ième, qui vaut $1/\varepsilon$, avec $\varepsilon > 0$. Pour un champ U donné, on note u^ε la vitesse projetée selon la norme pondérée. Décrire, en le justifiant aussi précisément que possible, le comportement de u^ε lorsque ε tend vers 0.

6. Précisons que l'analyse rigoureuse du problème d'évolution dépasse largement le cadre de ce cours.

II Équilibres de Nash, interactions symétriques

On se place maintenant dans une optique plus “comportementale”. On considère que chaque individu a conscience des vitesses effectives instantanées de ses voisins, et adopte la vitesse la plus proche de sa vitesse souhaitée parmi celles qui sont compatibles avec celles des voisins. Considérer que chaque individu se comporte de la sorte conduit à la notion d'*équilibre de Nash* décrite ci-dessous, dans le contexte particulier des foules. On considère comme précédemment une rangée de N personnes (toujours identifiées à des disques). On se donne pour chaque j une vitesse souhaitée U_j , et l'on appelle équilibre de Nash tout champ de vitesses $v \in \mathbb{R}^N$ tel que

$$v_j = \arg \min_{w \in C_j(v)} \frac{1}{2} |w - U_j|^2,$$

avec

$$C_j(v) = \{w \in \mathbb{R}, v_{j-1} \leq w \leq v_{j+1}\} \subset \mathbb{R}.$$

(Nous laissons au lecteur le soin de définir C_j pour $j = 1$ ou $j = N$, dans les cas où il n'y a qu'un seul voisin). Nous avons écrit C_j comme dépendant de v , mais on prendra garde au fait qu'il ne dépend que des v_k pour k différent de j , et même plus précisément pour $k = j + 1$ et $k = j - 1$. Il est en tout cas essentiel qu'il ne dépende pas de v_j , sinon le problème de minimisation ci-dessus n'aurait pas de sens. On note Λ l'ensemble des équilibre de Nash ainsi définis.

- 8)a) Montrer que la solution du problème de minimisation de la question 1 est un équilibre de Nash, et donc que Λ est non vide.
 - b) Montrer que l'on n'a pas toujours unicité. On pourra considérer par exemple le cas $N = 2$, deux personnes en contact, avec la personne de droite qui veut aller vers la gauche, celle de gauche qui veut aller vers la droite, et montrer qu'il y a dans ce cas une infinité non dénombrable d'équilibres de Nash.
 - c) Montrer que, pour tout j , il y a au moins un équilibre de Nash pour lequel j fait ce qu'il veut, i.e. tel que $u_j = U_j$.
 - d) Montrer qu'un équilibre de Nash ne conserve pas forcément la quantité de mouvement (i.e. la somme des u_j peut être différente de la somme des U_j), et que l'énergie cinétique n'est pas forcément diminuée (i.e. la somme des carrés des u_j peut être strictement supérieure à la somme des carrés des U_j).
 - e) Montrer qu'en revanche le principe du maximum de la question (5a) est préservé.
- 9)a) On se donne comme dans la question (7) une matrice M de masses strictement positives, et l'on note u la solution du problème de minimisation associé. Montrer que, pour tout M , le minimiseur associé u est un équilibre de Nash.
 - b) Tout équilibre de Nash (i.e. tout élément de Λ) est-il un minimiseur du problème granulaire (question 7) pour une certaine matrice de masse M ?
 - 10) On s'intéresse ici au cas périodique : les individus forment une chaîne qui se referme sur elle-même dans un couloir circulaire, de telle sorte que chaque individu a maintenant bien 2 voisins. Décrire l'ensemble des vitesses admissibles et montrer que, quelles que soient les vitesses souhaitées, tout champ de vitesse admissible est un équilibre de Nash.

Cette situation vous paraît-elle pertinente en termes de modélisation ?

III Hiérarchie

On considère maintenant des personnes qui souhaitent toutes aller vers la droite ($U_j \geq 0$), en regardant devant elles, et adaptent leur vitesse à celle de la personne devant seulement. On s'intéresse

donc ici à des champs de vitesse $u = (u_j)$ qui soient tels que chaque personne adopte une vitesse qui est la plus proche de sa vitesse souhaitée au vu de ce que fait la personne devant elle, si elle en a une, c'est à dire à des champs u tels que

$$v_j = \arg \min_{w \in C_j(v)} \frac{1}{2} |w - U_j|^2, \quad j = 1, \dots, N,$$

avec

$$C_j(v) = \{w \in \mathbb{R}, w \leq v_{j+1}\} \text{ pour } j = 1, \dots, N-1, \quad C_N(v) = \mathbb{R}.$$

11) Montrer qu'il existe un unique champ de vitesse, toujours noté $u = (u_1, \dots, u_N)$, qui vérifie les conditions ci-dessus, et proposer une stratégie permettant de le calculer exactement.

12) On s'intéresse dans cette question, comme dans la question (6) pour le modèle granulaire, au problème instationnaire. On suppose données une collection de positions initiales dans l'intérieur de K , ainsi que des vitesses souhaitées U_1, \dots, U_N positives ou nulles, qui restent constantes au cours du temps. Décrire le comportement des solutions du problème d'évolution. On pourra s'intéresser notamment au cas où les vitesses sont décroissantes vis à vis de j , i.e. $U_{j+1} < U_j$ pour $j = 1, \dots, N-1$.

13) Comparer ce modèle d'évolution avec le modèle de trafic étudié en cours, défini par

$$\dot{x}_j = \varphi_j(x_{j+1} - x_j),$$

où φ_j est défini par $w \mapsto \varphi_j(w) = U_j(1 - \exp((w - 2r)_+/\delta))$, où $\delta > 0$ est un paramètre de raideur. On pourra en particulier décrire qualitativement comment se comporte ce modèle lorsque δ tend vers 0.

14) (*) Montrer que la solution u de la question (11) peut-être obtenue comme limite quand ε tend vers 0 de solutions du modèle granulaire (partie I, question 7) associé à des matrices de masse M_ε qui dégénèrent d'une manière que l'on précisera.

IV (*)

Extension à la dimension 2 d'espace (question ouverte)

Proposer une extension des différentes démarches proposées précédemment à la dimension 2 : les personnes sont assimilées à des disques de même rayon $r > 0$, le centre de j est maintenant un point du plan : $x_j \in \mathbb{R}^2$, et les centres sont assujettis à rester à distance $\geq 2r$ les uns des autres. On pourra introduire les vecteurs unitaires associés aux centres de deux grains en contact :

$$e_{ij} = \frac{x_j - x_i}{|x_j - x_i|}.$$

On pourra en particulier s'interroger sur les versions bidimensionnelles des questions (3a), (5a), et (9b), pour lesquelles la situation est très différente de la version monodimensionnelle.

20.12 Mouvement de véhicules autonomes

On s'intéresse à la modélisation du mouvement de N véhicules sans conducteur, que l'on supposera évoluer sur une route circulaire. Les positions des véhicules, désignées par x_1, x_2, \dots, x_N , appartiennent à l'intervalle $[0, L[$ périodisé, avec la convention usuelle⁷ pour désigner la distance entre deux véhicules : si x_i est la position la plus à droite, x_{i+1} est la plus à gauche, et $x_{i+1} - x_i$ désigne en fait $L + x_{i+1} - x_i$. Les véhicules sans conducteur ont des caméras à l'avant et à l'arrière qui leur

⁷. Comme dans le problème I, on considérera comme acquises ces questions de prise en compte de la périodicité dans les notations, sans qu'il soit donc besoin de détailler ces considérations dans la rédaction.

permettent d'estimer la distance aux deux véhicules voisins, et d'adapter leur vitesse en fonction de ces distances. On suppose que la vitesse de chaque véhicule relaxe vers une vitesse souhaitée commune $V > 0$, avec des termes de correction qui prennent en compte les distances aux véhicules de devant et derrière. Plus précisément, on écrit

$$\ddot{x}_i = \frac{1}{\tau}(V - \dot{x}_i) - \beta\varphi(x_{i+1} - x_i) + \beta\varphi(x_i - x_{i-1}),$$

où $\tau > 0$ et β sont donnés, et $\varphi = -\psi'$, où ψ est une fonction de $]0, +\infty[$ dans \mathbb{R} . On suppose que ψ est C^2 , strictement convexe, et que $\psi(u)$ tend vers $+\infty$ quand u tend vers 0^+ .

On notera U l'ouvert de $[0, L]^N$ périodisé qui contient les $x = (x_1, \dots, x_N)$ tels que

$$x_2 - x_1 > 0, \dots, x_N - x_{N-1} > 0, x_1 - x_N > 0,$$

avec prise en compte de la périodicité comme indiqué précédemment.

1) Écrire le modèle sous la forme d'un système d'ordre 1 en temps sur les variables

$$(x, v) = (x_1, \dots, x_N, v_1, \dots, v_N) \in U \times \mathbb{R}^N, \quad \text{avec } v_i = \dot{x}_i.$$

2) On se donne une collection de positions initiales $x^0 = (x_1^0, x_N^0) \in U$, et un vecteur de vitesses initiales $v^0 = (v_1^0, \dots, v_N^0)$ quelconques.

a) Montrer que le système écrit à la question précédente admet une solution maximale unique définie sur $[0, T^*]$ à valeurs dans U , avec $T^* \in]0, +\infty]$.

b) Montrer que la solution maximale est en fait globale (c'est à dire que $T^* = +\infty$).

3) Écrire le système d'ordre 1 vérifié par les variables de distance et leurs dérivées, i.e.

$$u_i = x_{i+1} - x_i, \quad w_i = \dot{u}_i.$$

4) a) Montrer que le système en (u, w) écrit précédemment admet un unique point d'équilibre défini par

$$(u^e, w^e) = (L/N, \dots, L/N, 0, \dots, 0).$$

b) A quelle situation pour le problème en (x, v) ce point d'équilibre correspond-il ?

5) Étude de stabilité du point d'équilibre.

a) Écrire le gradient de la fonction $F(u, w)$ qui définit le système différentiel, au point d'équilibre (u^e, w^e) . On pourra noter $A = (a_{ij})$ la matrice du Laplacien discret avec conditions périodiques, i.e. telle que $a_{ii} = 2$, $a_{i,i+1} = a_{i,i-1} = -1$ (avec conditions de périodicité), les autres coefficients étant nuls. On pourra noter $\alpha = -\varphi'(L/N) = \psi''(L/N) > 0$ pour alléger les notations.

b) Montrer que la matrice A définie ci-dessus est diagonalisable, de valeurs propres $4 \sin(k\pi/N)^2$, pour $k = 0, 1, \dots, N-1$.

c) Montrer que les valeurs propres du gradient de F sont de partie réelle négative.

d) Sous quelle condition reliant τ , β , et α aura-t-on des oscillations amorties ?

6) (*) Commenter de façon critique et libre ce modèle (réalisme, pertinence du dispositif choisi pour contrôler des véhicules sans conducteurs, ...), et proposer des extensions ou améliorations.

20.13 Transport partiel

On considère X et Y deux ensembles finis, de cardinaux respectifs N et M , deux mesures discrètes $\mu = (\mu_i)$ et $\nu = (\nu_j)$ sur X et Y , respectivement, et une famille de coûts $(c_{ij}) \in \mathbb{R}^{NM}$ donnée. On

supposera toutes les masses élémentaires μ_i et ν_j strictement positives. On suppose que la masse totale de ν est *supérieure ou égale* à celle de μ , qui vaut 1 :

$$|\mu| = \sum_{i=1}^N \mu_i = 1 \leq |\nu| = \sum_{j=1}^M \nu_j. \quad (20.11)$$

On s'intéresse aux plans γ qui transportent μ vers une mesure portée par Y dominée par ν . L'ensemble des plans de transports admissibles est donc

$$\Pi = \left\{ \gamma = (\gamma_{ij}) \in \mathbb{R}_+^{NM}, \sum_j \gamma_{ij} = \mu_i \quad \forall i, \sum_i \gamma_{ij} \leq \nu_j \quad \forall j \right\}. \quad (20.12)$$

On s'intéresse à la minimisation sur Π du coût associé à un plan γ , défini classiquement par

$$C(\gamma) = \sum_{i,j} c_{ij} \gamma_{ij}.$$

Pour tout $\gamma \in \Pi$, on notera $\gamma_\sharp \mu$ la mesure image, définie sur Y par

$$(\gamma_\sharp \mu)_j = \sum_i \gamma_{ij}.$$

- 1) On suppose dans cette première question que ν_j est égal à une constante $\beta > 0$ pour tout j .
 - a) Que devient le problème de minimisation décrit ci-dessus lorsque β tend vers $+\infty$?
 - b) Pour toute valeur de β admissible, on note C_β le coût optimal correspondant. Que peut-on dire de la manière dont C_β dépend de β ?
- 2) On se replace dans le cas général. Montrer que, sous l'hypothèse (20.11), Π n'est pas vide, et qu'il existe $\gamma \in \Pi$ qui minimise C sur Π .
- 3) Montrer que γ minimise C sur Π si et seulement s'il existe $p = (p_i) \in \mathbb{R}^N$ et $q = (q_j) \in \mathbb{R}^M$ tels que $p_i + q_j \leq c_{ij}$ pour tous i, j , avec égalité sur le support de γ , et $q_j = 0$ pour tous les j correspondant à des points non saturés, i.e. tels que $(\gamma_\sharp \mu)_j < \nu_j$.
- 5) On considère la situation où $\nu_j = 1$ pour tout j (en supposant que la masse portée par Y est supérieure à la masse portée par X). Pour tout $\varepsilon > 0$, on considère la fonctionnelle

$$C^\varepsilon : \gamma \in \mathbb{R}_+^{N \times M} \mapsto \sum c_{ij} \gamma_{ij} + \varepsilon \sum_{j=1}^M \eta_j^{1/\varepsilon}.$$
 où l'on a noté pour alléger l'écriture η_j la masse transportée en j , i.e. $\eta_j = (\gamma_\sharp \mu)_j$ (il s'agit donc bien d'une fonction de γ).
 - a) Montrer que C^ε admet un minimiseur γ^ε sur Π_μ , ensemble des plans de transports qui admettent μ comme première marginale ($\sum_j \gamma_{ij} = \mu_i$ pour tout i), sans aucune contrainte sur la mesure image.
 - b) Montrer que, quand ε tend vers 0, γ^ε converge (à sous-suite extraite près) vers une solution du problème initial, c'est à dire un minimiseur de C sur Π (défini par l'équation (20.12)).
- 6) On considère la situation $Y = \mathbb{Z}^2$, et $X = \{x\} = \{(0,0)\} \subset Y$. On place une mesure $\mu > 0$ en x , on considère que ν est la mesure qui donne un poids 1 à chaque point de $Y = \mathbb{Z}^2$, et l'on suppose que les coûts c_j (on omet l'indice i puisqu'il n'y a qu'un point de départ) correspondent à la distance euclidienne ($c_j = |y_j - 0| = |y_j|$, où y_j décrit Y). Expliquer pourquoi le fait que ν soit de masse totale infinie ne pose pas de problème, et décrire aussi précisément que possible la ou les solutions du problème de transport partiel décrit précédemment lorsque μ croît (de 0^+ à $+\infty$).

7) Donner un ou des exemples de situations de la vie “réelle” qui rentrent dans le cadre du problème théorique considéré précédemment. (Sans que ce soit obligatoire, on pourra renverser le problème en considérant que des agents de X associent à chaque $y \in Y$ une utilité, et que l’on cherche à maximiser l’utilité globale.) On s’attachera alors à donner une interprétation des potentiels de Kantorovich q_j (multiplicateurs de Lagrange associés à la contrainte sur la mesure d’arrivée).

20.14 Transport sous contraintes

On considère deux ensembles X et Y de même cardinal $N \in \mathbb{N}$, et l’on note μ (resp. ν) la mesure sur X (resp. Y) qui attribue à chaque point une masse 1 (la masse totale de chaque mesure est donc N). On s’intéresse au problème de transport optimal entre μ et ν sous la contrainte que seuls certains déplacements $i \mapsto j$ sont autorisés. On note comme dans le cours Π l’ensemble des plans sans contraintes :

$$\Pi = \left\{ \gamma_{ij} \in \mathbb{R}_+^{N \times N}, \sum_j \gamma_{ij} = \mu_i \quad \forall i, \sum_i \gamma_{ij} = \nu_j \quad \forall j \right\}$$

On note $I \subset \{(i, j), 1 \leq i \leq N, 1 \leq j \leq N\}$ l’ensemble des couples pour lesquels le transport est autorisé, et l’on définit l’ensemble des plans de transport admissibles associé

$$\Pi_I = \{\gamma \in \Pi, (i, j) \notin I \implies \gamma_{ij} = 0\}.$$

On notera T la matrice carrée associée à I , i.e. définie par $T_{ij} = 1$ si $(i, j) \in I$, $T_{ij} = 0$ sinon.

- 1) a) Quel est le plus petit nombre k qui assure : $[card(I) \geq k \implies \Pi_I \neq \emptyset]$?
- b) Une condition nécessaire pour que Π_I soit non vide est de façon évidente que chaque ligne et chaque colonne de T ait au moins un élément non nul. Montrer que cette condition est suffisante pour $N \leq 2$, mais ne l’est plus dès que $N \geq 3$.
- c) (*) Montrer que Π_I est non vide si et seulement si on peut obtenir à partir de T une matrice de permutation en remplaçant des 1 par des 0.
- 2) On se donne une famille de coûts sur l’ensemble des chemins permis $(c_{ij})_{(i,j) \in I}$, et l’on considère le problème consistant à minimiser

$$C(\gamma) = \sum_{(i,j) \in I} c_{ij} \gamma_{ij} \quad \text{sur } \Pi_I.$$

- a) Montrer que ce problème de minimisation (on parlera de *transport constraint*) admet une solution dès que Π_I est non vide. On notera \bar{C}_I la valeur minimale correspondante. On considérera que $\bar{C}_I = +\infty$ si $\Pi_I = \emptyset$.

On s’intéresse à une version pénalisée de ce problème : on introduit $\varepsilon > 0$, et l’on définit les coûts

$$c_{ij}^\varepsilon = \begin{cases} c_{ij} & \text{si } (i, j) \in I \\ 1/\varepsilon & \text{sinon.} \end{cases}$$

On s’intéresse maintenant au problème de minimisation du coût associé C^ε sur l’ensemble Π des transports admissibles Π *sans la contrainte d’appartenir à Π_I* .

- b) Montrer que ce nouveau problème admet au moins une solution γ^ε .
- c) Dans le cas où Π_I est non vide, montrer que le minimum réalisé $C^\varepsilon(u^\varepsilon)$ converge vers le minimum \bar{C}_I du problème contraint, et que la suite des plans γ^ε converge au sens suivant : l’ensemble des valeurs d’adhérence de la suite γ^ε est inclus dans l’ensemble des solutions du problème de transport contraint.
- d) Montrer que, si Π_I est vide, alors $C^\varepsilon(u^\varepsilon)$ tend vers $+\infty$ quand ε tend vers 0.

20.15 Décomposition polaire discrète

Contexte : dans un espace de Hilbert, si l'on se donne un point x_0 et vecteur v , on peut s'éloigner de x_0 vers l'infini à la vitesse constante $|v|$, en considérant $x_t = x_0 + tv$. C'est moins évident si l'on considère plusieurs points vus comme support d'une mesure atomique, dans l'espace de Wasserstein. On se propose de montrer ici que cet "éloignement de soi-même vers l'infini à vitesse constante" est néanmoins possible dans l'espace de Wasserstein, de façon assez générale. Cette démarche est liée à une version discrète du Théorème de Brenier, qui fait l'objet des dernières questions du problème.

On considère μ et ν les deux mesures de probabilité uniformes associées à des ensembles finis $X = \{x_i\}$ et $Y = \{y_j\}$ dans \mathbb{R}^d , de même cardinal N :

$$\mu = \frac{1}{N} \sum_{i=1}^N \delta_{x_i}, \quad \nu = \frac{1}{N} \sum_{j=1}^N \delta_{y_j}$$

On se place dans le cas du coût quadratique $c_{ij} = |y_j - x_i|^2$, et l'on note γ un minimiseur du coût de transport de μ vers ν :

$$\gamma \in \arg \min_{\Pi(\mu, \nu)} C(\gamma), \quad C(\gamma) = \sum \gamma_{ij} c_{ij}, \quad \Pi(\mu, \nu) = \left\{ \gamma \in \mathbb{R}_+^{N \times N}, \quad \sum_j \gamma_{ij} = \mu_i, \quad \sum_i \gamma_{ij} = \nu_j \right\}$$

On introduit Γ , l'application multivaluée associée au support de γ , qui à x_i fait correspondre l'ensemble des y_j tels que $\gamma_{ij} > 0$ (Γ est une application de X dans l'ensemble des parties de Y).

1) Montrer que Γ est monotone, c'est à dire que

$$(y_{j'} - y_j) \cdot (x_{i'} - x_i) \geq 0 \quad \forall x_i, x_{i'} \in X, y_j \in \Gamma(x_i), y_{j'} \in \Gamma(x_{i'}).$$

Indication : on pourra considérer, en justifiant soigneusement la démarche, des variations de γ du type $\gamma^\varepsilon = \gamma - \varepsilon \delta_{ij} + \varepsilon \delta_{ij'} - \varepsilon \delta_{i'j'} + \varepsilon \delta_{i'j}$, où $\delta_{ij} \in \mathbb{R}^{N \times N}$ correspond au plan qui envoie une masse unitaire de x_i vers y_j , et $\varepsilon > 0$.

2) Montrer que Γ est cycliquement monotone, c'est à dire que

$$\sum_{k=1}^p x_{i_k} \cdot (y_{j_k} - y_{j_{k-1}}) \geq 0,$$

pour tous $(i_1, j_1), \dots, (i_p, j_p)$, avec $y_{j_k} \in \Gamma(x_{i_k})$ pour $k = 1, \dots, p$, avec la convention $j_0 = j_p$.

3) Montrer la réciproque de la question précédente, c'est à dire que si un plan $\gamma \in \Pi(\mu, \nu)$ est tel que son support Γ est cycliquement monotone, alors il est optimal.

(On pourra commencer par étudier le cas où γ est associé à une bijection de X vers Y .)

4) Les notions de monotonie et cyclique monotonie pour une application multivaluée de X vers Y ont été définies dans les questions précédentes. Montrer que la cyclique monotonie implique la monotonie, mais qu'il existe des applications monotones qui ne sont pas cycliquement monotones (on pourra proposer un exemple pour $N = 4$, avec $X = Y$ l'ensemble des 4 sommets d'un carré).

On note μ_0 la mesure de probabilité uniforme sur X (notée précédemment μ), et l'on considère une collection de N vecteurs $v_1, \dots, v_N \in \mathbb{R}^d$. Pour toute permutation $\varphi \in S_N$, tout $t \geq 0$, on note μ_t^φ la mesure obtenu par transport de μ_0 selon le champ de déplacement $(tv_{\varphi(i)})$:

$$\mu_t^\varphi = \frac{1}{N} \sum_{i=1}^N \delta_{x_i + tv_{\varphi(i)}}.$$

5) Montrer que, si les v_i sont non tous nuls, la quantité $W_2(\mu_0, \mu_t^\varphi)$ (distance de Wasserstein entre μ_0 et μ_t^φ , définie comme la racine carrée du coût minimal) tend vers $+\infty$ avec t , quel que soit $\varphi \in S_N$.

Donner un exemple de situation pour laquelle, selon le choix de φ , cette quantité est strictement croissante sur $[0, +\infty]$, ou alors décroissante sur certaines plages de temps.
(*On pourra considérer par exemple $X = \{-1, 1\} \in \mathbb{R}$, $v_1 = -1$, $v_2 = 1$.*)

6) Montrer qu'il existe une permutation $\varphi \in S_N$ telle que

$$W_2(\mu_0, \mu_t^\varphi) = t \frac{1}{\sqrt{N}} |v|_2 \quad \forall t \geq 0.$$

(*On pourra s'intéresser au problème de transport optimal de μ_0 vers ν , mesure uniforme sur l'ensemble des v_i , identifiés à des points de \mathbb{R}^d .*)

7) Montrer que, pour le φ trouvé à la question précédente, l'évolution est expansive, au sens où le mouvement des points défini par $t \mapsto x_i + tv_{\varphi(i)}$, pour t croissant, augmente toutes les distances entre les points. Montrer (par un contre-exemple) que cette propriété d'expansion ne suffit pas à assurer la propriété de croissance linéaire de la distance $W_2(\mu_0, \mu_t^\varphi)$.

Soit Ψ une application de \mathbb{R}^d dans $\mathbb{R} \cup \{+\infty\}$, convexe⁸. On la suppose *propre*, c'est à dire non identiquement égale à $+\infty$. On définit son sous-différentiel en un point x comme l'ensemble

$$\partial\Psi(x) = \{h \in \mathbb{R}^d, \Psi(x) + h \cdot (x' - x) \leq \Psi(x') \quad \forall x' \in \mathbb{R}^d\}.$$

Le sous-différentiel est donc défini comme une application multivaluée, i.e. une application qui à chaque x associe une partie de \mathbb{R}^d , éventuellement vide ou réduite à un singleton.

8) Montrer que l'application multivaluée $x_i \longmapsto \partial\Psi(x_i)$ définie sur X est cycliquement monotone (au sens défini dans la question 1) .

9) Réciproquement, on considère Γ une application multivaluée de \mathbb{R}^d dans l'ensemble des parties de \mathbb{R}^d , cycliquement monotone. On souhaite montrer que Γ est incluse dans le sous-différentiel d'une fonction convexe propre, au sens où il existe Ψ non identiquement égale à $+\infty$, telle que, pour tout x , tout $y \in \Gamma(x)$

$$y \in \partial\Psi(x).$$

On considère $x_0 \in \mathbb{R}^d$, $y_0 \in \Gamma(x_0)$, et on définit $\Psi : \mathbb{R}^d \rightarrow \mathbb{R}$ par

$$\Psi(x) = \sup \{(x - x_n) \cdot y_n + (x_n - x_{n-1}) \cdot y_{n-1} + \dots + (x_1 - x_0) \cdot y_0\},$$

où le sup est pris sur l'ensemble des familles de n couples $(x_1, y_1), \dots, (x_n, y_n)$, avec $y_k \in \Gamma(x_k)$, et n un entier naturel non nul.

a) Montrer que Ψ est convexe, que $\Psi(x_0) = 0$, et donc que Ψ est non identiquement égale à $+\infty$.

b) Soit $x \in \mathbb{R}^d$ tel que $\Gamma(x) \neq \emptyset$, et $y \in \Gamma(x)$. Pour tout $x' \in \mathbb{R}^d$, pour toute famille de couples $(x_1, y_1), \dots, (x_n, y_n)$, montrer que

$$(x - x_n) \cdot y_n + (x_n - x_{n-1}) \cdot y_{n-1} + \dots + (x_1 - x_0) \cdot y_0 \leq \Psi(x') - (x' - x) \cdot y,$$

et en déduire que $y \in \partial\Psi(x)$.

10) Démontrer la version discrète du théorème de Brenier : on considère $X \subset \mathbb{R}^d$ de cardinal N , et une collection de vecteurs v_1, \dots, v_N associés aux points de X . Il existe une permutation φ et une application convexe propre Ψ telle que $v_{\varphi(i)}$ soit dans $\partial\Psi(x_i)$ pour tout $i = 1, \dots, N$.

20.16 Entropie relative

On note

$$\mathcal{P}_N = \left\{ \mu = (\mu_1, \dots, \mu_N) \in \mathbb{R}_+^N, \sum_i \mu_i = 1 \right\}$$

⁸. Pour une fonction à valeurs dans $\mathbb{R} \cup \{+\infty\}$, la notion de convexité se définit selon l'inégalité usuelle, avec la convention naturelle $a \leq +\infty$ pour tout $a \in \mathbb{R}$, et $+\infty \leq +\infty$.

l'espace des lois de probabilités sur l'ensemble à N éléments, et $\mathring{\mathcal{P}}_N$ le sous-ensemble des μ qui chargent tous les points, i.e. tels que $\mu_i > 0$ pour tout i . On définit l'entropie d'une loi par

$$S(\mu) = \sum \mu_i \log \mu_i,$$

et, pour $\eta \in \mathring{\mathcal{P}}$, l'entropie relative de μ par rapport à η par

$$\mu \in \mathcal{P}_N \longmapsto S_\eta(\mu) = \sum_i \mu_i \log \frac{\mu_i}{\eta_i} = \sum_i \frac{\mu_i}{\eta_i} \log \left(\frac{\mu_i}{\eta_i} \right) \eta_i.$$

1) Montrer que l'entropie relative de $\mu \in \mathcal{P}$ relativement à la mesure uniforme ρ est égale l'entropie de μ , à une constante additive près.

2) Soit $\eta \in \mathring{\mathcal{P}}_N$ donné. Montrer que S_η admet un minimiseur unique sur \mathcal{P}_N , qui est η .

On se donne $\eta \in \mathring{\mathcal{P}}$, un vecteur $E = (E_i)$ de \mathbb{R}^N , $\bar{E} \in \mathbb{R}$, et l'on s'intéresse au problème de minimisation de $S_\eta(\mu)$ sur l'ensemble

$$K = \left\{ \mu \in \mathcal{P}_N, \sum_i \mu_i E_i = \bar{E} \right\}.$$

3) Que peut on dire si $\bar{E} \notin [\min(E_i), \max(E_i)]$?

4) Faire l'analyse du problème dans le cas où $\bar{E} = \min(E_i)$ ou $\bar{E} = \max(E_i)$.

On fait maintenant l'hypothèse que $\bar{E} \in [\min(E_i), \max(E_i)]$.

5) Montrer que S_η admet un minimiseur unique $\bar{\mu}$ sur K .

6) a) Montrer que $K \cap \mathring{\mathcal{P}}_N \neq \emptyset$. (*Suggestion : on pourra supposer que max et min sont atteints en 1 et N, écrire \bar{E} comme combinaison barycentrique de E_1 et E_N , et montrer que l'on peut rajouter une masse ε sur les points entre 2 et $N - 1$ en corrigeant μ_1 et μ_N de façon à préserver les deux contraintes.*)

b) Montrer que $\bar{\mu} \in \mathring{\mathcal{P}}_N$. (*Indication : on pourra montrer que, si ça n'est pas le cas, alors on peut en perturbant $\bar{\mu}$ obtenir une valeur strictement inférieure de la fonctionnelle.*)

7) Justifier soigneusement que l'on peut utiliser les conditions nécessaires d'optimalité sous contrainte, et montrer qu'il existe $\beta \in \mathbb{R}$ tel que $\bar{\mu}$ s'écrit sous la forme (distribution de Gibbs)

$$\bar{\mu}_i = \frac{1}{Z} \eta_i e^{-\beta E_i} \text{ pour } i = 1, \dots, N, \text{ avec } Z = \sum_i \eta_i e^{-\beta E_i}.$$

On considère maintenant 2 ensembles finis X et Y , on se donne deux mesures de probabilité $\mu \in \mathcal{P}(X)$ et $\nu \in \mathcal{P}(Y)$, et l'on note Λ_μ^ν l'espace des mesures de probabilité sur le produit, qui admettent μ et ν comme marginales, c'est à dire

$$\Pi_\mu^\nu = \left\{ \gamma \in \mathcal{P}(X \times Y), \sum_x \gamma_{xy} = \nu_y, \sum_y \gamma_{xy} = \mu_x \right\}.$$

On se donne une collection $c = (c_{xy})$ de coûts⁹. On s'intéresse au problème de minimisation sur Π_μ^ν de

$$C(\gamma) = \sum_{xy} \gamma_{xy} c_{xy}.$$

9. Le terme vient du fait que l'on peut voir $\gamma \in \Lambda_\mu^\nu$ comme un *plan de transport* entre μ et ν , μ correspondant à une masse unitaire répartie sur X , et ν une collection de capacités d'accueil dont la somme est elle-même unitaire, γ_{xy} est la masse transportée de x vers y , et c_{xy} est le coût de transport d'une quantité unitaire de masse de x vers y .

8) Montrer que $C(\cdot)$ admet un minimiseur sur Π_μ^ν .

9) a) Montrer sur un exemple que ce minimiseur peut ne pas être unique, et sur un autre exemple qu'il peut l'être.

(*Suggestion : on pourra considérer que X et Y sont des ensembles de points du plan, que $c_{xy} = |y - x|$, dans le cas où X et Y sont par exemple tous deux de cardinal 2.*)

b) Montrer que l'ensemble Λ des minimiseurs de $C(\cdot)$ sur Π_μ^ν est un convexe fermé.

On s'intéresse dans la suite à une version dite *entropique* du problème. On se donne un paramètre $\varepsilon > 0$ (destiné à tendre vers 0 comme son nom le suggère), et l'on définit

$$C_\varepsilon(\gamma) = C(\gamma) + \varepsilon S(\gamma) = \sum_{(x,y) \in X \times Y} \gamma_{xy} c_{xy} + \varepsilon \sum_{(x,y) \in X \times Y} \gamma_{xy} \log \gamma_{xy}.$$

10) Montrer que C_ε admet un minimiseur unique γ^ε sur Π_μ^ν , et que $\gamma^\varepsilon \in \mathring{\mathcal{P}}(X, Y)$ (i.e. $\gamma_{xy} > 0$ pour tous x, y).

11) Montrer que (γ^ε) est bornée, et que toute valeur d'adhérence de (γ^ε) minimise C sur Π_μ^ν .

12) a) Montrer que $S(\cdot)$ admet un minimiseur γ^{opt} unique sur Λ (ensemble des minimiseurs de C sur Π_μ^ν).

b) Montrer que minimiser C_ε revient à minimiser (on note $C^{opt} = C(\gamma^{opt})$)

$$\gamma \mapsto S_\varepsilon(\gamma) = \frac{1}{\varepsilon} (C(\gamma) - C^{opt}) + S(\gamma).$$

c) Montrer que l'on a, pour tout $\varepsilon > 0$, $S(\gamma^\varepsilon) \leq S_\varepsilon(\gamma^\varepsilon) \leq S_\varepsilon(\gamma^{opt}) = S(\gamma^{opt})$.

d) Montrer que γ^ε converge vers γ^{opt} .

13) a) Montrer que $C_\varepsilon(\gamma)$ est, à une constante additive près et à un facteur multiplicatif strictement positif près, l'entropie relative de γ relativement à $\eta^\varepsilon \in \mathcal{P}(X \times Y)$ définie par

$$\eta_{xy}^\varepsilon = \frac{1}{Z} e^{-c_{xy}/\varepsilon}$$

où $Z > 0$ est une constante de normalisation (qui dépend bien sûr de ε).

b) En déduire que η^ε minimise C_ε sur $\mathcal{P}(X \times Y)$. Pourquoi η^ε n'est-il pas égal à γ^ε en général ?

c) Préciser la limite de la mesure η^ε quand ε tend vers 0.

14) a) Montrer que γ^ε s'écrit

$$\gamma_{xy}^\varepsilon = a_x b_y \eta_{xy},$$

où $a = (a_x) \in]0, +\infty[^X$, et $b = (b_x) \in]0, +\infty[^Y$.

b) Réciproquement, montrer que si $\gamma \in \Pi_\mu^\nu$ s'écrit comme ci-dessus, alors $\gamma = \gamma^\varepsilon$.

15) On se donne $\bar{\gamma} \in \mathring{\mathcal{P}}(X \times Y)$, et l'on considère les ensembles

$$\Pi_\mu = \left\{ \gamma \in \mathcal{P}(X \times Y), \sum_y \gamma_{xy} = \mu_x \right\}, \quad \Pi^\nu = \left\{ \gamma \in \mathcal{P}(X \times Y), \sum_x \gamma_{xy} = \nu_y \right\}$$

Montrer que $S_{\bar{\gamma}}(\cdot)$ admet un unique minimiseur sur Π_μ , qui s'écrit

$$\gamma_{xy} = \bar{\gamma}_{xy} \frac{\mu_x}{\sum_{y'} \bar{\gamma}_{xy'}},$$

et énoncer un résultat analogue portant sur la minimisation de $S_{\bar{\gamma}}(\cdot)$ sur Π^ν .

16) Partant de $\gamma^0 = \eta$, on construit une suite $\gamma^0 = \eta, \gamma^{1/2}, \gamma^1, \dots, \gamma^k, \gamma^{k+1/2}, \gamma^{k+1}, \dots$ selon

$$\gamma^{k+1/2} = \arg \min_{\Pi_\mu} S_{\gamma^k}(\gamma), \quad \gamma^{k+1} = \arg \min_{\Pi_\nu} S_{\gamma^{k+1/2}}(\gamma).$$

Montrer que γ^k peut se mettre sous la forme $\gamma_{xy}^k = a_x^k b_y^k \eta_{xy}$, où les a^k et b^k sont définis par des relations de récurrence que l'on explicitera.

17) Montrer que si (a^k, b^k) converge vers $(a, b) \in]0, +\infty[^{X \times Y}$, alors γ^k converge vers γ^ε , minimiseur de C_ε sur Π_μ^ν .

20.17 Décroissance de l'entropie pour les schémas de différences finies

On considère l'équation de la chaleur sur l'intervalle $]0, 1[$ avec conditions périodiques (0 est identifié à 1) :

$$\partial_t u(x, t) - D \partial_{xx} u(x, t) = 0.$$

On discrétise l'intervalle en espace de façon uniforme avec les points $x_j = j \Delta x$ (avec $\Delta x = 1/J$), pour $j = 0, \dots, J$. Du fait de la condition périodique, on identifie x_0 et x_J . De la même manière, pour $\Delta t > 0$ choisi, on définit les temps discrets $t^0, \dots, t^n = n \Delta t, \dots$. On ne s'attachera pas, dans la rédaction, à détailler les considérations liées aux fait que le domaine est périodique ; par exemple on considérera comme acquis, sans qu'il soit besoin de le préciser, que u_{j+1}^n , pour $J = N$, correspond à u_1^n , etc...

On s'intéresse dans un premier temps au schéma explicite

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} - D \frac{u_{j-1}^n - 2u_j^n + u_{j+1}^n}{(\Delta x)^2} = 0 \quad \forall j = 1, \dots, J,$$

avec la condition initiale (u_j^0) donnée.

1) On suppose que $(u_j^0)_j$ est la loi d'une variable aléatoire à valeurs dans $1, \dots, J$ (on rappelle que 0 est identifié à J), c'est à dire que les u_j^0 sont positifs, et leur somme vaut 1. Montrer que les vecteurs $u^n = (u_j^n)_j$ obtenus par application du schéma sont aussi des lois de probabilité sur $\{1, \dots, J\}$, sous réserve qu'une certaine condition (que l'on précisera) soit vérifiée par Δx et Δt .

2) On définit l'entropie d'une loi de probabilité donnée sous la forme d'un vecteur $u = (u_j)_{1 \leq j \leq J}$ par

$$S(u) = \sum_j \varphi(u_j),$$

où φ est une fonction définie sur $[0, 1]$, régulière sur $]0, 1[$, strictement convexe (on pourra penser par exemple à $\varphi(a) = a \log a$). Montrer que, sous la condition obtenue à la question 1, l'entropie des u^n décroît.

A quelle condition sur u^n a-t-on $S(u^{n+1}) < S(u^n)$?

On s'intéresse maintenant au schéma implicite

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} - D \frac{u_{j-1}^{n+1} - 2u_j^{n+1} + u_{j+1}^{n+1}}{(\Delta x)^2} = 0 \quad \forall j = 1, \dots, J.$$

3) Montrer qu'il s'agit bien d'un "algorithme", au sens où le schéma permet de calculer sans ambiguïté u^{n+1} en fonction de u^n .

4) Montrer que, quels que soient $\Delta t > 0$ et $\Delta x > 0$, partant d'une loi de probabilité u^0 , les vecteurs u^n obtenus par application de ce schéma sont aussi des lois de probabilité.

5) Montrer, comme dans le cas explicite, que l'entropie $S(u^n)$ décroît (cette fois-ci sans condition sur les pas d'espace et de temps).

A quelle condition sur u^n a-t-on $S(u^{n+1}) < S(u^n)$?

6) Que peut on déduire des propriétés de décroissance de l'entropie (démontrée ci-dessus) sur le comportement des vecteurs solutions $u^n = (u_j^n)_j$, quand n tend vers $+\infty$?

20.18 Flots de gradients discrets dans l'espace de Wasserstein, équation de la chaleur comme flot de gradient de l'entropie

0) Questions préliminaires : le cas euclidien

On note $H = \mathbb{R}^d$, $d \geq 1$, et on considère Ψ une fonctionnelle continûment différentiable de H dans \mathbb{R} . Pour toute donnée initiale $u^0 \in H$, pour tout $\tau > 0$ on définit l'algorithme suivant (le fait qu'il permette effectivement de définir une suite d'itérés, de façon unique ou pas, fera l'objet des questions qui suivent)

$$u^{k+1} \in \arg \min \left(\Psi(u) + \frac{1}{2\tau} |u - u^k|^2 \right) \quad (JKOh)$$

a) Montrer par un exemple que $(JKOh)$ ne permet pas toujours de définir une suite d'itérés. Donner des conditions suffisantes sur Ψ pour que la démarche assure l'existence d'itérés successifs, et des conditions suffisantes sur Ψ qui assurent l'existence *et* l'unicité de ces itérés, éventuellement conditionnées à une condition sur τ . La convergence de u^k vers u implique-t-elle que u soit un minimum local de Ψ ?

b) Dans le cas où l'on a existence et unicité des itérés, écrire les conditions nécessaires d'optimalité, et expliquer pourquoi ce procédé peut être vu comme une discrétisation en temps du flot de gradient

$$\frac{du}{dt} = -\nabla \Psi(u).$$

c) Montrer que, pour toute solution régulière $t \mapsto u(t)$ de l'équation différentielle ci-dessus, $\Psi(u(t))$ est décroissant. Montrer que, pour toute suite (u^k) obtenue suivant le procédé $(JKOh)$, la suite $\Psi(u^k)$ est décroissante, et même strictement décroissante tant qu'elle n'est pas stationnaire.

Soit $N \geq 1$. On se place maintenant sur $\mathcal{A}_N = \mathcal{A}_N(\mathbb{R}^d)$ l'ensemble des mesures de probabilités atomiques sur \mathbb{R}^d supportées de façon uniforme sur N points (non nécessairement distincts deux à deux)

$$\mathcal{A}_N = \left\{ \mu = h_N \sum_{i=1}^N \delta_{x_i}, \quad x_1, \dots, x_N \in \mathbb{R}^d \right\},$$

où $h_N = 1/N$ est la masse portée par chaque atome. Pour μ^0 et μ^1 dans \mathcal{A}_N , on définit

$$W_2(\mu^0, \mu^1) = \left(h_N \min_{g \in S_N} \sum_{i=1}^N |x_{g(i)}^1 - x_i^0|^2 \right)^{1/2},$$

où S_N est le groupe symétrique des permutations sur l'ensemble à N éléments.

1) Rappeler pourquoi W_2 est une distance sur \mathcal{A}_N . Préciser le lien entre l'espace métrique \mathcal{A}_N muni de cette distance et l'espace \mathbb{R}^N muni de la distance euclidienne canonique.

Soit Ψ une fonctionnelle continue de \mathcal{A}_N dans \mathbb{R} . Pour toute donnée initiale $\mu^0 \in \mathcal{A}_N$, on définit l'algorithme suivant, que l'on appellera algorithme JKO,

$$\mu^{k+1} \in \arg \min_{\mu \in \mathcal{A}_N} \left(\Psi(\mu) + \frac{1}{2\tau} W_2(\mu^k, \mu)^2 \right) \quad (JKO)$$

2) Montrer que

$$\inf_{\mu \in \mathcal{A}_N} \left(\Psi(\mu) + \frac{1}{2\tau} W_2(\mu^k, \mu)^2 \right) = \inf_{x \in \mathbb{R}^{dN}} \left(\Psi \left(h_N \sum_{j=1}^N \delta_{x_j} \right) + \frac{h_N}{2\tau} \sum_{j=1}^N |x_j - x_j^k|^2 \right).$$

On s'intéresse ici au cas particulier d'une fonctionnelle du type

$$\mu \in \mathcal{A} \longmapsto \Psi(\mu) = \langle \mu, D \rangle = h_N \sum_{i=1}^N D(x_i),$$

où D est une fonctionnelle C^1 de \mathbb{R}^d dans \mathbb{R} .

3) Montrer que Ψ est bien continue sur \mathcal{A}_N , et faire l'analyse la plus complète possible du problème

$$\min_{\mu \in \mathcal{A}_N} \Psi(\mu)$$

(existence d'un minimum, d'un minimiseur, unicité éventuelle du minimiseur, nombre de minimiseurs en cas de non unicité, etc ... en fonction des hypothèses que l'on fait sur $D(\cdot)$).

4) On suppose ici D continûment différentiable et α -convexe, sans condition de signe sur α . Montrer que, pour τ assez petit, (JKO) permet bien de construire de façon unique une suite d'itérés (μ^k) . Préciser le lien entre la suite de mesures obtenues et le flot de gradient sur l'espace euclidien \mathbb{R}^d associé à la fonction D , et décrire les différents comportements asymptotiques possibles de la suite des itérés, selon le choix de D .

Équation de la chaleur "lagrangienne" et flot de gradient de l'entropie

On s'intéresse maintenant au cas de fonctionnelles Ψ qui font "interagir" les atomes entre eux. On considère le cas $d = 1$: \mathcal{A}_N représente ici l'ensemble des mesures atomiques qui sont uniformément réparties sur N atomes de l'intervalle $]0, 1]$, *en supposant que le dernier atome est fixé à l'extrémité droite*. On note ainsi

$$U = \{x = (x_1, \dots, x_{N-1}) \in]0, 1[^{N-1}, x_0 = 0 < x_1 < x_2 < \dots < x_{N-1} < x_N = 1\}$$

l'ensemble des degrés de liberté. On associe à la mesure $\mu = h_N \sum \delta_{x_j} \in \mathcal{A}_N$ la fonction en escalier ρ obtenue en étalant la masse de chaque atome entre 1 et N sur l'intervalle à sa gauche :

$$\rho(x) = \frac{h_N}{x_j - x_{j-1}} \quad \forall x \in]x_{j-1}, x_j[,$$

(on pourra noter cette valeur $\rho_{j-1/2}$), et l'on définit $\Psi(\mu)$ comme l'entropie (mathématique) de ρ , i.e.

$$\Psi(\mu) = \int \rho \log(\rho).$$

5) On considère une mesure $\mu^k \in \mathcal{A}_N$, caractérisée par les positions de ses atomes $x^k = (x_1^k, \dots, x_N^k) \in U$. Montrer qu'une étape de (JKO) peut s'écrire

$$\mu^{k+1} \in \arg \min_{\mu \in \mathcal{A}_N} \left(\Psi(\mu) + \frac{h_N}{2\tau} \sum_{j=1}^{N-1} |x_j - x_j^k|^2 \right)$$

(sans qu'il soit besoin de considérer toutes les permutations pour le second terme).

6) Écrire précisément la fonctionnelle Ψ en fonction des x_i qui constituent le support de μ , et montrer que l'algorithme permet de construire une suite d'itérés dans \mathcal{A}_N définis de façon unique.

7) Écrire les conditions d'optimalité vérifiées par les points du support de μ^{k+1} , et montrer que le schéma peut s'interpréter comme une discrétisation en temps d'un système d'équations du type

$$\dot{x}_j = F(x_{j-1}, x_j, x_{j+1}), \quad j = 1, \dots, N-1, \quad \text{avec } x_0 = 0, x_N = 1.$$

pour une certaine fonction $F(\cdot, \cdot, \cdot)$ de \mathbb{R}^3 dans \mathbb{R} . Montrer que ce système est un flot de gradient au sens hilbertien pour une certaine fonctionnelle Φ définie sur U , que l'on explicitera.

8) Montrer que, pour toute condition initiale μ^0 supporté par les points $0 < x_1^0 < \dots < x_N^0 = 1$, le système différentiel obtenu dans la question précédente admet une solution maximale unique, qui est globale.

9) Montrer que l'approche peut être interprétée comme la version discrète en espace et en temps d'un transport conservatif de densité à la vitesse $-\partial_x \rho / \rho$.

10) Expliquer pourquoi ce qui précède permet d'interpréter l'équation de la chaleur comme flot de gradient de la fonctionnelle d'entropie. Quelle est le type de conditions aux limites correspondant à l'équation de la chaleur obtenue formellement de cette manière ?

11)(*) Adapter ce qui précède au cas d'une fonctionnelle générale de type $\Psi(\rho) = \int \varphi(\rho)$, où φ est une fonction de \mathbb{R}_+ dans \mathbb{R} , et montrer que l'équation aux dérivées partielles formellement correspondante est l'équation de transport conservatif à la vitesse $-\partial_x \varphi'(\rho)$.

Flot de gradient avec contrainte

Toujours dans le cas $d = 1$, on revient maintenant au cas d'une fonctionnelle de type $\langle \mu, D \rangle$, où D est une fonction C^1 sur \mathbb{R} . Une mesure de \mathcal{A}_N répartie uniformément sur N atomes est toujours caractérisée par les points de son support x_1, \dots, x_N . Dans l'optique de modéliser une foule dont la densité ne peut dépasser une certaine valeur, on impose aux points du support d'appartenir à l'ensemble

$$K = \{x = (x_1, \dots, x_N), x_j - x_{j-1} \geq 1 \quad \forall j = 2, \dots, N\}.$$

On se permettra de noter K le sous-ensemble des mesures de \mathcal{A}_N dont les points du support vérifient les contraintes d'espacement minimal ci-dessus. On s'intéresse donc maintenant, partant d'une mesure initiale μ^0 dans K , à l'algorithme de JKO sous contrainte (que l'on écrit ici comme formulé sur le vecteur des positions)

$$x^{k+1} \in \arg \min_{\mu \in K} \left(\sum_{j=1}^N D(x_j) + \frac{1}{2\tau} \sum_{j=1}^N |x_j - x_j^k|^2 \right) \quad (\text{JKOc}).$$

On suppose que la fonction $D(\cdot)$ est α -convexe, avec $\alpha \in \mathbb{R}$.

12) Montrer que, pour τ suffisamment petit, l'algorithme permet de définir de façon unique les itérés successifs, et écrire les conditions d'optimalité sous contrainte vérifiées par $x^{k+1} = (x_j^{k+1})$ (on pourra noter $p_{12}, p_{23}, \dots, p_{N-1,N}$ les multiplicateurs de Lagrange associés aux contraintes).

13) Décrire l'évolution de ce système dynamique discret lorsque l'on part d'une mesure initiale dans l'intérieur de K , pour la fonction $x \mapsto D(x) = x^2/2$, ainsi que pour la fonction $D(x) = e^x$.

(On pourra commencer par le cas $N = 1$, puis $N = 2$, avant de décrire la situation générale.)

On s'intéresse maintenant au cas de masses non uniformes. Plus précisément, on considère N atomes (identifiés à des piétons), auxquels on affecte une fois pour toutes les masses $1, \varepsilon, \varepsilon^2, \dots$ etc. L'ensemble K s'écrit maintenant (en tant qu'ensemble de mesures)

$$K = \left\{ \mu = \sum_{j=1}^N m_j \delta_{x_j}, x_j - x_{j-1} \geq 1 \quad \forall j = 2, \dots, N \right\}, \quad m_j = \varepsilon^{j-1},$$

pour $\varepsilon > 0$ fixé. On choisit pour simplifier l'écriture de ne pas normaliser la masse totale à 1, il ne s'agit donc plus de mesures de probabilité.

- 14) On se place dans le cas d'une condition initiale qui sature la contrainte : $x_j - x_{j-1} = 1$ pour $j = 2, \dots, N$, et un champ D strictement croissant (de telle sorte que les atomes ont tendance à se déplacer vers la gauche) et strictement convexe. Pour $\tau > 0$ fixé, l'algorithme (JKOc) permet pour tout $\varepsilon > 0$ de calculer un premier itéré caractérisé par le vecteur des positions des masses, que l'on notera x^ε au lieu de x^1 pour souligner la dépendance en ε . Décrire, en justifiant le plus précisément possible la réponse, le comportement de x^ε lorsque ε tend vers 0.
- 15) De façon plus générale, décrire le comportement du système au fil des itérés, pour une condition initiale dans K , sous l'hypothèse d'une fonction D strictement croissante et strictement convexe. Proposer une interprétation du modèle asymptotique obtenu en faisant tendre ε vers 0, en termes de modélisation de mouvements de piétons.

Index

- Élémentaire (chemin), 226
- Énergie
 - cinétique, 31
 - interne, 31
- Épigraphe, 273
- Équation
 - de continuité, 85
 - de la chaleur, 14
 - de Navier-Stokes, 212
 - de Stokes, 212, 217
- Équilibre
 - de Nash, 282
- Équilibre (point d'), 239
- aaPoumEFL, 423
- aaPoumNonLIn, 420
- Amblygone (triangle), 196, 205
- Amplification (coefficient), 330
- Arbre, 227
- Arbre
 - couvrant, 227
- Asymétrique (graphe), 225
- Automorphisme (de graphe), 225
- Biparti (graphe), 224
- Bouts (espace des), 400
- Cantor (Processus d'extraction diagonale), 380, 400
- Capacité (d'un sous-domaine), 249
- Capacité thermique, 19
- Cauchy-Lipschitz (Théorème), 237
- Cercles de Gerschgorin, 395
- Chaleur (équation), 14
- Champ harmonique, 229
- Charismatique (réseau), 140
- Charisme, 416
- Chemin
 - simple, 226
 - élémentaire, 226
- Cinétique (énergie), 28
- Clivage, 414
- Coefficient
 - de diffusion, 14
- Complet (graphe), 224
- Composante connexe, 227
- Conditions aux limites
 - de Dirichlet, 90
 - de Neumann, 90
- Conditions aux limites
 - de Dirichlet, 213
 - de Neumann, 43, 74
 - de Robin, 408
- Continuité (équation d'e), 85
- Contrainte (tenseur des), 206
- Convection naturelle, 32
- Courant-Fisher (théorème de), 369
- Coût marginal, 28
- Céa (lemme de), 351
- Darcy (Loi de)
 - Isotrope, 215
- Densité
 - de population, 22
 - du bâti, 22
- Description
 - eulérienne, 117
 - lagrangienne, 117
- Diagramme fondamental, 13
- Diesel particle, 102
- Diffusion (coefficient), 14
- Dirichlet (conditions aux limites de), 213
- Dirichlet (conditions aux limites de), 90
- Dispersion, 180
- Distance
 - géodésique, 144
- Divergence discrète, 45
- Divergence de Kullback-Leibler, 144
- Dual (graphe), 225
- Dérivée
 - lagrangienne, 14, 17
 - particulaire, 14, 17
 - totale, 14, 17
- Entropie, 33
- Entropie
 - mathématique, 33
 - physique, 33
- Entropie relative, 98
- Equation
 - Stokes, 219
- Équilibre
 - de Nash, 118
 - de Wardrop, 118
- Espace
 - H^2 , 248

des bouts, 400
 Espérance de vie, 15
 Euler (schéma), 283
 Eulérienne (description), 117
 Extensive (variable), 17, 129

 Faible (solution), 87, 198, 255
 Farkas (lemme), 266
 Fick (loi de), 89
 Finesse, 28
 Finesse (d'un avion), 215
 Fisher (information de), 144
 Fisher (information), 92
 Flot, 17
 Flux (vecteur), 84
 Fonction
 indicatrice, 273
 Fonctions implicites (théorème), 376
 Formulation variationnelle, 43
 Formule
 de Green, 393
 de Green (deuxième), 252
 de Green (discrète), 49
 de Green (première), 252
 Forçage radiatif, 32
 Frustration, 118

 Gerschgorin (cercles de), 395
 Graphe
 asymétrique, 225
 biparti, 224
 complet, 224
 dual, 225
 hiérarchique, 227
 non orienté, 39, 224
 orienté, 223
 pondéré, 224
 simple, 224
 vide, 224, 226
 Green (formule de), 252, 393

 Hahn-Banach (théorème de), 361
 Harmonique (champ), 229
 Harmonique (mesure), 94, 129
 Hiérarchie, 227
 Hyperstatique, 195, 205, 282

 Identité du parallélogramme , 359
 Inégalité
 de Poincaré, 253
 de Poincaré généralisée, 254
 Indicatrice (fonction), 273
 Information de Fisher, 92
 Information mutuelle, 26
 Intensive (variable), 17, 129
 Interpolation (opérateur d'), 352, 354
 Inversion locale (théorème), 379

 Inégalité
 de Cauchy-Schwarz, 358
 de Poincaré, 42
 Irréductible (matrice), 228
 Isomorphisme (de graphe), 225

 Kirchhof's law, 72
 Krein-Milman (théorème), 376
 Kullback-Leibler (divergence), 144

 Lagrange (multiplicateurs de), 281
 Lagrangien, 270, 320
 Lagrangienne (description), 117
 Lagrangienne (dérivée), 14
 Laplacien
 combinatoire, 229
 discret, 40
 pondéré, 229
 Laplacien non symétrique, 230
 Law
 Kirchhof's, 72
 Lax-Milgram (théorème de), 363
 Lemme
 de Aubin-Nitsche, 355
 de Céa (cas non symétrique), 351
 de Céa (cas symétrique), 351
 Lemme de Aubin-Nitsche, 355
 Lemme de Farkas, 266
 Loi
 d'Ohm, 37
 de Fick, 89

 Marginal (coût), 28
 masse ponctuelle, 28
 Matrice
 d'information de Fisher, 144
 irréductible, 228
 à diagonale strictement dominante, 395
 Maximum (principe), 124, 126
 Mesure
 harmonique, 94, 129
 Modèle
 de Verhulst, 244
 Multiplicateurs de Lagrange, 281

 Nash (équilibre), 118, 282
 Navier-Stokes (équations), 212
 Neumann (conditions aux limites), 43, 74, 90
 Nombre
 de Reynolds, 212
 de Péclét, 95
 de Reynolds, 29
 de Stokes, 29, 103
 Non orienté (graphe), 224

 Ordre (d'un graphe), 225
 Orienté (graphe), 223
 Ostrogradsky (théorème), 84

Parallélogramme (identité, 359
 Particulaire (dérivée), 14, 17
 Plancherel (théorème de), 258
 Poincaré (inégalité de), 253
 Poincaré généralisée (inégalité de), 254
 Poincaré (inégalité), 42
 Point
 d'équilibre , 239
 stationnaire , 239
 Point fixe (Brouwer), 375
 Point fixe (Picard), 375
 Point-selle, 269, 270
 Poiseuille's Law, 219
 Poisson (problème de), 92, 255, 344
 Pondéré (graphe), 224
 PoumStab, 418
 Poupée (russe), 397
 Principe
 des travaux virtuels, 277
 du maximum, 124, 126
 Problème
 de Poisson, 92, 255, 344
 Projection, 359
 Prolongement (opérateur), 251
 Pyramide des âges, 15
 Péclet (nombre de), 95
 Radiatif (Forçage), 32
 Rellich (Théorème), 253
 Resistance
 équivalente, 48
 Resistance équivalente, 75
 Reynolds (nombre de), 29, 212
 Riesz-Fréchet (théorème de représentation de), 361
 Russe (poupée), 397
 Réseau
 charismatique, 140
 de transport, 115
 résistif, 39
 Séparabilité, 359
 Schéma d'Euler
 implicite, 283
 Simple (chemin), 226
 Simple (graphe), 224
 Sobolev (espaces de), 247
 Solution
 faible, 87, 198, 255
 Sous-différentiel, 313
 Sous-différentiel
 au sens de Fréchet, 274
 d'une fonction convexe, 274
 Stationnaire (point) , 239
 Stoke's equations, 219
 Stokes
 nombre, 103
 équation, 217
 Stokes (nombre de), 29, 103
 Stokes (équations), 212
 Sédimentation (vitesse de), 102
 Taux de reproduction
 comme spectre, 158
 de base, 150, 246
 individuel, 157
 Temps
 de parcours, 118
 Tenseur
 des contraintes, 206
 des taux de déformation, 211
 Tension surfacique, 78
 Théorème
 de point fixe de Picard, 375
 d'inversion locale, 379
 d'Ostrogradsky, 84
 de Cauchy Lipschitz, 237
 de Courant-Fisher, 369
 de Hahn-Banach, 361
 de Krein-Milman, 376
 de la divergence, 84
 de Lax-Milgram, 363
 de Plancherel, 258
 de point fixe de Brouwer, 375
 de Rellich, 253
 de représentation de Riesz-Fréchet, 361
 des fonctions implicites, 376
 Totale (dérivée), 14, 17
 Tournoi, 231
 Trace
 d'une fonction de H^1 , 251
 Traceur passif, 29, 35
 Trafic
 piéton, 13, 197
 routier, 13, 197
 Trajectoire, 17
 Transport
 branché, 59
 Triangulation, 354
 Valuation 2-adique, 396
 Variable
 extensive , 17, 129
 intensive, 17, 129
 Variationnelle (formulation), 43
 Vide (graphe), 224, 226
 Viscosité, 212
 Wardrop (équilibre), 118