Modélisation macroscopique de mouvements de foule à deux types, modèles SIR condensés

Félicien Bourdin

▶ **To cite this version:**

Félicien Bourdin. Modélisation macroscopique de mouvements de foule à deux types, modèles SIR condensés. Equations aux dérivées partielles [math.AP]. Université Paris-Saclay, 2022. Français. NNT : 2022UPASM013 . tel-03783598

HAL Id: tel-03783598

https://theses.hal.science/tel-03783598v1

Submitted on 22 Sep 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

université PARIS-SACLAY

Modélisation macroscopique de mouvements de foule à deux types, modèles SIR condensés

Macroscopic Modeling of the Motion of a Crowd with Two Types, Condensed SIR Models

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°574 : École Doctorale de Mathématiques Hadamard (EDMH)
Spécialité de doctorat : Mathématiques Appliquées
Graduate School : Mathématiques, Référent : Faculté des sciences d'Orsay

Thèse préparée dans les unités de recherche **Département de mathématiques et applications (ENS, CNRS)** et **Laboratoire de mathématiques d'Orsay (Université Paris-Saclay, CNRS)**, sous la direction de **Bertrand MAURY**, professeur et la co-direction de **Gabriel PERYE**, directeur de recherche.

Thèse soutenue à Paris, le 13 juillet 2022, par

Félicien BOURDIN

Composition du jury

Noureddine IGBIDA Professeur des Universités, Université de Limoges	Président
Vincent CALVEZ Directeur de recherche, Université Claude Bernard Lyon1	Rapporteur & Examinateur
Clément CANCES Chargé de recherche, INRIA Lille, HDR	Rapporteur & Examinateur
Anne-Laure DALIBARD Professeure, Sorbonne Université	Examinatrice
Quentin MERIGOT Professeur, Université Paris-Saclay	Examinateur
Bertrand MAURY Professeur associé, École Normale Supérieure	Directeur de thèse

THESE DE DOCTORAT
NNT : 2022UPASM013


1794
ENS
ÉCOLE NORMALE
SUPÉRIEURE


université
PARIS-SACLAY
ÉCOLE DOCTORALE
de mathématiques
Hadamard (EDMH)


DMA
UMR 8553
Département de mathématiques et applications
de l'École normale supérieure


cnrs


erc


Fondation mathématique
FMJH
Jacques Hadamard

Remerciements

Je remercie en premier lieu Bertrand Maury, qui m'a accompagné au long de cette thèse malgré un contexte parfois difficile, sans jamais faillir à un encadrement rigoureux, stimulant et toujours bienveillant. Son aide constante aura été précieuse sur le plan professionnel aussi bien que sur le plan personnel. Je remercie également Gabriel Peyré - sans qui cette thèse n'aurait pas été possible - pour ses conseils avisés et les discussions enrichissantes que j'ai pu avoir avec lui. Je remercie par ailleurs chaudement Sylvain Faure pour sa disponibilité permanente pour toutes mes questions d'implémentation et les nombreuses heures passées en visio ou au téléphone qu'il a consenti à m'accorder.
Je tiens à remercier Vincent Calvez et Clément Cancès pour le temps qu'ils ont passé à rapporter ma thèse, ainsi que les autres membres de mon jury, Anne-Laure Dalibard, Noureddine Igbida et Quentin Mérigot.

Je remercie l'ensemble du personnel de l'Université Paris-Saclay et de l'École Normale Supérieure pour avoir contribué à cette thèse en me garantissant un espace de travail de qualité. J'ai une pensée particulière pour Fabienne Renia et Zaina Elmir qui m'ont accompagné dans toutes mes démarches administratives ; leur assistance m'a été primordiale. Je remercie tout·e·s ceux·elles avec qui j'ai pu échanger lors de cette thèse de doctorat, en particulier toutes les personnes qui ont travaillé à mes côtés - officiellement ou non - dans le bureau du DMA, et notamment Yann qui a souvent ranimé ma motivation et ma confiance en moi vacillantes. Merci à Max de m'avoir donné 7 euros pour figurer dans mes remerciements.

Enfin, je tiens à remercier ma famille pour leur soutien, notamment mes parents pour m'avoir insufflé une curiosité et une persévérance sans lesquelles mon parcours universitaire n'aurait pas été si riche. Je remercie l'ensemble de mes ami·e·s et proches pour leur soutien psychologique indéfectible. Je vous dois beaucoup dans l'aboutissement de cette thèse.

This thesis was funded by the ERC project NORIA.

Table des Matières

Chapitre 1

Introduction

On étudie dans cette thèse deux problématiques indépendantes : d'une part la modélisation macroscopique du mouvement d'une foule divisée en deux types d'individus différents, de l'autre l'élaboration de modèles d'épidémiologie adaptés à l'étude de la dynamique infectieuse dans une école.

1.1 Modélisation des mouvements de foules

La modélisation des comportements collectifs - d'humains, de cellules ou de particules - donne lieu à une grande variété d'approches en fonction des hypothèses de modélisation utilisées pour décrire la population. On distingue d'une part les modèles "microscopiques", où chaque entité est numérotée, repérée et suivie lors de sa trajectoire dans l'espace. Le système est alors usuellement décrit par un ensemble d'équations différentielles ordinaires portant sur la position de chaque individu. Cette modélisation est dite "lagrangienne" car elle donne accès à l'ensemble des trajectoires individuelles. Très pratiques pour des foules relativement peu denses, les modèles microscopiques deviennent difficiles d'utilisation lorsque l'effectif devient très important, car ils impliquent le calcul d'un grand nombre d'interactions (d'ordre du carré de la taille de la population).
Les modèles dits "macroscopiques" décrivent en revanche une population par sa densité, c'est-à-dire par l'information d'un nombre d'entités par unité de surface. Le système est alors encodé par une équation aux dérivées partielles portant sur la densité de population, de type transport :

$$\partial_t \rho + \nabla \cdot (\rho u) = 0, \tag{1.1}$$

où u est la vitesse "eulérienne" de la foule dans l'espace. Le calcul du déplacement d'une foule arbitrairement dense est alors possible, mais la donnée de chaque trajectoire n'est plus inhérente au modèle.

1.1.1 Congestion

Un phénomène majeur intervenant dans le mouvement des foules est le principe de congestion, traduisant l'impossibilité des particules à s'interpénétrer ou se déformer à l'échelle microscopique. On fait alors la distinction entre la vitesse souhaitée des individus (ou vitesse spontanée dans le cas de particules ou de cellules) et la vitesse effective résultant de la prise en compte de la congestion dans le système. Cette vitesse est un compromis entre l'ensemble de la population pour avoir une vitesse la plus proche de sa vitesse souhaitée sans pour autant qu'il y ait interpénétration ou superposition. Cette congestion peut être introduite dans le système de différentes façons. Certains modèles encodent la contrainte d'impénétrabilité de manière molle, en pénalisant les configurations impossibles. Pour les modèles microscopiques, un fort potentiel répulsif à courte portée entre particules est souvent introduit afin de prévenir l'interpénétration, comme dans [34] dans le contexte du déplacement des piétons, ou [8, 64] pour la collision des solides. Pour les modèles macroscopiques, on peut ajouter à la vitesse un potentiel d'interaction répulsif à noyau, de la forme $u = \nabla(G * \rho)$ où G est un noyau répulsif d'interaction à courte portée (voir par exemple [48, 49] pour des choix de noyau). L'équation des milieux poreux [20, 37, 55], où la vitesse spontanée est de la forme $m\rho^{m-2}\nabla\rho$, par sa structure de flot gradient pour la fonctionnelle $\int \rho^m$, peut également être vue comme une version molle de la congestion pénalisant les hautes valeurs de la densité.
Dans ce qui suit, on s'intéresse aux modèles où la congestion est gérée de manière forte. Dans ces modèles qui n'autorisent aucune violation de la contrainte, on astreint la vitesse effective à préserver la contrainte de congestion (pas de pénétration dans les modèles microscopiques, une densité en deçà d'une valeur maximale pour les modèles macroscopiques). Pour ce faire, on projette à tout moment la vitesse souhaitée sur l'ensemble des vitesses admissibles qui respectent la congestion. La complexité de ce type de modèle réside alors dans le fait de transporter les particules par un champ de vitesse qui est la projection à chaque instant des vitesses spontanées sur l'ensemble des vitesses permises par la configuration instantanée de la population. Il est néanmoins démontré dans [46] que le modèle macroscopique de mouvement de foule à un type introduit dans [45] entre dans le cadre du "sweeping process" introduit par Moreau dans [51], et est donc bien posé.

1.1.2 Modèles à deux types

Un des enjeux principaux de cette thèse est l'étude de l'extension du modèle macroscopique de mouvement de foule sous contrainte de congestion introduit dans [45] au cas d'une population à deux types. Les écoulements multiphasiques ont été largement étudiés en physique (voir [13, 24]), même sous contrainte de densité maximale [17], mais ne peuvent être appliqués aux mouvements de foule car ils induisent une séparation de phase et non un mélange total des types dans le milieu. Si la généralisation d'un modèle microscopique à une population bipartie est immédiate - il suffit a priori de définir une hétérogénéité

dans les vitesses spontanées - le modèle macroscopique se prête moins directement à cette généralisation. En particulier, la contrainte portant sur la somme des densités de chaque type ne définit plus un ensemble géodésiquement convexe dans l'espace de Wasserstein, rendant plus difficile la projection sur l'ensemble des paires de densités admissibles.

1.2 Modélisation d'épidémie en milieu scolaire

1.2.1 Modèles SIR

La récente pandémie de Covid-19 a popularisé l'utilisation et la compréhension de modèles dits "SIR", décrivant la propagation d'une épidémie dans une population initialement saine. Dans le modèle original, introduit dans l'article historique [36], la population est divisée en trois catégories : Sain·e·s (S) - Infecté·e·s (I) - Remis·e·s (R). Sous des hypothèses de panmixie de la population, on modélise la dynamique infectieuse par le système d'équations différentielles ordinaires suivant :

$$\begin{cases} \dfrac{\mathrm{d}S}{\mathrm{d}t} &= -\beta SI \\ \dfrac{\mathrm{d}I}{\mathrm{d}t} &= \beta SI - \gamma I \\ \dfrac{\mathrm{d}R}{\mathrm{d}t} &= \gamma I \end{cases} \tag{1.2}$$

où $\gamma > 0$ est le taux de rémission de la maladie, et $\beta > 0$ est le produit de la probabilité de rencontre par paire d'individus par unité de temps et de la probabilité d'infection par rencontre. Dans le cas d'une grande population et au début de l'émergence de l'épidémie, on définit le taux de reproduction de la maladie par $R_0 = \dfrac{\beta}{\gamma}$. Si $R_0 > 1$, il y a croissance exponentielle au début de l'infection jusqu'à un maximum puis décroissance exponentielle ; si $R_0 < 1$, l'épidémie décroît exponentiellement dès son apparition. D'autres modèles à compartiments plus raffinés font apparaître des catégories E ("exposed", sont infecté·e·s mais pas encore infectieux·ses) ou A (asymptomatiques), voire D (pour "dead" quand la maladie peut être mortelle). Dans le cas d'une maladie qui ne confère pas d'immunité, on parle de modèle SIS (les infecté·e·s qui se remettent deviennent à nouveau susceptibles d'être contaminé·e·s). On renvoie à [53] pour une introduction générale des différents modèles d'épidémie à compartiments.

1.2.2 Écoles

L'hypothèse de mélange uniforme de la population sous-jacente au modèle SIR (1.2) apparaît invalide dans un établissement fortement structuré tel qu'une école. On souhaite écrire un

modèle respectueux de cette structure en classes ; on suppose donc que la population est divisée en un ensemble de particules - ou communautés -, groupes de personnes que l'on suppose rester ensemble lors de l'emploi du temps et avoir les mêmes contacts au sein de la population. Une telle communauté est typiquement une classe, ou une partie d'une classe ayant les mêmes options (par exemple la partie de la 6eme1 ayant pour LV2 l'espagnol et mangeant à la cantine). On inclut également dans le modèle des particules singletons, dans le cas de professeur·e·s ou d'agent·e·s.

1.2.3 Modèles

Deux points de vue sont possibles lors de la modélisation d'épidémie dans une population dont on connaît le parcours au cours de la journée. On peut envisager de définir l'état de chaque personne comme une variable aléatoire à 3 états (S, I et R) et de définir un modèle de transmission probabiliste de l'épidémie, avec infection à probabilité fixée lors d'un contact entre une personne saine et une infectée, rémission suivant une variable aléatoire exponentielle, comme dans [40, 72]. On obtient des résultats avec ce type de modèle par des arguments de type Monte-Carlo, en moyennant sur un grand nombre de tirages du modèle. On cherche ici à construire un modèle en l'absence du graphe exact des contacts, mais où l'on dispose de données de croisement entre les différentes communautés. On suppose que l'on a typiquement accès au temps de croisement entre chaque paire de particules, d'où l'on dérive un nombre de contact de chacun·e avec des élèves de l'autre particule. De cette manière, on s'affranchit de différencier les élèves au sein d'une particule en choisissant un graphe de contact parmi d'autres. On choisit dès lors une modélisation déterministe de l'évolution de l'épidémie au sein de la population où l'on décrit l'évolution des fractions d'infecté·e·s au sein de chaque particule en fonction des contacts journaliers. Les quantités épidémiologiques sont définies à l'échelle du groupe de personnes, comme dans [1, 18]. L'avantage d'un tel modèle est notamment son temps de calcul : on n'a plus besoin d'itérer sur les réalisations d'un modèle aléatoire et on ne calcule plus les interactions entre chaque paire d'individus mais entre chaque paire de particules.

1.3 Contributions

On étudie dans la première partie différentes problématiques relatives à l'étude de la modélisation macroscopique des mouvements de foule, notamment dans le cas où la foule est divisée en types distincts.

Le chapitre 2 est consacré à l'extension formelle du modèle initial introduit dans [45] à une population à deux types (problème 2), lorsque la vitesse confère au problème une structure de flot gradient. À l'instar de l'analyse du modèle à un type, on introduit en section 2.3 un schéma de type JKO et pointe les difficultés de la preuve de sa convergence

vers une solution du modèle macroscopique à deux types. On présente alors en section 2.4 une implémentation efficace de ce schéma qui utilise une version vectorielle du transport optimal régularisé.

On étudie dans le chapitre 3 l'extension d'un schéma de splitting introduit dans [46] permettant de considérer n'importe quel champ de vitesse (définition 1). On étudie la projection dans l'espace de Wasserstein sur l'ensemble des paires de densités à somme majorée et on établit un résultat permettant son calcul à partir de l'opérateur de projection sur l'ensemble des densités admissibles dans le cas monotype (proposition 1, corollaire 1). On montre alors le caractère bien posé du schéma pour des temps arbitrairement grands (théorème 1), et on présente une implémentation de celui-ci dans la section 3.4.

Le chapitre 4 est dédié à l'étude d'un schéma de volumes finis pour lequel un résultat de convergence des vitesses vers la vitesse effective peut être établi en dimension 1 (théorème 2). On étudie alors l'extension de ce théorème à des cas plus complexes (dimension 2, multi-type) - le résultat est a priori négatif car on utilise un théorème de régularité propre à la dimension 1 (théorème 3) et des arguments propres au cas saturé. On analyse enfin en section 4.7 une implémentation numérique de ce schéma.

Le chapitre 5 est consacré à l'étude de résultats partiels d'homogénéisation de modèles microscopiques vers les modèles macroscopiques. On montre que la projection sur l'ensemble des vitesses admissibles dans le cas macroscopique à un type (i.e. l'ensemble des vitesses à divergence positive) ne peut pas être obtenue génériquement comme limite d'un opérateur de projection microscopique (voir section 5.3). On exhibe alors des situations très structurées - où les particules sont placées sur un réseau - où une procédure d'homogénéisation est possible (proposition 9). Le travail présenté dans ce chapitre a été réalisé en collaboration avec Bertrand Maury et a fait l'objet d'une publication (voir [16]).

Enfin, le chapitre 6 présente une application des modèles macroscopiques à l'identification de paramètres. Étant donné une vidéo de mouvements de foule, on essaye d'inférer le champ de vitesse sous-jacent au déplacement global. Pour ce faire, on considère le modèle comme une fonction qui à un champ de vitesses souhaitées associe une vidéo. Sous ce formalisme, le problème d'identification de paramètres devient un problème inverse qui peut être approché numériquement en dérivant formellement à l'aide de la librairie python *pytorch* à travers le modèle. On identifie dans un premier temps un potentiel paramétrique pour un modèle microscopique (section 6.1) avant de pointer les problèmes d'apprentissage d'un potentiel non-paramétrique dans un cadre macroscopique dans la section 6.2, que l'on contourne en paramétrisant le champ inconnu par un réseau de neurones.

La deuxième partie est dédiée à l'élaboration, l'analyse et l'implémentation d'un modèle SIR de propagation d'une épidémie dans une école. Ces travaux font suite à une sollicitation de MODCOV19[1] nous proposant de travailler sur ce sujet conjointement à Bertrand Maury et Sylvain Faure (Laboratoire de Mathématiques d'Orsay).

[1] https://modcov19.math.cnrs.fr/

On définit dans le chapitre 7 le modèle SIR (7.6) dit "condensé", dont la granularité est intermédiaire, i.e. définie à l'échelle du groupe d'individus. On compare ce modèle déterministe à des modèles microscopiques probabilistes pour montrer que le modèle n'exprime pas l'évolution de la loi de probabilité d'infection d'un modèle aléatoire défini à l'échelle individuelle (section 7.2).

Le chapitre 8 analyse théoriquement le procédé de condensation du modèle SIR condensé, mais aussi de modèles de propagation de courant dans des réseaux électriques (l'analogie entre les deux types de modèles étant motivée par un opérateur elliptique discret de type laplacien qui intervient dans les deux cas). Plus précisément, on compare le modèle microscopique défini sur un graphe donné au modèle condensé défini sur le graphe plus grossier où l'on a regroupé certains ensembles de nœuds (proposition 21 pour les réseaux électriques, proposition 23 pour le modèle SIR). On étudie alors des conditions sur le graphe et les nœuds que l'on a choisi de condenser pour que ces deux modèles prédisent des résultats similaires.

Le chapitre 9 présente l'implémentation effective du modèle condensé défini dans le chapitre 7. Cette implémentation prend la forme d'une application web destinée à des chef·fe·s d'établissement désireux·se·s de comparer différents emploi du temps en termes de risques épidémiologiques. L'interface permet d'entrer les plans de l'établissement et l'emploi du temps. Les déplacements sont alors reconstitués, les contacts calculés et un modèle SIR condensé est ensuite simulé pour définir un score de risque ainsi que des visuels permettant une analyse plus fine de la dynamique épidémiologique au sein de l'école.

Partie I

Modélisation macroscopique de mouvements de foule

Chapitre 2

Modélisation, premier algorithme

On introduit dans ce chapitre le modèle macroscopique de mouvements de foule pour une foule composée de deux types. Chaque type a une vitesse souhaitée qui lui est propre. L'écriture du modèle est directement l'adaptation du modèle à un type introduit dans [45] : on projette à chaque instant les vitesses souhaitées sur l'ensemble des vitesses tendant à faire diminuer la densité globale sur la zone saturée $\rho_1 + \rho_2 = 1$. La contrainte portant sur la somme des densités n'étant pas géodésiquement convexe, l'existence ne peut pas être montrée avec les mêmes arguments que dans le cas monotype. En particulier, on ne peut pas montrer la convergence des vitesses induites par le schéma JKO vers la projection sur l'ensemble des vitesses admissibles. On propose néanmoins une implémentation efficace de ce schéma utilisant un algorithme de calcul vectorisé du transport optimal régularisé.

2.1 Motivations

2.1.1 Mouvement de cellules

La modélisation de la coopération entre des organismes unicellulaires ayant des comportements sociaux est importante pour comprendre la formation d'organismes multicellulaires dans l'apparition de la vie. *Dictyostelium Discoideum* est une amibe unicellulaire souvent qualifiée d'"amibe sociale". En effet, une colonie présente un comportement collectif lorsqu'elle est placée en état de stress nutritionnel : elle secrète de l'AMPc, un chemioattractant qui provoque l'agrégation de la colonie en un pseudo-plasmode, sorte de limace constituée de milliers de cellules et pouvant migrer à la recherche de nourriture. Une partie de cette limace se différencie alors en spores qui se répandent afin d'amorcer un cycle de reproduction. On réfère à [39] pour une introduction détaillée à *Dictyostelium Discoideum*. Son cycle de migration/reproduction est représenté sur la figure 2.1.

La compréhension du processus de différenciation entre cellules spores et cellules dites "stalks", qui dirigent la colonie lors de la migration est un sujet actif de recherche en écologie. Le développement de modèles macroscopiques multi-types permet un calcul global de

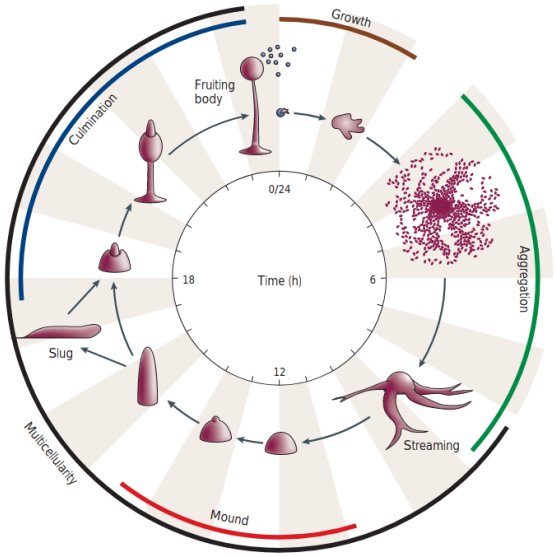


Figure 2.1: Cycle de vie de *Dictyostelium Discoideum*, extrait de [23].

l'interaction d'un grand nombre de particules sans augmentation de la complexité avec la densité, au contraire d'une modélisation microscopique qui encode la donnée de chaque interaction entre paire de cellules. On verra en section 2.2 qu'on peut inclure dans la modélisation différents effets, allant de la chemoattraction à l'interaction courte ou longue portée en passant par l'attraction/répulsion par un champ extérieur.

2.1.2 Croisements de foules

Si les modèles microscopiques incluent naturellement des différences comportementales entre les individus d'une foule, le modèle macroscopique de mouvement de foule à un type [45] est construit sur une population d'individus indiscernables. La modélisation d'un croisement de deux foules ne peut donc être intégrée directement au modèle macroscopique de mouvement de foule. Le développement d'un modèle multi-type permettrait de garder les avantages d'un calcul efficace du mouvement d'une foule composée d'un grand nombre d'individus, tout en introduisant des hétérogénéités de comportement au sein de la population.

2.2 Modèle à un type, modèle à deux types

2.2.1 Modèle à un type

On décrit ici le modèle macroscopique de mouvement de foule, introduit dans [45] dont on étudiera l'extension à deux types. On se donne un domaine borné $\Omega \subset \mathbb{R}^2$, $\rho_0 \in \mathbb{L}^\infty(\Omega)$ une

densité initiale de population, et $U \in \mathbb{L}^2(\Omega, \mathbb{R}^2)$ une vitesse souhaitée de déplacement. En l'absence de congestion, le modèle de mouvement de foule s'écrit :

$$\begin{cases} \partial_t \rho + \nabla \cdot (\rho U) = 0 \\ \qquad \rho(t=0) = \rho_0. \end{cases} \tag{2.1}$$

Afin de modéliser la saturation de la foule (en effet, l'équation (2.1) n'interdit pas la densité d'atteindre des valeurs arbitrairement grandes), on souhaite imposer une contrainte de densité maximale $\rho \leqslant \rho_M$. Dans la suite, on prendra par convention $\rho_M = 1$. Étant donné une densité ρ, on peut écrire l'ensemble des vitesses qui conduisent à une densité admissible :

$$C_\rho = \left\{ v \in \mathbb{L}^2(\Omega, \mathbb{R}^2),\ \nabla \cdot v \geqslant 0 \text{ où } \rho = 1 \right\}. \tag{2.2}$$

La positivité de la divergence d'un champ de $\mathbb{L}^2(\Omega)$ est à comprendre au sens faible : on veut que pour tout champ de pression test $q \in \mathbb{H}^1(\Omega)$ positif, et non nul uniquement sur la zone saturée $\{\rho = 1\}$, on ait

$$\int_\Omega v \cdot \nabla q \leqslant 0. \tag{2.3}$$

On note $\mathbb{H}^1_\rho(\Omega)$ un tel ensemble de pressions. Le modèle macroscopique de mouvement de foule s'écrit alors

Problème 1 (Modèle macroscopique de mouvement de foule)**.** *Pour $T > 0$, une densité initiale $\rho_0 \in \mathbb{L}^\infty(\Omega)$ admissible, $U \in \mathbb{L}^2(\Omega, \mathbb{R}^2)$, on cherche $\rho \in \mathbb{L}^2([0,T] \times \Omega)$ qui vérifie*

$$\begin{cases} \partial_t \rho + \nabla \cdot (\rho u) &= 0 \\ u &= P_{C_\rho}(U) \\ \rho(t=0) &= \rho_0, \end{cases} \tag{2.4}$$

où P_{C_ρ} est l'opérateur de projection orthogonale dans $\mathbb{L}^2(\Omega, \mathbb{R}^2)$ sur C_ρ.

Il n'est pas immédiat que le problème 1 soit bien posé, car la vitesse par laquelle est transportée la densité ρ est elle-même la projection de la vitesse souhaitée sur l'ensemble des vitesses admissibles, qui dépend de la densité en tout temps. Il est cependant prouvé dans [45] que dans le cas d'une vitesse souhaitée de type $U = -\nabla D$ où D est une fonction λ-convexe, l'équation a une structure de flot-gradient dans l'espace de Wasserstein $\mathbb{W}_2(\Omega)$. En étudiant la convergence d'un schéma de type JKO, les auteur·e·s montrent l'existence d'une solution faible au problème 1.

2.2.2 Modèle à deux types

On veut à présent décrire le mouvement d'une foule à deux types par un modèle macroscopique. Pour un espace d'évolution ouvert $\Omega \subset \mathbb{R}^2$, on veut modéliser l'évolution de

$\rho_1, \rho_2 \in \mathbb{L}^\infty(\Omega)$, en intégrant une contrainte de congestion maximale portant sur la densité totale en tout point : $\rho_1 + \rho_2 \leqslant \rho_M$, avec $\rho_M > 0$ une constante (dans la suite, on suppose $\rho_M = 1$). On suppose qu'en tout point de Ω une cellule de type j est mue par une vitesse $U_j \in \mathbb{L}^2(\Omega, \mathbb{R}^2)$. Étant donné deux densités ρ_1, ρ_2 admissibles, le cône des vitesses conduisant à un respect de la contrainte est donné par :

$$C_{\rho_1,\rho_2} = \left\{ u_1, u_2 \in \mathbb{L}^2(\Omega, \mathbb{R}^2), \nabla \cdot (\rho_1 u_1 + \rho_2 u_2) \geqslant 0 \text{ où } \rho_1 + \rho_2 = 1 \right\}. \tag{2.5}$$

On souhaite donc étudier le problème suivant :

Problème 2. *Pour $\rho_1^0, \rho_2^0 \in \mathbb{L}^\infty(\Omega)$ tels que $\rho_1^0 + \rho_2^0 \leqslant 1$ et pour $U_1, U_2 \in \mathbb{L}^2(\Omega, \mathbb{R}^2)$, on cherche $\rho_1, \rho_2 \in \mathbb{L}^2([0,T] \times \Omega)$ telles que faiblement*

$$\begin{cases} \partial_t \rho_1 + \nabla \cdot (\rho_1 u_1) &= 0 \\ \partial_t \rho_2 + \nabla \cdot (\rho_2 u_2) &= 0 \\ (u_1, u_2) &= P_{C_{\rho_1,\rho_2}}(U_1, U_2) \\ \rho_1(t=0) &= \rho_1^0, \\ \rho_2(t=0) &= \rho_2^0 \end{cases} \tag{2.6}$$

où $P_{C_{\rho_1,\rho_2}}$ est la projection au sens $\mathbb{L}^2$ sur C_{ρ_1,ρ_2}.

Remarque 1. *On verra par la suite qu'une grande différence entre les problèmes 1 et 2 est que l'ensemble des densités admissibles dans le cas à deux espèces n'est pas géodésiquement convexe dans l'espace produit $\mathbb{W}_2(\Omega) \times \mathbb{W}_2(\Omega)$, où $\mathbb{W}_2(\Omega)$ est l'espace des mesures de probabilités sur Ω muni de la distance de Wassestein W_2, ce qui rend difficile la projection d'une paire de densités quelconque sur cet ensemble (même s'il est fermé). En revanche, C_{ρ_1,ρ_2} étant un cône convexe fermé, l'opérateur $P_{C_{\rho_1,\rho_2}}$ est toujours bien défini.*

2.3 Convergence du schéma JKO

On reprend à présent la preuve d'existence du problème d'évolution pour une espèce présente dans [45]. Si le schéma JKO est bien posé et permet de dériver les mêmes estimées que dans [45], on ne peut pas montrer directement que les vitesses limites obtenues sont bien la projection des vitesses souhaitées sur l'ensemble des vitesses admissibles.

On suppose que les champs de vitesses souhaitées sont de la forme $U_j = -\nabla D_j$, où les $D_j : \mathbb{R}^2 \longrightarrow \mathbb{R}$ sont des fonctions continues λ-convexes, et on se place dans $\mathbb{W}_2(\Omega)$ l'ensemble des mesures de probabilité sur Ω muni de la distance de Wasserstein W_2. On note $K_2 \subset \mathbb{W}_2(\Omega) \times \mathbb{W}_2(\Omega)$ l'ensemble des densités admissibles :

$$K_2 = \{(\rho_1, \rho_2) \in \mathbb{W}_2(\Omega), \rho_1 + \rho_2 \leqslant 1 \text{ p.p.}\}. \tag{2.7}$$

Commençons par un résultat qui garantit que le problème de projection est bien posé.

Proposition 1. *C_{ρ_1,ρ_2} est un cône convexe fermé : $P_{C_{\rho_1,\rho_2}}$ est donc bien définie.*

Démonstration. Le fait que C_{ρ_1,ρ_2} est un cône convexe est immédiat. Montrons que c'est un fermé : soit $(u_1^n, u_2^n)_n \in (C_{\rho_1,\rho_2})^{\mathbb{N}}$ qui tend vers (u_1, u_2). Soit $q \in \mathbb{H}^1_{\rho_1+\rho_2}(\Omega)$. On a

$$\int_\Omega (\rho_1 u_1 + \rho_2 u_2) \cdot \nabla q = \lim_{n\to\infty} \int_\Omega (\rho_1 u_1^n + \rho_2 u_2^n) \cdot \nabla q \leqslant 0. \tag{2.8}$$

□

On cherche à construire un schéma de type JKO, afin d'en étudier la convergence. Notons

$$J(\rho_1, \rho_2) = \int_\Omega D_1\rho_1 + D_2\rho_2 \tag{2.9}$$

la fonctionnelle que les champs de vecteurs U_j visent à minimiser, et définissons le schéma suivant, de pas de temps $\tau > 0$:

$$\begin{cases} (\rho_1^{k,\tau}, \rho_2^{k,\tau}) & \in & \underset{(\rho_1,\rho_2)\in K_2}{\operatorname{argmin}} J(\rho_1, \rho_2) + \dfrac{1}{2\tau}\left(W_2^2\left(\rho_1, \rho_1^{k-1,\tau}\right) + W_2^2\left(\rho_2, \rho_2^{k-1,\tau}\right)\right) \\ \rho_1^{0,\tau}, \rho_2^{0,\tau} & = & \rho_1^0, \rho_2^0. \end{cases} \tag{2.10}$$

On définit également pour $j = 1, 2$:

- $t_j^{k,\tau}$ l'unique plan de transport de $\rho_j^{k,\tau}$ à $\rho_j^{k-1,\tau}$,
- $v_j^{k,\tau} = \dfrac{Id - t_j^{k,\tau}}{\tau}$ la vitesse,
- $E_j^{k,\tau} = \rho_j^{k,\tau} v_j^{k,\tau}$ la quantité de mouvement,
- $\rho_j^\tau, v_j^\tau, E_j^\tau$ les interpolées constantes par morceaux de quantités précédentes.

On cherche à montrer la convergence étroite de $\left(\rho_j^\tau, E_j^\tau\right)$ vers une limite de la forme $(\rho_j, \rho_j u_j)$. Commençons par montrer que le schéma est bien défini :

Lemme 1. *Soit $(\mu_1, \mu_2) \in K_2$. Alors*

$$\phi(\rho_1, \rho_2) := J(\rho_1, \rho_2) + \mathbf{1}_{K_2}(\rho_1, \rho_2) + \frac{1}{2\tau}\left(W_2^2(\rho_1, \mu_1) + W_2^2(\rho_2, \mu_2)\right) \tag{2.11}$$

admet un minimum (ρ_1^m, ρ_2^m). Il existe alors deux potentiels de Kantorovich $\overline{\varphi}_1, \overline{\varphi_2}$ associés au transport des ρ_j^m vers les μ_j, tels que pour tout (ρ_1, ρ_2) admissibles,

$$\sum_{j=1}^2 \int_\Omega \rho_j \left(D_j + \frac{\overline{\varphi_j}}{\tau}\right) \geqslant \sum_{j=1}^2 \int_\Omega \rho_j^m \left(D_j + \frac{\overline{\varphi_j}}{\tau}\right). \tag{2.12}$$

Démonstration. On adapte la démonstration de [45] ; il n'y a cependant plus unicité des minimiseurs de ϕ.
ϕ est convexe s.c.i. sur un compact donc elle admet des minimiseurs (ρ_1^m, ρ_2^m). On suppose dans un premier temps que $\mu_j > 0$, pour $j = 1, 2$: les mêmes arguments que dans [45] montrent la formule d'optimalité (2.12). Pour le cas général, on se donne $(\rho_1^{m,\delta}, \rho_2^{m,\delta})$ minimiseur de

$$\begin{aligned}\psi_\delta(\rho_1, \rho_2) := J(\rho_1, \rho_2) + \mathbf{1}_{K_2}(\rho_1, \rho_2) + \frac{1}{2\tau}\left(W_2^2\left(\rho_1, \mu_1^\delta\right) + W_2^2\left(\rho_2, \mu_2^\delta\right)\right) \\ + c_\delta\left(W_2^2(\rho_1, \rho_1^m) + W_2^2(\rho_2, \rho_2^m)\right)\end{aligned} \tag{2.13}$$

où $\left(\mu_1^\delta, \mu_2^\delta\right)$ sont admissibles, positives presque-partout et convergent vers (μ_1, μ_2) et $c_\delta > 0$ tend vers 0 quand δ tend vers 0.
Montrons que $\rho_j^{m,\delta}$ tend vers ρ_j^m pour $j = 1, 2$: supposons qu'il existe $\epsilon > 0$ tel que $W_2^2(\rho_j^{m,\delta}, \rho_j^m) \geqslant \epsilon$ pour $j = 1$ ou 2. On a alors $\psi_\delta(\rho_1^{m,\delta}, \rho_2^{m,\delta}) \geqslant \phi_\delta(\rho_1^{m,\delta}, \rho_2^{m,\delta}) + \epsilon c_\delta$, avec

$$\phi_\delta(\rho_1, \rho_2) = J(\rho_1, \rho_2) + \mathbf{1}_{K_2}(\rho_1, \rho_2) + \frac{1}{2\tau}\left(W_2^2\left(\rho_1, \mu_1^\delta\right) + W_2^2\left(\rho_2, \mu_2^\delta\right)\right). \tag{2.14}$$

On a donc, par optimalité de $(\rho_1^{m,\delta}, \rho_2^{m,\delta})$ par rapport à (ρ_1^m, ρ_2^m) :

$$\epsilon c_\delta + \phi_\delta(\rho_1^{m,\delta}, \rho_2^{m,\delta}) \leqslant \psi_\delta(\rho_1^m, \rho_2^m) = \phi_\delta(\rho_1^m, \rho_2^m). \tag{2.15}$$

Or, on a l'estimation

$$|\phi_\delta(\rho_1, \rho_2) - \phi(\rho_1, \rho_2)| \leqslant \sum_{j=1}^{2} \frac{1}{2\tau}\left(W_2(\rho_j, \mu_j) + W_2\left(\rho_j, \mu_j^\delta\right)\right) W_2\left(\mu_j, \mu_j^\delta\right). \tag{2.16}$$

On peut donc écrire

$$\epsilon c_\delta + \phi(\rho_1^{m,\delta}, \rho_2^{m,\delta}) \leqslant \phi(\rho_1^m, \rho_2^m) + h_\delta \tag{2.17}$$

où h_δ tend vers 0 et ne dépend pas de c_δ. On fixe maintenant $c_\delta \gg h_\delta$ pour aboutir à une contradiction.
D'après le premier cas, on a pour toute paire (ρ_1, ρ_2) admissible

$$\sum_{j=1}^{2} \int_\Omega D_j\left(\rho_j - \rho_j^{m,\delta}\right) + \frac{1}{\tau}\int_\Omega \overline{\varphi}_j^\delta\left(\rho_j - \rho_j^{m,\delta}\right) + o(\delta) \geqslant 0 \tag{2.18}$$

avec $\overline{\varphi}_j^\delta$ potentiel de Kantorovich de $\rho_j^{m,\delta}$ à μ_j. Par continuité des potentiels de Kantorovich et d'après la convergence précédemment établie, on peut passer à la limite en $\delta \longrightarrow 0$ pour dériver la condition d'optimalité (2.12). □

Remarque 2. *On n'a plus unicité des minimiseurs à chaque étape. Cela vient du fait qu'on n'a plus de propriété de convexité le long des géodésiques généralisées dans le cas d'un minimiseur sous la forme d'une paire. On peut néanmoins définir le schéma en prenant n'importe que minimiseur de la fonctionnelle JKO à chaque pas.*

On dérive à présent de ces conditions d'optimalité une formulation de type Darcy faisant apparaître une pression.

Lemme 2. *Il existe* $p^{k,\tau} \in \mathbb{H}^1_{\rho_1^{k,\tau}+\rho_2^{k,\tau}}(\Omega)$ *telle que :*

- *Sur* $\left\{\rho_1^{k,\tau}=0\right\} \cap \left\{\rho_2^{k,\tau}>0\right\}$, $U_2 = v_2^{k,\tau} + \nabla p^{k,\tau}$.
- *Sur* $\left\{\rho_1^{k,\tau}>0\right\} \cap \left\{\rho_2^{k,\tau}=0\right\}$, $U_1 = v_1^{k,\tau} + \nabla p^{k,\tau}$.
- *Sur* $\left\{\rho_1^{k,\tau}>0\right\} \cap \left\{\rho_2^{k,\tau}>0\right\}$, $U_1 - v_1^{k,\tau} = U_2 - v_2^{k,\tau} = \nabla p^{k,\tau}$.

Démonstration. D'après le lemme précédent, on a

$$\left(\rho_1^{k,\tau}, \rho_2^{k,\tau}\right) \in \underset{(\rho_1,\rho_2)\in K_2}{\operatorname{argmin}} \int_\Omega F_1\rho_1 + F_2\rho_2 \tag{2.19}$$

avec $F_j = D_j + \dfrac{\overline{\varphi}_j}{\tau}$, pour $j=1,2$, soit encore

$$\left(1-\rho_1^{k,\tau}-\rho_2^{k,\tau}, \rho_1^{k,\tau}, \rho_2^{k,\tau}\right) \in \underset{\rho_0+\rho_1+\rho_2=1}{\operatorname{argmin}} \int_\Omega 0\rho_0 + F_1\rho_1 + F_2\rho_2. \tag{2.20}$$

D'après le théorème B.1. dans [17], il existe $\alpha_0, \alpha_1, \alpha_2 \in \mathbb{R}$ tels qu'en posant

$$\lambda(x) = \min_{j=0,1,2} F_j(x) + \alpha_j, \tag{2.21}$$

on a pour $j=0,1,2$

$$\lambda(x) = F_j(x) + \alpha_j \text{ sur } \{\rho_j > 0\}. \tag{2.22}$$

Soit $p = \alpha_0 - \lambda$; p est positive presque-partout et $p=0$ si $\rho_1^{k,\tau}+\rho_2^{k,\tau}<1$. Calculons à présent le gradient de p. D'après le lemme 5.2.24 de [3], pour $v \in \mathbb{H}^1_0(\Omega)$, $\nabla(v_+) = \nabla v \mathbf{1}_{v>0}$. On applique ce lemme à

$$p = \alpha_0 - \left(\alpha_0 - (\alpha_1+F_1) + (\alpha_1+F_1-\alpha_2-F_2)_+\right)_+ \tag{2.23}$$

pour établir (au sens des distributions) :

$$\nabla p = \begin{cases} 0 & \text{si } \alpha_0 \leqslant \min\left(F_1+\alpha_1, F_2+\alpha_2\right), \\ -\nabla F_1 & \text{si } F_1+\alpha_1 \leqslant \min\left(\alpha_0, F_2+\alpha_2\right), \\ -\nabla F_2 & \text{si } F_2+\alpha_2 \leqslant \min\left(\alpha_0, F_1+\alpha_1\right). \end{cases} \tag{2.24}$$

On a donc $\nabla p = -\nabla F_j$ sur $\{\rho_j > 0\}$: le gradient de p est dans $\mathbb{L}^2(\Omega)$ et $p \in \mathbb{H}^1_{\rho_1^{k,\tau}+\rho_2^{k,\tau}}(\Omega)$. □

On poursuit la démonstration d'existence de [45] et on définit à présent $\tilde{\rho}_j^\tau$ les interpolées entre les $(\rho_j^{k,\tau})$ le long des géodésiques de $\mathbb{W}_2(\Omega)$. Conformément au théorème 5.14 de [62], on note $\tilde{v}_j^\tau$ les champs de vitesses advectant ces géodésiques, et on pose $\tilde{E}_j^\tau = \tilde{\rho}_j^\tau \tilde{v}_j^\tau$. Comme dans [45], on obtient la borne :

$$\sum_{k=1}^{N}\sum_{j=1}^{2} \tau \left(\frac{W_2\left(\rho_j^{k,\tau}, \rho_j^{k-1,\tau}\right)}{\tau} \right)^2 \leqslant C \tag{2.25}$$

où $C > 0$ est une constante indépendante de τ. On en déduit les estimations suivantes, en notant p^τ l'interpolation constante par morceaux de $p^{k,\tau}$:

Lemme 3. *Pour $j = 1, 2$,*

- *v_j^τ est τ-uniformément bornée dans $L^2\left([0,T], L^2_{\rho_j^\tau}\left(\Omega, \mathbb{R}^2\right)\right)$,*
- *p^τ est τ-uniformément bornée dans $L^2\left([0,T], \mathbb{H}^1\left(\Omega\right)\right)$,*
- *E_j^τ et $\tilde{E}_j^\tau$ sont des mesures τ-uniformément bornées.*

Démonstration. La démonstration découle de l'estimation (2.25) et est la même que celle du lemme 3.4 de [45]. □

Grâce à ces estimées, on peut extraire une limite faible (ρ_j, E_j) aux mesures uniformément bornées $\left(\tilde{\rho}_j^\tau, \tilde{E}_j^\tau\right)$. On applique les mêmes arguments que dans [45] pour établir que $\left(\rho_j^\tau, E_j^\tau\right)$ tendent vers la même limite (ρ_j, E_j) et qu'il existe une vitesse limite telle que $E_j = \rho_j u_j$; on montre de même qu'il existe une pression limite dans $\mathbb{H}^1_{\rho_1+\rho_2}(\Omega)$.
Cependant, contrairement au cas mono-espèce, (u_1, u_2) n'est pas immédiatement la projection de (U_1, U_2) sur C_{ρ_1,ρ_2}. En effet, dans le cas à une espèce, on peut écrire

$$E^\tau = -\rho^\tau U - \rho^\tau \nabla p^\tau = -\rho^\tau U - \nabla p^\tau, \tag{2.26}$$

car la pression est nulle là où $\rho < 1$. Le premier membre tend vers ρu, le second vers $-\rho U - \nabla p = -\rho U - \rho \nabla p$: on a la formulation de Darcy $u = U - \nabla p$ sur $\{\rho = 1\}$ et on peut donc déduire que u est la projection de U sur l'ensemble des vitesses admissibles. Dans le cas à deux espèces, on n'a pas $E_j^\tau = -\rho_j^\tau U_j - \nabla p^\tau$ mais seulement $E_1^\tau + E_2^\tau = -\rho_1^\tau U_1 - \rho_2^\tau U_2 - \nabla p^\tau$, d'où l'on déduit :

$$\rho_1 u_1 + \rho_2 u_2 = \rho_1 \left(U_1 - \nabla p\right) + \rho_2 \left(U_2 - \nabla p\right). \tag{2.27}$$

Par conséquent, la simplification utilisée pour passer à la limite la relation entre le champ souhaité, la vitesse effective et la pression n'est plus valide : le produit $\rho_j^\tau \nabla p^\tau$ ne converge pas directement vers $\rho_j \nabla p$. En effet, on a

- $\nabla p^\tau \rightharpoonup \nabla p$ dans $\mathbb{L}^2\left([0,T], \mathbb{L}^2(\Omega)\right)$
- $\rho_j^\tau \overset{*}{\rightharpoonup} \rho_j$ dans $\mathbb{L}^1([0,T]\times\Omega)$ et $\rho_j^\tau \to \rho_j$ dans $\mathbb{W}_2(\Omega)$,

ce qui n'est pas suffisant pour la convergence, même faible, du produit $\rho_j^\tau \nabla p^\tau$.

2.4 Aspects numériques

2.4.1 Schéma JKO

Le schéma JKO, uniquement convergent a priori dans le cadre monotype, peut néanmoins être utilisé comme un schéma numérique pour calculer des solutions à (2.6). Pour ce faire, on réécrit l'équation (2.10) sous la forme d'un unique problème de minimisation portant sur les plans de transport de $\rho_j^{k,\tau}$ à $\rho_j^{k-1,\tau}$:

$$\begin{cases} \gamma_1^{k,\tau}, \gamma_2^{k,\tau} & \in \underset{\gamma_1,\gamma_2\in\mathcal{M}(\Omega\times\Omega)}{\operatorname{argmin}} J\left(\pi^x_\#\gamma_1, \pi^x_\#\gamma_2\right) + \dfrac{1}{2\tau}\left(\displaystyle\int_{\Omega\times\Omega}\|x-y\|^2\,\mathrm{d}\gamma_1 + \int_{\Omega\times\Omega}\|x-y\|^2\,\mathrm{d}\gamma_2\right) \\ & \qquad\qquad +\mathbf{1}_{\left(\pi^x_\#\gamma_1,\pi^x_\#\gamma_2\right)\in K_2} + \mathbf{1}_{\left(\pi^y_\#\gamma_1,\pi^y_\#\gamma_2\right)=\left(\rho_1^{k-1,\tau},\rho_2^{k-1,\tau}\right)} \\ \rho_j^{k,\tau} & = \pi^x_\#\gamma_j^{k,\tau} \\ \rho_1^{0,\tau}, \rho_2^{0,\tau} & = \rho_1^0, \rho_2^0. \end{cases} \tag{2.28}$$

Discrétisons à présent Ω en N cases. On peut réécrire simplement l'équation d'évolution (2.28) sous forme matricielle, en notant $\mathbf{q_j^{k,\tau}}$ la discrétisation de $\rho_j^{k,\tau}$ sur le maillage, $\mathbf{\Lambda_j^{k,\tau}}$ le plan de transport discret entre $\mathbf{q_j^{k,\tau}}$ et $\mathbf{q_j^{k-1,\tau}}$, $\mathbf{C}$ la matrice de coût de transport, $\mathbf{D_j}$ le vecteur potentiel qui discrétise D_j et $\mathbf{p}\in\mathbb{R}^N$ le vecteur composé de 1 permettant par multiplication de projeter sur une marginale :

$$\begin{cases} \mathbf{\Lambda_1^{k,\tau}}, \mathbf{\Lambda_2^{k,\tau}} & \in \underset{\mathbf{\Lambda_1},\mathbf{\Lambda_2}\in\mathbb{R}^{N^2}}{\operatorname{argmin}} \langle \mathbf{D_1}, \mathbf{p\Lambda_1}\rangle + \langle \mathbf{D_2}, \mathbf{p\Lambda_2}\rangle + \dfrac{1}{2\tau}\langle \mathbf{C}, \mathbf{\Lambda_1}+\mathbf{\Lambda_2}\rangle \\ & \qquad\qquad +\mathbf{1}_{\mathbf{p}(\mathbf{\Lambda_1}+\mathbf{\Lambda_2})\leqslant \mathbf{p}^T} + \mathbf{1}_{(\mathbf{\Lambda_1}\mathbf{p}^T, \mathbf{\Lambda_2}\mathbf{p}^T)=(\mathbf{q_1^{k-1,\tau}}, \mathbf{q_2^{k-1,\tau}})} \\ \mathbf{q_j^{k,\tau}} & = \mathbf{p\Lambda_j^{k,\tau}}. \end{cases} \tag{2.29}$$

On voit alors apparaître un problème d'optimisation linéaire, qui peut être résolu numériquement - par exemple par l'algorithme du simplexe ou par un algorithme de point intérieur.

L'inconvénient de ces méthodes de calcul est qu'elles sont lentes et dégénèrent quand la taille du maillage augmente.

2.4.2 Schéma JKO régularisé

Conformément à ce qui est fait dans [57] dans le cadre monotype, on peut améliorer l'efficacité du calcul de l'approche JKO développée ci-dessus en considérant une régularisation entropique du coût Wasserstein dans (2.10), c'est-à-dire en définissant la distance régularisée par entropie entre deux mesures discrètes $\mathbf{q}$ et $\mathbf{r}$ par

$$W_\alpha(\mathbf{q},\mathbf{r}) = \min_{\mathbf{\Lambda}\in\Pi_{\mathbf{q},\mathbf{r}}} \langle \mathbf{C},\mathbf{\Lambda}\rangle + \alpha E(\mathbf{\Lambda}) \tag{2.30}$$

où $\alpha>0$ est un paramètre de régularisation, $E(\mathbf{\Lambda}) = \sum_{i,j=1}^{N} \mathbf{\Lambda_{i,j}}(\log(\mathbf{\Lambda_{i,j}})-1)$ est une entropie et $\Pi_{\mathbf{q},\mathbf{r}}$ l'ensemble des plans de transport entre $\mathbf{q}$ et $\mathbf{r}$. L'avantage de l'introduction d'une telle entropie est le calcul rapide d'un pas du schéma JKO par un algorithme itératif, voir [57]. Conformément au cas à un type, on réécrit la version régularisée de (2.29) sous la forme

$$\underset{\bar{\mathbf{\Lambda}}=(\mathbf{\Lambda_1},\mathbf{\Lambda_2})}{\operatorname{argmin}} \ \mathrm{Div}_\Gamma(\bar{\mathbf{\Lambda}}\,|\,\bar{\xi}) + \varphi_1(\bar{\mathbf{\Lambda}}) + \varphi_2(\bar{\mathbf{\Lambda}}) \tag{2.31}$$

où Div_Γ est la divergence de Bregman associée à $\Gamma(\bar{\mathbf{\Lambda}}) = E(\mathbf{\Lambda_1}) + E(\mathbf{\Lambda_2})$, $\bar{\xi} = (\xi,\xi)$ avec $\xi = e^{-\mathbf{C}/\alpha}$ et

$$\begin{aligned} \varphi_1(\bar{\mathbf{\Lambda}}) &= \mathbf{1}_{(\mathbf{\Lambda_1 p}^T,\mathbf{\Lambda_2 p}^T)=(\mathbf{q_1^{k-1,\tau}},\mathbf{q_2^{k-1,\tau}})} \\ \varphi_2(\bar{\mathbf{\Lambda}}) &= 2\tau\,\langle \mathbf{D_1},\mathbf{p\Lambda_1}\rangle + \langle \mathbf{D_2},\mathbf{p\Lambda_2}\rangle + \mathbf{1}_{\mathbf{p}(\mathbf{\Lambda_1}+\mathbf{\Lambda_2})\leqslant \mathbf{p}^T}. \end{aligned} \tag{2.32}$$

On utilise alors les lemmes suivants, pendants à deux types des propositions 3.2 et 2 dans [57] :

Lemme 4. *Pour tout $\bar{\mathbf{\Lambda}}\in(\mathbb{R}^{N\times N})^2$, on a*

$$\begin{aligned} \mathrm{Prox}^{\mathrm{Div}_\Gamma}_{\varphi_1}(\bar{\mathbf{\Lambda}}) &= \bar{\mathbf{\Lambda}}\mathrm{Diag}\left(\frac{(\mathbf{q_1^{k-1,\tau}},\mathbf{q_2^{k-1,\tau}})}{(\mathbf{\Lambda_1 p}^T,\mathbf{\Lambda_2 p}^T)}\right) \\ \mathrm{Prox}^{\mathrm{Div}_\Gamma}_{\varphi_2}(\bar{\mathbf{\Lambda}}) &= \mathrm{Diag}\left(\frac{\mathrm{Prox}^{\mathrm{Div}_\Gamma}_{\frac{\tau f}{\alpha}}(\mathbf{p\Lambda_1},\mathbf{p\Lambda_2})}{(\mathbf{p\Lambda_1},\mathbf{p\Lambda_2})}\right)\bar{\mathbf{\Lambda}} \end{aligned} \tag{2.33}$$

où

$$f:\begin{cases} \mathbb{R}^N\times\mathbb{R}^N & \longrightarrow \quad \mathbb{R} \\ (\mathbf{p_1},\mathbf{p_2}) & \longmapsto \quad \mathbf{1}_{\mathbf{p_1}+\mathbf{p_2}\leqslant 1} + \langle \mathbf{D_1},\mathbf{p_1}\rangle + \langle \mathbf{D_2},\mathbf{p_2}\rangle\,. \end{cases} \tag{2.34}$$

Lemme 5. *Pour tous* $\mathbf{p_1}, \mathbf{p_2} \in (\mathbb{R}^N)^2$*, on a*

$$\mathrm{Prox}^{\mathrm{Div}_\Gamma}_{\sigma f}(\mathbf{p_1}, \mathbf{p_2}) = \frac{\min\left(\mathbf{p_1} \odot e^{-\sigma \mathbf{D_1}} + \mathbf{p_2} \odot e^{-\sigma \mathbf{D_2}}, 1\right)}{\mathbf{p_1} \odot e^{-\sigma \mathbf{D_1}} + \mathbf{p_2} \odot e^{-\sigma \mathbf{D_2}}} \left(\mathbf{p_1} \odot e^{-\sigma \mathbf{D_1}}, \mathbf{p_2} \odot e^{-\sigma \mathbf{D_2}}\right). \tag{2.35}$$

Lemme 6. *Il existe* $\mathbf{a_1}, \mathbf{a_2}, \mathbf{b_1}, \mathbf{b_2} \in \mathbb{R}^N$ *tels que la solution* $(\mathbf{\Lambda_1}, \mathbf{\Lambda_2})$ *du problème régularisé* (2.31) *vérifie*

$$\begin{aligned} \mathbf{\Lambda_1} &= \mathrm{Diag}(\mathbf{a_1}) \xi \mathrm{Diag}(\mathbf{b_1}) \\ \mathbf{\Lambda_2} &= \mathrm{Diag}(\mathbf{a_2}) \xi \mathrm{Diag}(\mathbf{b_2}). \end{aligned} \tag{2.36}$$

L'algorithme d'itérations de Dykstra s'écrit alors

Itérations de Dykstra pour deux types

- On part de $\mathbf{a_j^0} = \mathbf{b_j^0} = \mathbf{u_j^0} = \mathbf{v_j^0}$ pour $j = 1, 2$,
- Pour l impair,

$$\mathbf{a_j^l} = \mathbf{a_j^{l-1}} \odot \mathbf{u_j^{l-2}} \tag{2.37}$$

$$\mathbf{b_j^l} = \frac{\mathbf{q_j^{k-1,\tau}}}{\xi^{\mathbf{T}} \mathbf{a_j^l}}, \tag{2.38}$$

- Pour l pair,

$$\mathbf{b_j^l} = \mathbf{b_j^{l-1}} \odot \mathbf{v_j^{l-2}} \tag{2.39}$$

$$\mathbf{a_j^l} = \frac{\mathrm{Prox}^{\mathrm{Div}_\Gamma}\left(\mathbf{a_1^{l-1}} \odot \mathbf{u_1^{l-1}} \odot \xi \mathbf{b_1^l}, \mathbf{a_2^{l-1}} \odot \mathbf{u_2^{l-1}} \odot \xi \mathbf{b_2^l}\right)}{\xi \mathbf{b_j^l}}, \tag{2.40}$$

- Et dans tous les cas

$$\mathbf{u_j^l} = \mathbf{u_j^{l-2}} \frac{\mathbf{a_j^{l-1}}}{\mathbf{a_j^l}} \tag{2.41}$$

$$\mathbf{v_j^l} = \mathbf{v_j^{l-2}} \frac{\mathbf{b_j^{l-1}}}{\mathbf{b_j^l}}. \tag{2.42}$$

Une implémentation de cet algorithme est présentée sur la figure 2.2 : on a simulé le mouvement de deux disques de population initialement superposés, l'un attiré par leur centre, l'autre repoussé. On voit que les deux types se séparent : l'un se concentre au centre et l'autre s'agrège sur les bords de l'espace disponible. L'implémentation en python

de l'algorithme JKO calculé par transport régularisé à un ou deux types est disponible en ligne[1].

2.4.3 Limites

L'algorithme précédemment présenté comporte deux inconvénients majeurs. L'un est qu'il exploite la structure de flot gradient de l'équation lorsque le champ des vitesses souhaité est de type $-\nabla D$, où D est une fonction potentiel indépendante de ρ. Dans l'optique de modéliser des phénomènes plus complexes - notamment en biologie - on aimerait pouvoir introduire dans le modèle des vitesses souhaitées dépendant du champ de densité lui même, par exemple du type $U = \nabla c$ où c est la concentration d'un chemoattractant vérifiant $-\kappa\Delta c = \rho$, voire $U = \nabla(V * \rho)$, où V est un noyau d'interaction.
Le deuxième point négatif vient du phénomène de "freezing", relevé dans [61] : le choix d'un petit pas de temps fait exploser l'importance de la partie Wasserstein dans (2.29). Lorsqu'on fait tendre le pas de temps vers 0 à pas d'espace fixé, la pénalisation par le coût force les plans de transport à prendre une forme triviale $\mathbf{\Lambda_j} = \mathrm{Id}_{N,N}$, ce qui aboutit à un bloquage et aucune évolution des densités.

[1] https://gitlab.math.u-psud.fr/bourdin/jko-crowd-motion-model-for-two-typed-population

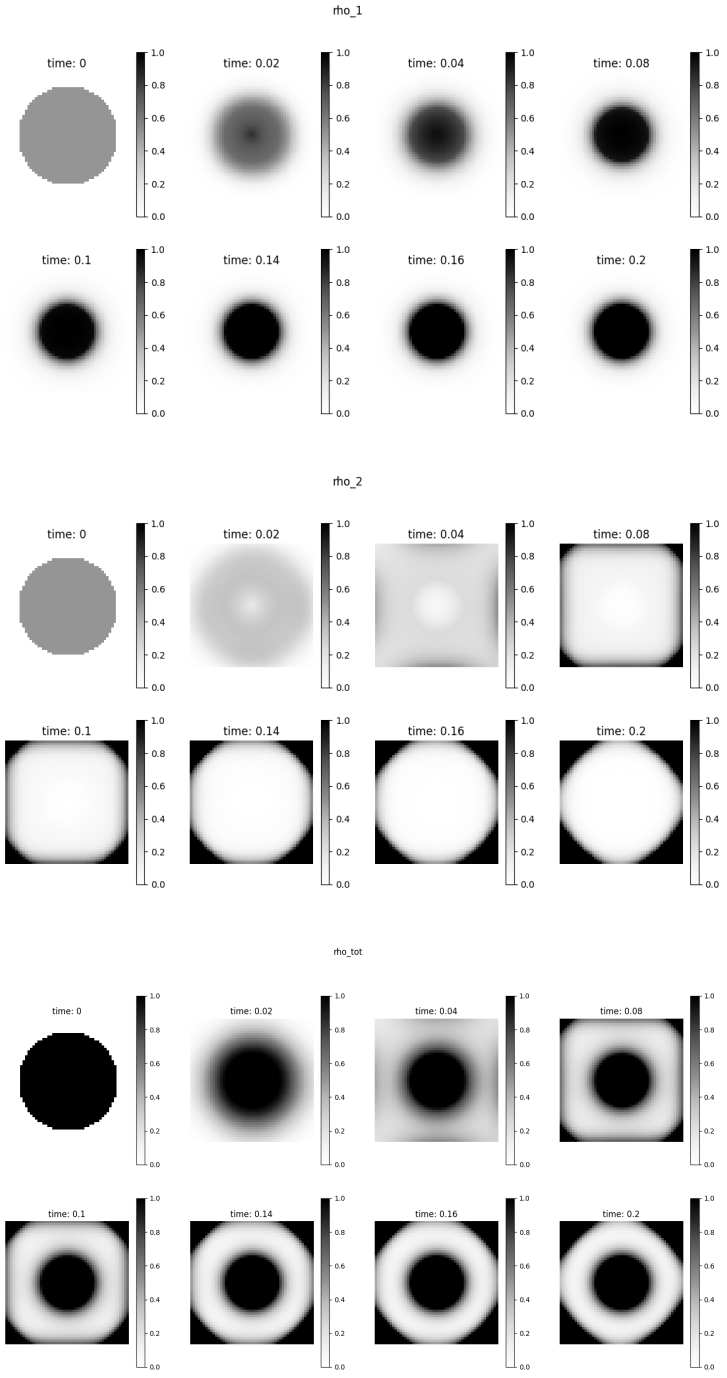


Figure 2.2: Le mouvement de deux disques initialement superposés, l'un formé d'une population attirée par le centre, l'autre formé d'une population repoussée par celui-ci.

Chapitre 3

Splitting

On étudie dans ce chapitre l'adaptation d'un algorithme dit de "splitting" introduit dans [46] à une population à deux types. Comme l'ensemble des paires de mesures à densité dont la somme est inférieure à 1 n'est plus géodésiquement convexe dans l'espace de Wasserstein, la deuxième étape de projection dans le pas du schéma n'est pas automatiquement bien posée. On montre que la projection de mesures à densité est bien définie et qu'elle peut être calculée en projetant la masse totale sur l'ensemble des mesures à densité bornée par 1. On montre ensuite le caractère bien posé du schéma pour des temps arbitrairement longs. On dérive enfin un schéma numérique permettant d'intégrer des vitesses plus génériques que le schéma JKO introduit dans le chapitre 2.

3.1 Présentation du schéma

On cherche à développer un schéma de splitting qui approche le problème 2 du chapitre 2. Conformément au cadre monotype présenté dans [46], on découple un pas de temps en deux étapes :

- **Prédiction** : on applique pendant un temps τ les champs de vitesses souhaitées, indifféremment du fait qu'elles conduisent à des configurations qui violent la contrainte de congestion,
- **Correction** : on projette les deux densités ainsi obtenues sur l'ensemble des densités admissibles.

Définition 1. *Étant donnés $\rho_1^0, \rho_2^0 \in \mathbb{W}_2(\Omega)$ tels que $\rho_1^0 + \rho_2^0 \leqslant 1$, $U_1, U_2 \in \mathbb{L}^2(\Omega, \mathbb{R}^2)$ et un pas de temps $\tau > 0$, on définit le schéma de splitting à deux types par*

$$\begin{cases} \mu_j^{k+1} = (Id + \tau U_j)_{\#}\rho_j^k \quad \textit{pour } j = 1, 2 & \textit{(prédiction)} \\ (\rho_1^{k+1}, \rho_2^{k+1}) \in \underset{(\rho_1,\rho_2)\in K_2}{\operatorname{argmin}} W_2^2(\rho_1, \mu_1^{k+1}) + W_2^2(\rho_2, \mu_2^{k+1}) & \textit{(correction)} \end{cases} \tag{3.1}$$

où $K_2 = \{(\rho_1, \rho_2) \in \mathbb{W}_2(\Omega), \rho_1 + \rho_2 \leqslant 1\}$ *est l'ensemble des densités admissibles.*

On rappelle le théorème 3.5 de [46] dans le cas monotype : si le champ de vitesse souhaitée est C^1, l'interpolation constante par morceaux entre les densités du schéma converge faiblement quand τ tend vers 0 vers une solution du problème d'évolution correspondant. La preuve reposant sur les mêmes estimées et arguments que celle de la convergence du schéma JKO, on a le même problème de convergence faible que pointé dans le chapitre 2 si on souhaite l'adapter dans le cadre à deux types.

Remarque 3. *Le minimum qui définit l'étape de correction* (3.1)*, s'il est forcément atteint comme minimum d'une fonction semi-continue inférieurement sur un compact, n'est pas forcément atteint en un seul point comme c'était le cas avec un seul type. Cela vient du fait que* K_2 *n'est pas géodésiquement convexe dans* $\mathbb{W}_2(\Omega) \times \mathbb{W}_2(\Omega)$ *muni de la distance produit*

$$\mathrm{d}((\mu_1, \mu_2), (\nu_1, \nu_2))^2 = \mathrm{d}(\mu_1, \nu_1)^2 + \mathrm{d}(\mu_2, \nu_2)^2, \tag{3.2}$$

comme on le voit dans l'exemple 1.

Remarque 4. *Soit* (μ_1, μ_2) *et* $(\nu_1, \nu_2) \in \mathbb{W}_2(\Omega) \times \mathbb{W}_2(\Omega)$*. Considérons deux géodésiques de* $\mathbb{W}_2(\Omega)$*, de vitesse constante* $x_t : [0,1] \to \mathbb{W}_2(\Omega)$*,* $y_t : [0,1] \to \mathbb{W}_2(\Omega)$*, respectivement de* μ_1 *à* ν_1 *et de* μ_2 *à* ν_2*. Alors* $z_t = (x_t, y_t)$ *est une géodésique de vitesse constante de* (μ_1, μ_2) *à* (ν_1, ν_2) *dans* $\mathbb{W}_2(\Omega) \times \mathbb{W}_2(\Omega)$*.*

Exemple 1. *Soit* $\mu_1 = \nu_2$ *la fonction indicatrice de* $B(0, \epsilon)$*, et* $\mu_2 = \nu_1$ *la fonction indicatrice de* $B(1, \epsilon)$*. D'après la remarque précédente, la géodésique de vitesse constante menant de* (μ_1, μ_2) *à* (ν_1, ν_2) *est*

$$\phi : \left\{ \begin{array}{ccc} [0,1] & \longrightarrow & \mathbb{W}_2(\Omega) \times \mathbb{W}_2(\Omega) \\ t & \longmapsto & \left(\mathbf{1}_{B(t,\epsilon)}, \mathbf{1}_{B(1-t,\epsilon)}\right). \end{array} \right. \tag{3.3}$$

On voit sur la figure 3.1 qu'une telle géodésique ne reste pas dans K_2 *pour* $t = 1/2$*.*

On va néanmoins montrer que la projection d'une paire de mesure à densité est uniquement définie, et que le schéma de splitting (3.1) est bien posé pour des temps arbitrairement longs.

3.2 Propriétés de la projection sur K_2

Commençons par un lemme qui relie la projection sur K_2 à la projection de la densité totale sur

$$K_1 = \{\rho \in \mathbb{W}_2(\Omega), \rho \leqslant 1\}. \tag{3.4}$$

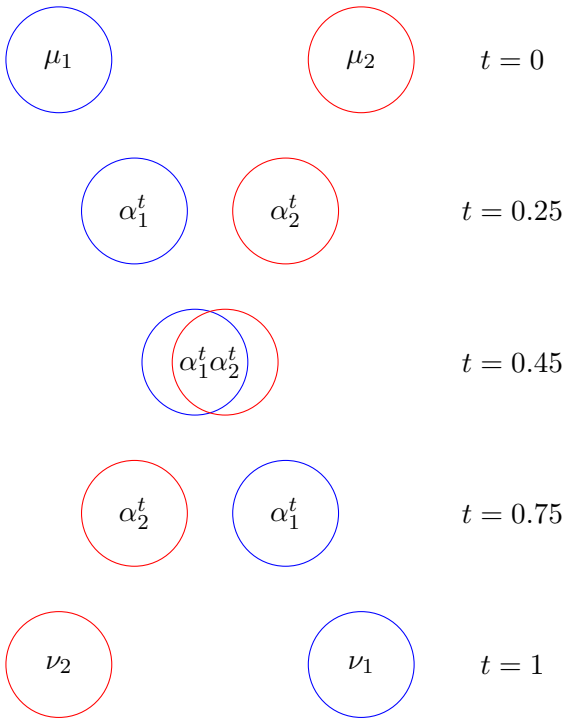


Figure 3.1: L'interpolation entre deux configurations de sphères opposées le long de géodésiques de $\mathbb{W}_2(\Omega) \times \mathbb{W}_2(\Omega)$. En particulier, pour $t = 0.45$ la paire (α_1^t, α_2^t) n'est pas dans K_2.

Lemme 7. *Soit $\mu_1, \mu_2 \in \mathbb{W}_2(\Omega)$. On note ρ la projection de $\mu_1 + \mu_2$ sur K_1. Soit*

$$(\rho_1, \rho_2) \in \underset{(\nu_1,\nu_2)\in K_2}{\operatorname{argmin}} W_2^2(\nu_1, \mu_1) + W_2^2(\nu_2, \mu_2). \tag{3.5}$$

Alors $\rho_1 + \rho_2 = \rho$ et

$$W_2^2(\rho_1 + \rho_2, \mu_1 + \mu_2) = W_2^2(\rho_1, \mu_1) + W_2^2(\rho_2, \mu_2). \tag{3.6}$$

Démonstration. Par définition de ρ, on a

$$W_2^2\left(\rho, \mu_1 + \mu_2\right) \leqslant W_2^2\left(\rho_1 + \rho_2, \mu_1 + \mu_2\right). \tag{3.7}$$

Soit T le plan de transport optimal de $\rho_1 + \rho_2$ à $\mu_1 + \mu_2$, T_1 le plan de transport optimal de ρ_1 à μ_1 et T_2 de ρ_2 à μ_2. Soit $\gamma = (Id, T_1)_{\#}\,\rho_1 + (Id, T_2)_{\#}\,\rho_2$. Les marginales de γ sont respectivement $\rho_1 + \rho_2$ et $\mu_1 + \mu_2$. Le coût quadratique de γ est $W_2^2\left(\rho_1, \mu_1\right) + W_2^2\left(\rho_2, \mu_2\right)$. On a donc

$$W_2^2\left(\rho_1 + \rho_2, \mu_1 + \mu_2\right) \leqslant W_2^2\left(\rho_1, \mu_1\right) + W_2^2\left(\rho_2, \mu_2\right). \tag{3.8}$$

Soit ζ un plan de transport optimal (au sens de Kantorovich) entre $\mu_1 + \mu_2$ et ρ. On a $\mu_1 \ll \mu_1 + \mu_2$, $\mu_2 \ll \mu_1 + \mu_2$: soit f et g les dérivées de Radon-Nikodym correspondantes et

$$\begin{aligned} \mathrm{d}\zeta_1(x, y) &= f(x)\,\mathrm{d}\zeta(x, y) \\ \mathrm{d}\zeta_2(x, y) &= g(x)\,\mathrm{d}\zeta(x, y). \end{aligned} \tag{3.9}$$

On a $\zeta_1 + \zeta_2 = \zeta$, $(\pi^x)_{\#}\zeta_1 = \mu_1$ et $(\pi^x)_{\#}\zeta_2 = \mu_2$ (où π^x et π^y sont les applications projection sur les coordonnées). En notant $\nu_j = (\pi^y)_{\#}\zeta_j$ pour $j = 1, 2$, on obtient

$$\begin{aligned} W_2^2\left(\rho, \mu_1 + \mu_2\right) &= \int_{\Omega^2} |x - y|^2\,\mathrm{d}\zeta \\ &= \int_{\Omega^2} |x - y|^2\,\mathrm{d}\zeta_1 + \int_{\Omega^2} |x - y|^2\,\mathrm{d}\zeta_2 \\ &\geqslant W_2^2\left(\nu_1, \mu_1\right) + W_2^2\left(\nu_2, \mu_2\right). \end{aligned} \tag{3.10}$$

Par optimalité de la paire (ρ_1, ρ_2), on a

$$W_2^2\left(\rho_1, \mu_1\right) + W_2^2\left(\rho_2, \mu_2\right) \leqslant W_2^2\left(\rho, \mu_1 + \mu_2\right). \tag{3.11}$$

En combinant (3.7), (3.8), (3.11), on obtient $W_2^2\left(\rho, \mu_1 + \mu_2\right) = W_2^2\left(\rho_1 + \rho_2, \mu_1 + \mu_2\right)$. Comme la projection sur K_1 est uniquement définie (proposition 2 dans [46]), on obtient $\rho = \rho_1 + \rho_2$. □

Remarque 5. *Par le lemme précédent, on obtient que la somme de deux densités dans K_2 qui réalisent la distance à une paire de mesure (μ_1, μ_2) peut être calculée en projetant $\mu_1 + \mu_2$ sur K_1. Il peut cependant y avoir de nombreuses paires qui réalisent cette distance.*

Considérons par exemple en dimension 1 *le cas* $\mu_1 = \mu_2 = \delta_0$. *On a* $P_{K_1}(\mu_1 + \mu_2) = \mathbf{1}_{[-1,1]}$. *Soit* $f, g \in \mathbb{L}^\infty([-1,1])$ *n'importe quelle paire de fonctions positives telles que* $f + g = 1$ *preque-partout. En posant alors*

$$\begin{aligned} \rho_1 &= f \,\mathrm{d}\lambda \\ \rho_2 &= g \,\mathrm{d}\lambda, \end{aligned} \tag{3.12}$$

où λ *est la mesure de Lebesgue sur* $[-1,1]$, *on a que* (ρ_1, ρ_2) *réalise la distance de* (μ_1, μ_2) *à* K_2.

On a néanmoins unicité lorsque toutes les mesures sont à densité.

Proposition 1. *Soit* $\mu_1, \mu_2 \in \mathbb{W}_2(\Omega)$ *deux mesures absolument continues par rapport à la mesure de Lebesgue. Il existe alors une unique paire* (ρ_1, ρ_2) *dans* K_2 *qui minimise* $W_2^2(\rho_1, \mu_1) + W_2^2(\rho_2, \mu_2)$.

Démonstration. Comme toutes les mesures sont à densité, tous les plans de transport au sens de Kantorovich le sont au sens de Monge. Soit T l'unique plan de transport optimal de $\mu_1 + \mu_2$ à $\rho := P_{K_1}(\mu_1 + \mu_2)$ et (ρ_1, ρ_2) qui minimise la distance à K_2. Soit T_1 et T_2 les plans de transport de μ_1 à ρ_1 et de μ_2 à ρ_2. Les T_j sont uniquement définis sur les supports des μ_j.
Montrons que $T_1 = T_2$ presque-partout. En dehors de $\mathrm{supp}(\mu_1) \cap \mathrm{supp}(\mu_2)$, on peut modifier T_1 ou T_2. Soit $\gamma = (Id, T_1)_\# \mu_1 + (Id, T_2)_\# \mu_2$: les marginales de γ sont $\mu_1 + \mu_2$ et $\rho_1 + \rho_2 = \rho$, d'après le lemme 7. Le coût de γ est

$$\begin{aligned} \int_\Omega |x-y|^2 d\gamma &= \int_\Omega |x - T_1(x)|^2 \,\mathrm{d}\mu_1 + \int_\Omega |x - T_2(x)|^2 \,\mathrm{d}\mu_2 \\ &= W_2^2(\rho_1, \mu_1) + W_2^2(\rho_2, \mu_2) \\ &= W_2^2(\rho, \mu_1 + \mu_2). \end{aligned} \tag{3.13}$$

γ est donc un plan de transport optimal entre ses marginales. Par unicité, γ est de la forme $(Id, T)_\#(\rho_1 + \rho_2)$. Soit $z \in \mathrm{supp}(\rho_1) \cap \mathrm{supp}(\rho_2)$. On a

- $(z, T_1(z)) \in \mathrm{supp}(\gamma)$
- $(z, T_2(z)) \in \mathrm{supp}(\gamma)$

donc $T_1 = T_2 = T$ p.p. et $\rho_j = T_\# \mu_j$ pour $j = 1, 2$. Les ρ_j sont donc uniquement déterminés. □

D'après la preuve précédente, on obtient la forme de la projection sur K_2 à partir de la projection de la somme des densités sur K_1 :

Corollaire 1. *Soit* $\mu_1, \mu_2 \in \mathbb{W}_2(\Omega)$ *deux mesures absolument continues par rapport à la mesure de Lebesgue. Soit* T *un plan de transport optimal de* $\mu_1 + \mu_2$ *à sa projection sur* K_1. *Alors la projection de* (μ_1, μ_2) *sur* K_2 *est donnée par* $(T_\# \mu_1, T_\# \mu_2)$.

L'étape de correction de (3.1) est ainsi uniquement définie, pourvu que les densités prédites soient absolument continues.

3.3 Caractère bien posé du schéma

La question restante pour que le schéma soit bien posé est la suivante : à quelle condition les densités prédites dans (3.1) sont-elles absolument continues par rapport à la mesure de Lebesgue ? En effet, on a montré que l'étape de projection était bien posée si les densités à projeter étaient absolument continues. Il faut donc montrer que l'étape de prédiction renvoie de telles mesures afin de pouvoir itérer le schéma de splitting. Commençons par une définition pratique.

Définition 2. *On définit l'ordre partiel sur l'ensemble des mesures positives sur Ω par*

$$\mu_1 \leqslant_* \mu_2 \iff \mu_1 \ll \mu_2 \text{ et } \frac{\mathrm{d}\mu_1}{\mathrm{d}\mu_2} \leqslant 1 \ \mu_2 - p.p. \tag{3.14}$$

En d'autres termes, $\mu_1 \leqslant_* \mu_2$ si $\mu_2 - \mu_1$ définit une mesure positive. Le lemme suivant contrôle l'image de la mesure de Lebesgue par l'étape de prédiction.

Lemme 8. *Soit Ω un ouvert (assez grand) de $\mathbb{R}^2$, U un champ de vitesse C^1 sur Ω et Ω_0 un compact inclus dans Ω. On suppose les dérivées partielles de U bornées par une constante $c > 0$. Alors pour tout pas de temps $\tau < (5c)^{-1}$, en notant λ la mesure de Lebesgue sur $\mathbb{R}^2$,*

$$(Id + \tau U)_{\#}\lambda_{|\Omega_0} \leqslant_* (1 + 5c\tau)\lambda_{|{\Omega_0}^{\tau\|U\|_{\mathbb{L}^\infty(\Omega_0)}}}. \tag{3.15}$$

où ${\Omega_0}^{\tau\|U\|_{\mathbb{L}^\infty(\Omega_0)}}$ est l'ensemble des points à distance inférieure à $\tau\|U\|_{\mathbb{L}^\infty(\Omega_0)}$ de Ω_0.

Démonstration. Soit $\tau > 0$ et

$$f : \left\{ \begin{array}{ccc} \Omega_0 & \longrightarrow & \mathbb{R}^2 \\ x & \longmapsto & x + \tau U(x). \end{array} \right. \tag{3.16}$$

f est C^1 et injective pour $\tau < c^{-1}$: soit $x, y \in \Omega$ tels que $f(x) = f(y)$. On a

$$x - y = \tau \int_0^1 \mathrm{Jac}_U((1-t)x + ty) \cdot (y - x)\,\mathrm{d}t \tag{3.17}$$

donc $\|x - y\| \leqslant c\tau\|x - y\|$ et $x = y$.
Soit A un borélien de Ω.

$$\begin{aligned} (Id + \tau U)_{\#}\lambda_{|\Omega_0}(A) &= \int_{\Omega_0} \mathbf{1}_A\left((Id + \tau U)(x)\right)\mathrm{d}\lambda(x) \\ &= \int_{\Omega_0} \mathbf{1}_A(f(x))\frac{|\det(\mathrm{Jac}_f(x))|}{|\det(\mathrm{Jac}_f(x))|}\,\mathrm{d}\lambda(x) \end{aligned} \tag{3.18}$$

On a $|\det(\mathrm{Jac}_f)| \geqslant 1 - 2c\tau - 2c^2\tau^2$. Pour $\tau < \dfrac{\sqrt{2}-1}{2c}$ on a

$$\frac{1}{1-2c\tau-2c^2\tau^2} \leqslant 1 + 4c\tau + 4c^2\tau^2. \tag{3.19}$$

Dès que $\tau \leqslant (4c)^{-1}$, on obtient $1 + 4c\tau + 4c^2\tau^2 \leqslant 1 + 5c\tau$ puis

$$\begin{aligned}(Id + \tau U)_{\#}\lambda_{|\Omega_0}(A) &\leqslant (1+5c\tau)\int_{\Omega_0} \mathbf{1}_A(f(x))|\det(\mathrm{Jac}_f(x))|\,\mathrm{d}\lambda(x)\\ &= (1+5c\tau)\int_{f(\Omega_0)} \mathbf{1}_A(x)\,\mathrm{d}\lambda(x)\\ &\leqslant (1+5c\tau)\lambda\left(A \cap \Omega_0^{\tau\|U\|_{\mathbb{L}^\infty(\Omega_0)}}\right).\end{aligned} \tag{3.20}$$

□

On est à présent en mesure d'établir le caractère bien posé du schéma de splitting à deux types (3.1).

Théorème 1. *Soit $U_1, U_2 \in C^1(\mathbb{R}^2, \mathbb{R}^2)$ deux champs de vitesses. On suppose qu'ils sont bornés par $c_1 > 1$, et que leurs dérivées partielles sont bornées par une constante $c_2 > 0$. Soient ρ_1^0, ρ_2^0 deux mesures absolument continues par rapport à la mesure de Lebesgue supportées par une boule $B(x, r_0)$, $\tau > 0$ un pas de temps inférieur à $(5c_2)^{-1}$, et un temps total T. Alors, en posant $\Omega = B(x, R)$ pour $R > 0$ assez grand, le schéma de splitting* (3.1) *est uniquement défini pour $n := \left\lceil \dfrac{T}{\tau} \right\rceil$ itérations.*

Démonstration. D'après le lemme 8, la première paire de densités prédites est supportée par $B(x, r_0 + \tau c_1)$ et sa somme a une densité majorée par $1 + 5c_2\tau$. D'après la proposition 1, le premier pas de correction est uniquement défini. On utilise alors le lemme suivant, dont la preuve figure en annexe.

Lemme 9. *Soit Ω un compact et μ_1, μ_2 deux mesures absolument continues telles que $\mu_1 \leqslant_* \mu_2$. Alors $P_{K_1}(\mu_1) \leqslant_* P_{K_1}(\mu_2)$.*

Par ce lemme et le corollaire 1, on obtient que les densités corrigées sont supportées par la boule $B(x, r_1)$, avec $r_1 = \sqrt{1 + 5c_2\tau}(r_0 + c_1\tau)$. On montre par récurrence que les densités sont bien définies à chaque pas et supportées au k-ième pas par une boule de rayon

$$r_k = (1+5c_2\tau)^{\frac{k}{2}}\left(r_0 + \frac{c_1\tau}{\sqrt{1+5c_2\tau}-1}\right) - \frac{c_1\tau}{\sqrt{1+c_2\tau}-1}. \tag{3.21}$$

On conclut la preuve en posant $R = r_n$. □

Finissons cette section théorique par un résultat de continuité pour la projection sur K_2.

Proposition 2. *Supposons Ω compact. Alors la restriction de la projection sur K_2 à l'ensemble des mesures absolument continues par rapport à la mesure de Lebesgue est continue.*

Démonstration. Soit $(\mu_1^n, \mu_2^n)_n$ qui converge vers (μ_1, μ_2). Notons

$$\begin{aligned}(\rho_1^n, \rho_2^n) &= P_{K_2}(\mu_1^n, \mu_2^n),\\ (\rho_1, \rho_2) &= P_{K_2}(\mu_1, \mu_2).\end{aligned} \tag{3.22}$$

Soit T^n un plan de transport optimal de $\mu_1^n + \mu_2^n$ à $\rho_1^n + \rho_2^n$ et T de $\mu_1 + \mu_2$ à $\rho_1 + \rho_2$. Comme P_{K_1} est continue (proposition 2 de [46]), $(\rho_1^n + \rho_2^n)_n$ tend vers $\rho_1 + \rho_2$. Le théorème 1.50 de [62] assure - comme Ω est compact - que $(Id, T^n)_\#(\mu_1^n + \mu_2^n)$ converge vers $(Id, T)_\#(\mu_1 + \mu_2)$. Soit ν_1 une valeur d'adhérence de $(T^n_\#\mu_1^n)_n$; on note encore $(T^n_\#\mu_1^n)_n$ la sous-suite de limite ν_1. On a

$$(\pi^y)_\#\left((Id, T^n)_\#(\mu_1^n + \mu_2^n)\right) = T^n_\#\mu_1^n + T^n_\#\mu_2^n. \tag{3.23}$$

D'une part, ${\pi^y}_\#\left((Id, T^n)_\#(\mu_1^n + \mu_2^n)\right)$ tend vers $\rho_1 + \rho_2$ de l'autre $(T^n_\#\mu_1^n)_n$ converge vers ν_1. Par conséquent,

$$\lim_{n\to\infty} T^n_\#\mu_2^n = \rho_1 + \rho_2 - \nu_1. \tag{3.24}$$

On a donc en utilisant deux fois le lemme 7

$$\begin{aligned}W_2^2(\mu_1, \nu_1) + W_2^2(\mu_2, \rho_1 + \rho_2 - \nu_1) &= \lim_{n\to\infty} W_2^2(\mu_1^n, T^n_\#\mu_1^n) + W_2^2(\mu_2^n, T^n_\#\mu_2^n)\\ &= \lim_{n\to\infty} W_2^2(\mu_1^n + \mu_2^n, T^n_\#(\mu_1^n + \mu_2^n))\\ &= W_2^2(\mu_1 + \mu_2, \rho_1 + \rho_2)\\ &= W_2^2(\mu_1, \rho_1) + W_2^2(\mu_2, \rho_2).\end{aligned} \tag{3.25}$$

Comme $(\nu_1, \rho_1 + \rho_2 - \nu_1)$ est admissible, par unicité de la projection sur K_2, $\nu_1 = \rho_1$. Alors $(\rho_1^n)_n$ est une suite à une seule valeur d'adhérence dans $\mathbb{W}_2(\Omega)$ qui est compact : elle tend vers ρ_1 (et de même $(\rho_2^n)_n$ converge vers ρ_2). □

Remarque 6. *Dans ce qui précédait, Ω était ouvert, et la proposition précédente requiert que Ω soit compact. La proposition précédente est encore valide si Ω est relativement compact, et de frontière de mesure nulle. En effet, dans ce cas, si on se donne (ρ_1^n, ρ_2^n) convergeant vers (ρ_1, ρ_2) dans $\mathbb{W}_2(\Omega) \times \mathbb{W}_2(\Omega)$, alors leurs extensions canoniques à $\overline{\Omega}$ convergent étroitement dans $\mathbb{P}(\overline{\Omega})$ (donc dans $\mathbb{W}_2(\overline{\Omega})$). On peut alors appliquer la proposition précédente dans $K_2(\overline{\Omega})$, qui coïncide avec K_2 (car la frontière de Ω est de mesure nulle).*

3.4 Implémentation

On présente dans cette section l'implémentation effective du schéma de splitting (3.1), grandement inspirée des méthodes développées dans le cas monotype dans [46] et [61].

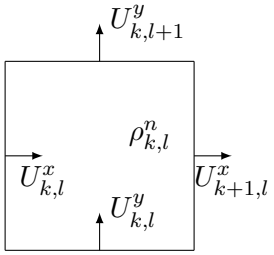


Figure 3.2: La cellule en position (k, l). Les densités sont définies au centre des cellules, alors que les vitesses sont définies sur les arêtes.

3.4.1 Prédiction

On présente ici deux méthodes pour implémenter l'étape de prédiction. On considère une seule densité, étant donné que les prédictions sont faites séparément pour chaque type.

Volumes finis

Dans [61], un schéma conservatif de type volumes finis est introduit, reposant sur un opérateur "upwind" pour discrétiser le flux. Étant donné un maillage de taille $h = \frac{1}{N}$, on commence par définir les vitesses sur les arêtes du maillage comme représenté sur la figure 3.2. L'opérateur upwind qui calcule le flux traversant une arête donnée est alors défini par

$$A^{\mathrm{up}}(U, \rho_-, \rho_+) = \begin{cases} U\rho_- \text{ si } U \geqslant 0 \\ U\rho_+ \text{ si } U < 0. \end{cases} \tag{3.26}$$

L'étape de prédiction se discrétise alors de la manière suivante :

$$\begin{aligned} \tilde{\rho}^{n+1}_{k,l} = \rho^n_{k,l} &- \frac{\tau}{h}\left(A^{\mathrm{up}}\left(U^x_{k+1,l}, \rho^n_{k,l}, \rho^n_{k+1,l}\right) - A^{\mathrm{up}}\left(U^x_{k,l}, \rho^n_{k-1,l}, \rho^n_{k,l}\right)\right) \\ &- \frac{\tau}{h}\left(A^{\mathrm{up}}\left(U^y_{k,l+1}, \rho^n_{k,l}, \rho^n_{k,l+1}\right) - A^{\mathrm{up}}\left(U^y_{k,l}, \rho^n_{k,l-1}, \rho^n_{k,l}\right)\right). \end{aligned} \tag{3.27}$$

Ce schéma est conservatif, stable sous la condition CFL

$$\frac{\tau}{h} \leqslant \frac{1}{4\|U\|_\infty}. \tag{3.28}$$

Cependant, cette condition contraint à choisir des petits pas de temps lorsqu'on raffine le maillage (augmentant ainsi le temps de calcul), et rend le schéma diffusif, comme on peut le voir sur l'exemple suivant en dimension 1.

Exemple 2. *Soit $\Omega = [0, 1]$, et considérons une seule densité $\rho = \epsilon \mathbf{1}_{\left[\frac{1}{4}, \frac{3}{4}\right]}$, sujette à une vitesse $U = \frac{5}{2} - 2x$ (de telle manière à ce que $U\left(\frac{1}{4}\right) = 2$ et $U\left(\frac{3}{4}\right) = 1$). On peut prendre*

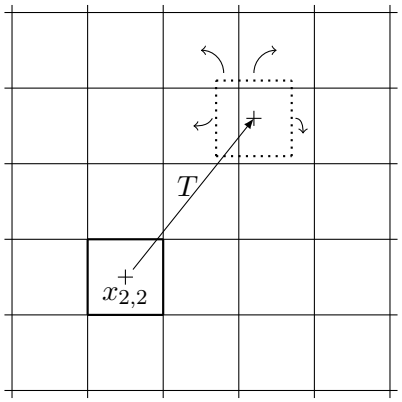


Figure 3.3: La répartition de la cellule $(2, 2)$ transportée par T. On calcule dans un premier temps l'image du centre, puis on dessine une case de taille $h \times h$. La partie inférieure gauche est distribuée sur la cellule en position $(3, 4)$, la partie supérieure gauche sur $(3, 5)$, la partie supérieure droite sur $(4, 5)$ et la partie inférieure droite sur $(4, 4)$.

un pas de temps maximal $\tau = \dfrac{h}{2}$ *pour préserver la positivité. Au* n*-ième pas, les* n *cellules à droite de* $x = \dfrac{3}{4}$ *supportent de la masse. Il y a donc de la masse qui se déplace à vitesse* $\dfrac{\delta_x}{\delta_t} = \dfrac{nh}{n\tau} = 2$*. Le schéma présente donc une diffusion plus rapide que le modèle, qui prédit une vitesse égale à* 1 *pour l'avant de la densité.*

Transport lagrangien

On présente ici une version lagrangienne de l'étape de prédiction, initialement introduite dans [46], qui a l'avantage de pouvoir utiliser des grands pas de temps. On transporte dans un premier temps la cellule $C_{k,l} = [kh, (k+1)h] \times [lh, (l+1)h]$ en calculant l'image de son centre par $T(x) := x + \tau U(x)$. On distribue ensuite la cellule transportée

$$B_{\|\cdot\|_\infty}\left(T(x), \frac{h}{2}\right) \tag{3.29}$$

sur les cellules voisines du maillage, comme illustré sur la figure 3.3.

Remarque 7. *Ce schéma est conservatif, stable et préserve la positivité pour n'importe quel pas de temps. L'avant de la densité étudiée dans l'exemple 2 avance avec ce schéma à une vitesse* $1 + \dfrac{dh}{\tau}$*, qui peut donc être réduite en choisissant une plus petite taille de maillage ou un pas de temps plus grand.*

3.4.2 Correction

Rappelons dans un premier temps comment la projection sur K_1 est obtenue dans le cas monotype dans [46]. On utilisera ensuite le corollaire 1 pour dériver un algorithme de calcul de la projection sur K_2.
Étant donné une densité μ discrétisée sur un maillage fixé, la projection sur K_1 est approchée par l'algorithme stochastique suivant :

- On part d'une case sur-saturée tirée uniformément, et on transporte l'excédent de masse par une marche aléatoire symétrique,
- Quand une case non saturée est atteinte, on délivre autant de masse que possible,
- Lorsque tout l'excédent de masse a été distribué, on recommence le processus en partant d'une autre case sur-saturée tirée aléatoirement.

On renvoie à [46] pour les heuristiques dont découle ce schéma et quelques résultats partiels de convergence. En utilisant le corollaire 1, on définit le schéma suivant pour l'étape de correction pour deux types.

Étape de correction pour deux types

- On projette la somme des densités $\mu_1 + \mu_2$ sur K_1 avec l'algorithme stochastique :
$$\rho = P_{K_1}(\mu_1 + \mu_2),$$
- On calcule à l'aide de l'algorithme de Sinkhorn le plan de transport régularisé γ^ϵ de $\mu_1 + \mu_2$ à ρ,
- On pose $f = \dfrac{\mathrm{d}\mu_1}{\mathrm{d}(\mu_1 + \mu_2)}$ et $g = \dfrac{\mathrm{d}\mu_2}{\mathrm{d}(\mu_1 + \mu_2)}$,
- En posant $\gamma_1^\epsilon(x, y) = f(x)\gamma^\epsilon(x, y)$ et $\gamma_2^\epsilon(x, y) = g(x)\gamma^\epsilon(x, y)$, la projection sur K_2 est donnée par
$$(\pi^y_\# \gamma_1^\epsilon, \pi^y_\# \gamma_2^\epsilon) \tag{3.30}$$
où π^y est la projection sur la coordonnée y.

Remarque 8. *Calculons* (3.30) *quand* $\gamma_T = (Id, T)_{\#}\mu$ *provient d'un plan de transport au sens de Monge et* $\mu = \mu_1 + \mu_2$*. Dans ce cas, on obtient*

$$\pi^y_{\#}\left(\frac{\mathrm{d}\mu_1}{\mathrm{d}\mu}(x)d\gamma_T(x,y)\right) = T_{\#}\mu_1. \tag{3.31}$$

On retrouve bien la projection exacte $T_{\#}\mu_1$ *dans le cas où on remplace l'étape Sinkhorn par un calcul exact de* T*. Dans le cas du transport régularisé, le plan de transport ne provient génériquement pas d'une application de transport. Il est néanmoins connu (voir [19]) que le plan de transport régularisé* γ^ϵ *converge vers le plan exact* γ_T *lorsque le paramètre de régularisation* ϵ *tend vers* 0*. On a alors que* γ^ϵ_j *tend vers* $\dfrac{\mathrm{d}\mu_j}{\mathrm{d}\mu}(x)d\gamma_T(x,y)$*, et donc que la projection approchée* (3.30) *converge vers la projection exacte.*

3.4.3 Vitesses considérées

Le schéma de splitting permet d'inclure dans le modèle des vitesses plus génériques que celles compatibles avec le schéma JKO étudié dans le chapitre 2. Dans l'implémentation que nous présentons, nous incluons la somme de différents termes :

- Un champ de vitesse extérieur U_i,
- L'attraction par un potentiel extérieur : $-\nabla\phi_i$,
- L'attraction vers un chemoattractant émis par les particules : ∇c où la concentration c vérifie une équation de Keller-Segel avec diffusion instantanée

$$-\kappa\Delta c = \rho_1 + \rho_2, \tag{3.32}$$

- Attraction/répulsion entre les particules : $\nabla(V * (\rho_1 + \rho_2))$ où $V(x,y)$ est un potentiel d'interaction.

Remarque 9. *Si on ne fait pas l'hypothèse d'une grande vitesse de diffusion du chemoattractant, il faut intégrer en tout temps l'évolution de cette concentration, c'est-à-dire faire un pas dans une équation de Keller-Segel :*

$$\left\{\begin{array}{lcl} \partial\rho + \nabla\cdot(\rho\nabla c) & = & 0 \\ \partial_t c - \kappa\Delta c & = & c_e\rho, \end{array}\right. \tag{3.33}$$

où κ *est la constante de diffusion de* c *et* c_e *le taux d'émission du chemoattractant.*

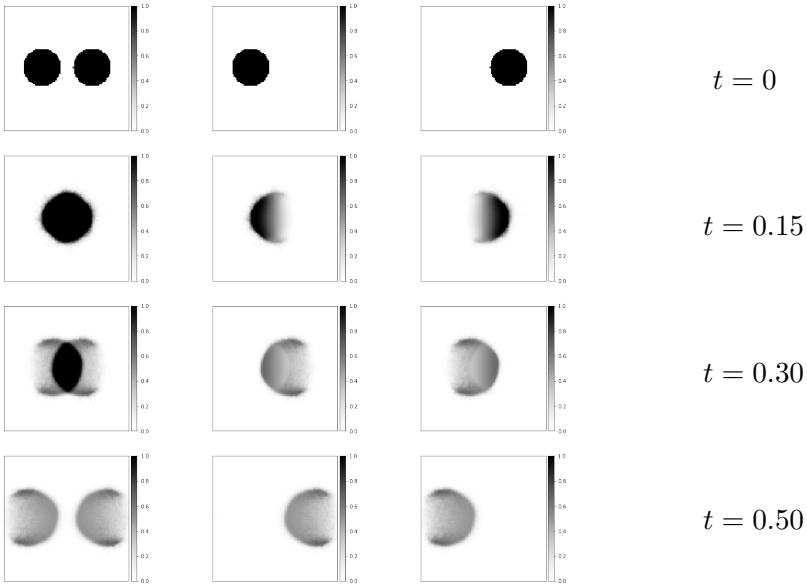


Figure 3.4: Le mouvement de deux disques se croisant. La première colonne représente la somme des deux densités, les deux autres les types séparés.

3.4.4 Simulations

L'implémentation de l'algorithme présenté dans ce qui précède est disponible[1] avec une démonstration prête à l'emploi. Le cas du croisement de disques saturés est représenté sur la figure 3.4. On voit que la foule trouve un compromis pour éviter la saturation en s'étalant sur une zone plus large lors du croisement. Les vitesses souhaitées sont ici constantes :

$$\begin{aligned} u_1 &= e_x \\ u_2 &= -e_x. \end{aligned} \tag{3.34}$$

Sur la figure 3.5, un effet de chemoattraction est introduit. Après s'être étalés sur des zones plus grandes pour se croiser, les deux types se réagrègent séparément en des foules compactes saturées. Les vitesses souhaitées sont ici choisies de la forme

$$\begin{aligned} u_1 &= e_x + \nabla c_1 \\ u_2 &= -e_x + \nabla c_2 \\ -\kappa \Delta c_1 &= \rho_1 \\ -\kappa \Delta c_2 &= \rho_2. \end{aligned} \tag{3.35}$$

[1] https://gitlab.math.u-psud.fr/bourdin/macroscopic-cell-motion

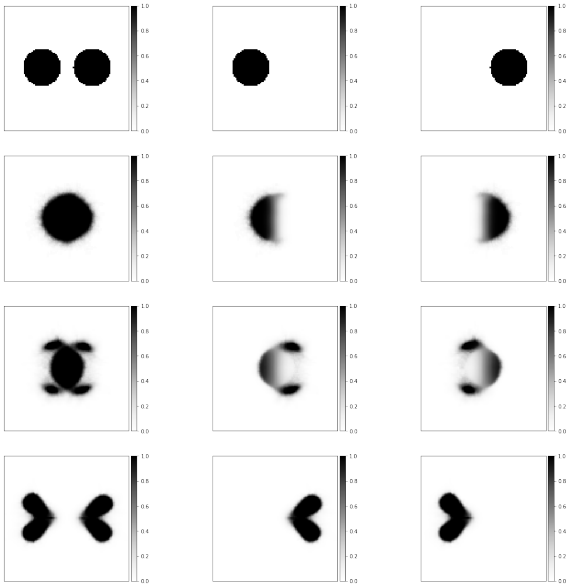


Figure 3.5: Le croisement de deux disques avec de la chemoattraction.

Un troisième exemple d'un mouvement plus complexe est illustré sur la figure 3.6. En absence de champ extérieur ou de potentiel, la foule est ici mue par de la chemoattraction et de l'attraction à courte portée avec son propre type. Le potentiel d'interaction a été choisi de la forme

$$V(x,y) = \mathbf{1}_{\|x-y\|<R}\left(1 - \frac{\|x-y\|^2}{R^2}\right)^3. \tag{3.36}$$

On a ajouté de la chemorépulsion pour l'autre type. La foule cherche alors un compromis pour se séparer en deux phases distinctes.
Les vitesses souhaitées s'écrivent ici

$$\begin{aligned} u_1 &= e_x + \nabla c_1 - \alpha \nabla c_2 + \nabla(V * \rho_1) \\ u_2 &= -e_x + \nabla c_2 - \alpha \nabla c_2 + \nabla(V * \rho_2) \\ -\kappa \Delta c_1 &= \rho_1 \\ -\kappa \Delta c_2 &= \rho_2. \end{aligned} \tag{3.37}$$

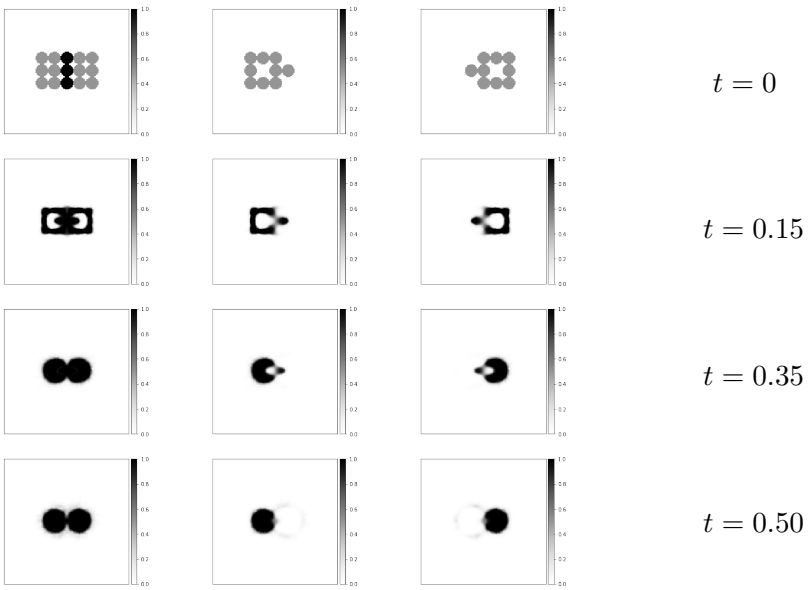


Figure 3.6: L'agrégation d'une foule composite en présence de chemoattraction et d'attraction courte distance pour son type.

Chapitre 4

Découplage

On présente dans ce chapitre un schéma de type volumes finis qui prédit des densités admissibles pour tout pas de temps sans étape de projection. Pour ce faire, on projette à chaque pas de temps les vitesses souhaitées sur l'ensemble des vitesses qui mènent à une paire de densités admissible par application d'un pas de volumes finis. Afin de contourner la non linéarité des flux traversant les arêtes en fonction des vitesses et de se ramener à un problème d'optimisation quadratique classique, on découple la variable vitesse à chaque arête en une vitesse positive et une vitesse négative. On montre la convergence du problème instantané dans le cas saturé à un type en dimension 1, puis on identifie les difficultés de l'extension de ce résultat à la dimension 2, ou au cas unilatéral en dimension 1. On analyse le cas du problème unilatéral en dimension 1 lorsque la zone saturée est un ouvert pour se ramener au cas saturé et donc utiliser le résultat d'homogénéisation. On présente enfin l'implémentation de ce schéma qui a l'avantage de produire pour tout pas de temps et d'espace des densités admissibles sans condition CFL.

4.1 Présentation du schéma

Étant donné un maillage carré de taille N de l'espace $\Omega = [0,1]$ ou $\Omega = [0,1]^2$, on se donne dans un premier temps une discrétisation des champs de vitesses souhaitées $(U_j)_j$ aux arêtes - en considérant par exemple une moyenne de U sur un voisinage de l'arête j. On se donne ensuite un opérateur $\mathrm{VF}_v^\tau(\mu)$ qui à une densité μ, un pas de temps τ et un champ de vitesse v défini sur les arêtes associe la densité obtenue par application d'un schéma de volumes finis pendant un pas de temps τ. Étant donné un ensemble de contraintes discrètes K^N portant sur les densités, on cherche à projeter au n-ième pas les vitesses souhaitées sur

$$\left\{(v_1, v_2),\ \text{vitesses t.q. } \left(\mathrm{VF}_{v_1}^\tau(\rho_1^n), \mathrm{VF}_{v_2}^\tau(\rho_2^n)\right) \in K^N\right\}. \tag{4.1}$$

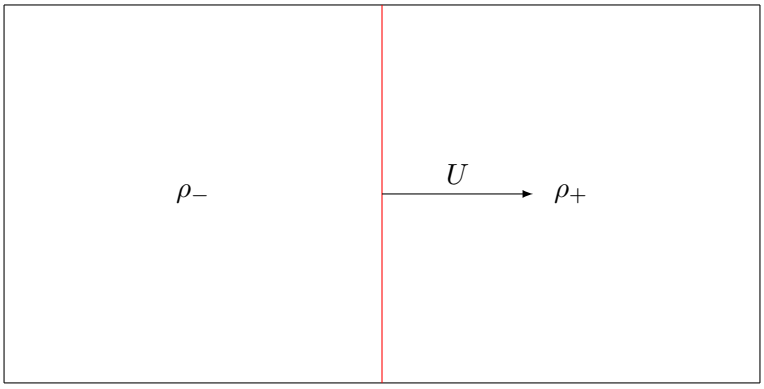


Figure 4.1: Le flux entre deux cases voisines. Lorsque $U > 0$, c'est la densité de gauche qui traverse l'arête, lorsque $U < 0$ c'est celle de droite.

Exemple 3. *En dimension* 1 *avec* $\Omega = [0,1]$ *et une seule densité, on peut écrire le schéma de volumes finis avec flux upwind comme en section 3, avec* $h = \dfrac{1}{N}$ *:*

$$(\mathrm{VF}_U^\tau(\rho^n))_i = \rho_i^n - \frac{\tau}{h}\left(A^{\mathrm{up}}\left(U_{i+1}, \rho_i^n, \rho_{i+1}^n\right) - A^{\mathrm{up}}\left(U_i, \rho_{i-1}^n, \rho_i^n\right)\right) \tag{4.2}$$

où

$$A^{\mathrm{up}}(U, \rho_-, \rho_+) = \begin{cases} U\rho_- \text{ si } U \geqslant 0 \\ U\rho_+ \text{ si } U < 0. \end{cases} \tag{4.3}$$

L'opérateur A^{up} *calcule le flux traversant une arête qui sépare deux cases portant les densités* ρ_- *et* ρ^+ *: si* $U > 0$*, une partie de la densité* ρ_- *traverse l'arête, et inversement si* $U > 0$*. Le comportement de l'opérateur* A^{up} *est illustré sur la figure 4.1.*

On cherche à se ramener à un problème d'optimisation quadratique sous contraintes affines. Afin de contourner la non-linéarité induite par le terme de flux "upwind", on découple la vitesse à chaque arête, comme sur la figure 4.3. Pour chaque arête séparant ρ^- (la case du bas si c'est une arête horizontale, de gauche si c'est une arête verticale) de ρ^+ (la case du haut ou de droite), on définit deux vitesses v^- et v^+, qui vont transporter les masses ρ^+ et ρ^- respectivement, comme représenté sur la figure 4.2. Le flux traversant l'arête est alors

$$B^{\mathrm{up}}(v^-, v^+, \rho_-, \rho_+) = -v^-\rho_+ + v^+\rho_-. \tag{4.4}$$

B^{up} est à présent un opérateur linéaire en (v^-, v^+). Dans ce qui suit, VF_v^τ est un opérateur de volumes finis avec flux découplés, c'est-à-dire que v est de la forme (v^-, v^+) et que le flux à travers les arêtes est calculé à l'aide de l'opérateur B^{up}.

On considère donc génériquement le problème instantané suivant :

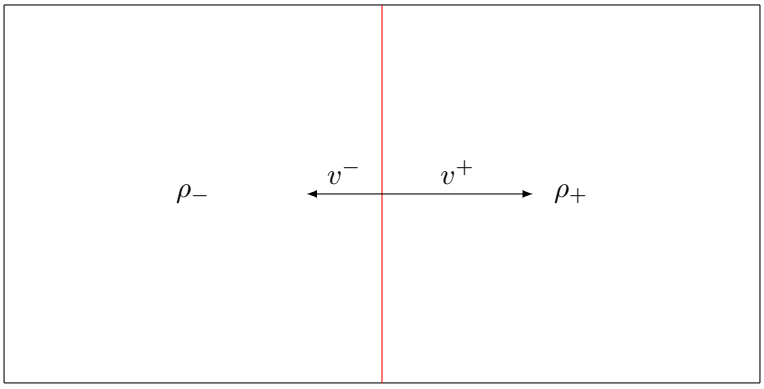


Figure 4.2: La linéarisation de l'opérateur flux par découplage des vitesses : ρ_- est transportée par v^+ et ρ_+ est transportée par v^-.

Problème 3. *Étant données ρ_1, ρ_2 deux densités définies sur un maillage d'ordre N, et des vitesses souhaitées U_1^N, U_2^N définies sur les arêtes, on cherche*

$$(u_1^-, u_1^+, u_2^-, u_2^+) \in \underset{C^N_{\rho_1,\rho_2}}{\operatorname{argmin}} \sum_{\text{arêtes}} \sum_{j=1}^{2} \rho_j^- |u_j^+ - U_j^N|^2 + \rho_j^+ |u_j^- + U_j^N|^2 \tag{4.5}$$

où $C^N_{\rho_1,\rho_2}$ est l'ensemble des vitesses dont l'application du schéma de volumes finis envoie les densités sur une paire de densités admissible.

Remarque 10. *Le problème 3 est à présent un problème d'optimisation quadratique sous contraintes d'inégalité affines. L'*argmin *est donc bien défini et son calcul effectif peu coûteux.*

Exemple 1. *On donne un exemple de contrainte pour $(u_1^-, u_1^+, u_2^-, u_2^+) \in C^N_{\rho_1,\rho_2}$ dans le cas d'une contrainte d'égalité $\rho_1 + \rho_2 = 1$, en dimension 1 :*

$$\rho_{1,i-1} u_{1,i}^+ + \rho_{2,i-1} u_{2,i}^+ + \rho_{1,i+1} u_{1,i}^- + \rho_{2,i+1} u_{2,i}^- = \rho_{1,i} \left(u_{1,i}^- + u_{2,i}^- + u_{1,i}^+ + u_{2,i}^+ \right), \forall 1 \leqslant i \leqslant N \tag{4.6}$$

c'est-à-dire que les flux entrant (terme de gauche) et sortant (terme de droite) de chaque cellule doivent coïncider.

4.2 Analyse du schéma dans le cas saturé en dimension 1

On commence l'analyse du schéma dans le cas 1D à une espèce, avec $\Omega = [0, 1]$, sous contrainte de saturation totale $\rho = 1$. Supposons donné un champ $U \in \mathbb{L}^2(\Omega)$. On découpe

Ω en N mailles $M_i = \left[\frac{i}{N}, \frac{i+1}{N}\right]$ pour $i = 0, \dots, N-1$.
À chaque arête, on associe le flux discrétisé

$$\bar{U}_i = \frac{1}{|L_i|} \int_{L_i} U(x)\,\mathrm{d}x, \tag{4.7}$$

où $L_i = M_i - \frac{1}{2N}$, avec pour convention $L_0 = \left[0, \frac{1}{2N}\right]$ et $L_N = \left[1 - \frac{1}{2N}, 1\right]$. L'ensemble de contraintes s'écrit alors :

$$\begin{aligned} C_\rho^N = \{(a_i, b_i)_{i=0,\dots,N} \in \mathbb{R}_+^{2N+2},\, \forall i = 1, \dots, N-1,\, a_i + b_{i+1} \geqslant a_{i+1} + b_i \\ a_0 + b_1 \geqslant a_1,\, a_{N-1} + b_N \geqslant b_{N-1}\}. \end{aligned} \tag{4.8}$$

En effet, si $(u_i^-, u_i^+) \in C_\rho^N$, le flux sortant de M_i sera $u_i^- + u_{i+1}^+$, supérieur au flux entrant $u_i^+ + u_{i+1}^-$. On suppose implicitement dans la formulation suivante qu'il n'y a pas de masse à gauche de 0 ou à droite de 1 ; le problème 3 devient ici

$$(a, b) \in \underset{(c,d) \in C_\rho^N}{\operatorname{argmin}} J_1(c, d) \tag{4.9}$$

où

$$J_1 : \begin{cases} \mathbb{R}^{2N+2} & \longrightarrow & \mathbb{R} \\ (a, b) & \longmapsto & \displaystyle\sum_{i=0}^{N} |a_i + \bar{U}_i|^2 + |b_i - \bar{U}_i|^2 \end{cases} \tag{4.10}$$

Comme il n'y a pas de contrainte sur b_0 et sur a_N, on a automatiquement $b_0 = \max(\bar{U}_0, 0)$ et $a_N = -\min(\bar{U}_N, 0)$. Le théorème suivant énonce que si le champ de vitesses est $\mathbb{H}^1$, les vitesses convergent vers la solution du problème continu quand $N \to +\infty$:

Théorème 2. *Soit $U \in \mathbb{H}^1(\Omega)$. Il existe une unique solution $(a_i, b_i)_{i=0,\dots,N}$ au problème 3 avec C_ρ^N de la forme (4.8). En notant v_N la fonction constante par morceaux valant $b_i - a_i$ sur L_i, alors v_N converge dans $\mathbb{L}^2(\Omega)$ vers la projection de U sur*

$$C_\rho = \left\{u \in \mathbb{L}^2(\Omega),\, \nabla \cdot u \geqslant 0\right\}. \tag{4.11}$$

De plus, on a

$$\|v_N - u\|_2^2 \leqslant \frac{c}{N}, \tag{4.12}$$

avec $c > 0$ une constante ne dépendant que de u.

Démonstration. On minimise une fonctionnelle quadratique sous contraintes linéaires d'égalité : il existe donc une unique solution (a, b) au problème 3. Notons u la projection de U sur C_ρ. Soit $V = \mathbb{L}_+^2(\Omega)^2$, et l'espace de contraintes macroscopiques

$$C_\rho^{\mathrm{dec}} = \{(v, w) \in V,\, \nabla \cdot (w - v) \geqslant 0\}. \tag{4.13}$$

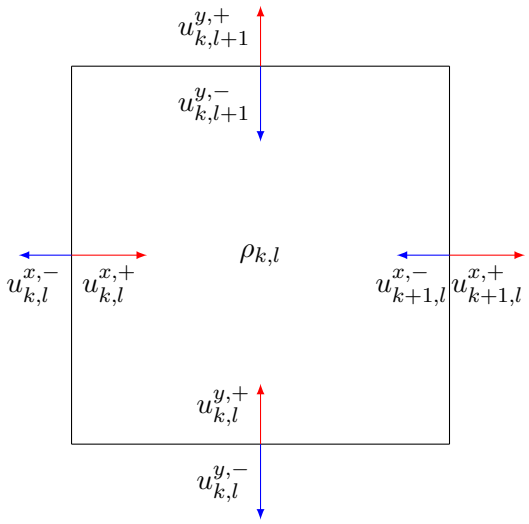


Figure 4.3: Une maille. La densité $\rho^{j,n}_{k,l}$ est définie à l'intérieur de la maille, tandis que les vitesses sont définies sur les arêtes.

En notant u_- et u_+ les parties négative et positive de u, (u_-, u_+) minimise la fonctionnelle

$$J_2 : \left\{ \begin{array}{ccc} L^2(\Omega) \times L^2(\Omega) & \longrightarrow & \mathbb{R} \\ (v, w) & \longmapsto & \|v\|_2^2 + \|w\|_2^2 - 2\langle w - v, U\rangle \end{array} \right. \tag{4.14}$$

sur C_ρ^{dec}. En effet, on a

$$J_2(v, w) = \|w - v - U\|_2^2 + 2\langle v, w\rangle - \|U\|_2^2. \tag{4.15}$$

Si on se donne $(v, w) \in C_\rho^{\text{dec}}$ tels que $J_2(v, w) < J_2(u_-, u_+)$, comme v et w sont positives, on obtient

$$\|w - v - U\|_2^2 < \|u - U\|_2^2, \tag{4.16}$$

ce qui est absurde car $w - v \in C_\rho$ et u est la projection de U sur C_ρ. En identifiant C_ρ^N avec l'ensemble des paires de fonctions constantes sur chaque L_i, on note que (a, b) minimise J_2 sur C_ρ^N : en effet, on a

$$J_2(a, b) = \frac{1}{N}\left(\sum_{i=0}^{N} |a_i + \bar{U}_i|^2 + |b_i - \bar{U}_i|^2\right) - \frac{2}{N}\sum_{i=0}^{N} \bar{U}_i^2. \tag{4.17}$$

On peut appliquer le théorème 1 de [29], pour obtenir l'estimation

$$\|v_N - u\|_2^2 \leqslant c\left(d\left(u, C_\rho^N\right)^2 + \|U\|_2\, d(u, C_\rho^N)\right) \tag{4.18}$$

où l'on a noté $d(u, C_\rho^N) = d\left((u_+, u_-), C_\rho^N\right)$ et $c > 0$ est une constante.
Posons alors v la fonction constante par morceaux valant les moyennes de u sur les L_i, puis v_- et v_+ ses parties négative et positive. Comme (u_-, u_+) est dans C_ρ^{dec}, (v_-, v_+) est dans C_ρ^N. On doit alors estimer

$$\sum_{i=0}^{N} \int_{L_i} |(u_+ - v_+)(x)|^2 \,\mathrm{d}x + \int_{L_i} |(u_- - v_-)(x)|^2 \,\mathrm{d}x \tag{4.19}$$

Pour conclure, on utilise les résultats suivants, dont les démonstrations figurent en annexe :

Théorème 3. *Soit $\Omega = [0,1]$. Si $U \in \mathbb{H}^1(\Omega)$, la projection en norme $\|\cdot\|_2$ de U sur $C_\rho = \left\{u \in \mathbb{L}^2(\Omega),\, \nabla \cdot u \geqslant 0\right\}$ est dans $\mathbb{H}^1(\Omega)$.*

Lemme 10. *Soit $u \in \mathbb{H}^1(\Omega)$, $1 \leqslant i \leqslant N-2$. Alors il existe c une constante ne dépendant que de u telle que*

$$\int_{L_i} \int_{L_i} |u(s) - u(t)|^2 \,\mathrm{d}s \,\mathrm{d}t \leqslant \frac{c}{N^3} \tag{4.20}$$

Si de plus $\lambda\left(L_i \cap \{u < 0\}\right) > 0$, alors

$$\int_{L_i} u_+^2(s) \,\mathrm{d}s \leqslant \frac{c}{N^2}. \tag{4.21}$$

Notons, pour $i = 1, \ldots, N-2$, $I_i = \displaystyle\int_{L_i} |(u_+ - v_+)(x)|^2 \,\mathrm{d}x$. Si $v \leqslant 0$ sur L_i, on a

$$I_i = \int_{L_i} |u_+(x)|^2 \,\mathrm{d}x, \tag{4.22}$$

et on peut directement appliquer le lemme 10 pour obtenir $I_i \leqslant \dfrac{c}{N^2}$. Si $v > 0$, on estime I_i de la façon suivante :

$$\begin{aligned} I_i &= \int_{L_i \cap \{u>0\}} |(u-v)(x)|^2 \,\mathrm{d}x + \int_{L_i \cap \{u<0\}} |v(x)|^2 \,\mathrm{d}x \\ &= \int_{L_i \cap \{u>0\}} \left(N \int_{L_i} u(x) - u(y) \,\mathrm{d}t\right)^2 \mathrm{d}x + \lambda(L_i \cap \{u<0\}) v_{|L_i}^2. \end{aligned} \tag{4.23}$$

On majore la première intégrale par $\dfrac{c}{N^2}$ à l'aide de l'inégalité de Jensen puis à l'aide du premier point du lemme (10). Pour la deuxième intégrale, ou bien $\lambda(J_i \cap \{u < 0\}) = 0$, ou bien on utilise l'inégalité de Jensen pour obtenir

$$\lambda(L_i \cap \{u<0\}) v_{|L_i}^2 \leqslant \left(\int_{L_i} u_+^2 + u_-^2 \right) \lambda(L_i \cap \{u<0\}) \tag{4.24}$$

et appliquer le deuxième point du lemme 10 pour majorer par un ordre $\frac{1}{N^3}$. On traite similairement le deuxième terme de (4.19), et on obtient $I_i \leqslant \frac{c}{N^2}$, soit une convergence du schéma à vitesse $\frac{1}{\sqrt{N}}$ (on majore brutalement I_0 et I_N par un terme d'ordre $\frac{1}{N}$). □

Finissons à présent l'analyse du cas 1D avec deux propositions éclairantes sur le comportement de la paire de flux optimale (a,b) donnée par (4.10). La proposition suivante affirme que soit le flux traversant une arête donnée par la gauche soit le flux la traversant par la droite est nul.

Proposition 2. *Soit (a,b) définie par (4.10). Alors, pour tout $i=0,\dots,N$, $a_i b_i = 0$.*

Démonstration. Définissons $c = a - a \wedge b$ et $d = b - a \wedge b$. On a $b-a = d-c$ donc la paire (c,d) est admissible dans C_ρ^N. Calculons maintenant

$$\begin{aligned}
J_1(c,d) &= \sum_{i=0}^N |c_i + \bar{U}_i|^2 + |d_i - \bar{U}_i|^2 - (\bar{U}_0)^2 - (\bar{U}_N)^2 \\
&= \sum_{i=0}^N |a_i + \bar{U}_i - a_i \wedge b_i|^2 + |b_i - \bar{U}_i - a_i \wedge b_i|^2 - (\bar{U}_0)^2 - (\bar{U}_N)^2 \\
&= \sum_{i=0}^N |a_i + \bar{U}_i|^2 + |b_i - \bar{U}_i|^2 + 2a_i \wedge b_i \left(a_i \wedge b_i - a_i - b_i\right) - (\bar{U}_0)^2 - (\bar{U}_N)^2 \\
&= J_1(a,b) + \sum_{i=0}^N 2a_i \wedge b_i \left(a_i \wedge b_i - a_i - b_i\right).
\end{aligned} \tag{4.25}$$

Comme le terme dans la somme est négatif, par optimalité de (a,b) pour J_1 sur C_ρ^N on obtient $a_i \wedge b_i = 0$ pour tout $i = 0,\dots,N$. □

Montrons maintenant une estimée sur (a,b) quand la vitesse souhaitée U est bornée en norme $\|\cdot\|_\infty$.

Proposition 3. *Supposons $U \in \mathbb{L}^\infty(\Omega)$. Alors il existe une constante $c(N)$ telle que*

$$\max_{i=0,\dots,N} \max(a_i, b_i) \leqslant c(N) \|U\|_\infty. \tag{4.26}$$

Démonstration. Par la formulation (4.10), on voit que $(a_0, \dots, a_N, b_0, \dots, b_N)$ est la projection de $(-\bar{U}_0, \dots, -\bar{U}_N, \bar{U}_0, \dots, \bar{U}_N)$ sur C_ρ^N pour la norme 2 discrète. Par équivalence des normes en dimension finie, soit $c_1, c_2 > 0$ telles que pour tout $x \in \mathbb{R}^{2N+2}$

$$\begin{aligned} \|x\|_\infty &\leqslant c_1 \|x\|_2 \\ \|x\|_2 &\leqslant c_2 \|x\|_\infty. \end{aligned} \tag{4.27}$$

Comme C_ρ^N est un convexe fermé qui contient 0, on a que la projection $P_{C_\rho^N}$ satisfait $\|P_{C_\rho^N}(x)\|_2 \leqslant \|x\|_2$. On a alors

$$\begin{aligned} \|(a,b)\|_\infty &\leqslant c_1 \|(a,b)\|_2 \\ &\leqslant c_1 \|(-\bar{U}_0, \dots, -\bar{U}_N, \bar{U}_0, \dots, \bar{U}_N)\|_2 \\ &\leqslant c_1 c_2 \|(-\bar{U}_0, \dots, -\bar{U}_N, \bar{U}_0, \dots, \bar{U}_N)\|_\infty. \end{aligned} \tag{4.28}$$

En posant $c(N) = c_1 c_2$, on a l'estimée voulue puisque les $(\bar{U}_i)_{i=0,\dots,N}$ sont des moyennes locales de U. □

Remarque 11. *Dans ce qui précède, on n'a pas inclus la contrainte $\rho > 0$ dans la définition de C_ρ^N. La proposition précédente justifie le fait que pour U et N donnés, il est possible de choisir τ assez petit pour que le flux sortant de chaque cellule soit plus petit que 1. Dans le cas saturé, c'est suffisant pour assurer que ρ reste positive partout.*

4.3 Analyse du cas saturé en dimension 2

La preuve menée en dimension 1 est plus délicate à adapter en dimension 2, avec $\Omega = [0,1]^2$. On se place dans le cas saturé $\rho = 1$ p.p. Définissons les cellules

$$\begin{aligned} M_{i,j} &= \left[\frac{i}{N}, \frac{i+1}{N}\right] \times \left[\frac{j}{N}, \frac{j+1}{N}\right] \quad i = 0, \dots, N-1,\ j = 0, \dots, N-1 \\ L_{i,j} &= \left[\frac{i-1/2}{N}, \frac{i+1/2}{N}\right] \times \left[\frac{j}{N}, \frac{j+1}{N}\right] \quad i = 0, \dots, N,\ j = 0, \dots, N-1 \\ J_{i,j} &= \left[\frac{i}{N}, \frac{i+1}{N}\right] \times \left[\frac{j-1/2}{N}, \frac{j+1/2}{N}\right] \quad i = 0, \dots, N-1,\ j = 0, \dots, N \end{aligned} \tag{4.29}$$

comme représenté sur la figure 4.4, avec des demi-cellules $L_{0,j}, L_{N,j}, J_{i,0}, J_{i,N}$ aux bords. On définit à présent l'approximation des flux de U à travers les arêtes de gauche et du bas de $M_{i,j}$ par

$$\begin{aligned} \bar{U}^x_{i,j} &= \frac{1}{|L_{i,j}|} \int_{L_{i,j}} U(x)\,\mathrm{d}x \\ \bar{U}^y_{i,j} &= \frac{1}{|J_{i,j}|} \int_{J_{i,j}} U(x)\,\mathrm{d}x. \end{aligned} \tag{4.30}$$

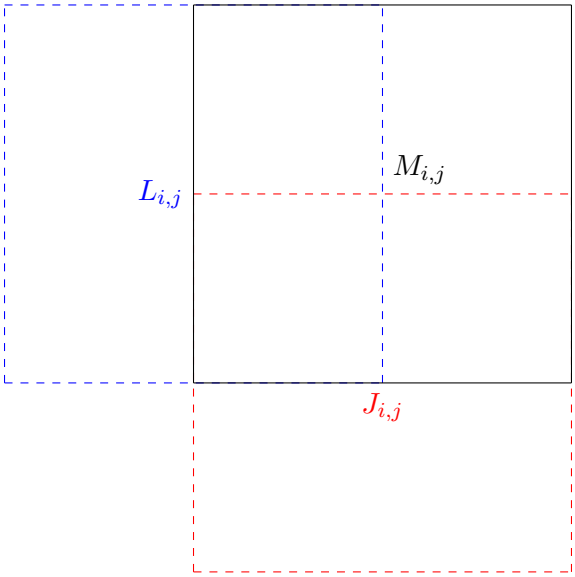


Figure 4.4: Les cellules $M_{i,j}, L_{i,j}, J_{i,j}$.

En dimension 2, le quadruplet de vitesses aux arêtes (a, b, c, d) appartient à C_ρ^N si

$$\tilde{a}_{i,j} + \tilde{b}_{i+1,j} + \tilde{c}_{i,j} + \tilde{d}_{i,j+1} \geqslant \tilde{a}_{i+1,j} + \tilde{b}_{i,j} + \tilde{c}_{i,j+1} + \tilde{d}_{i,j} \quad \forall i, j = 0, \dots, N-1 \tag{4.31}$$

avec

$$\begin{aligned}
\tilde{a}_{i,j} &= \begin{cases} a_{i,j} & \text{si } i \neq N \\ 0 & \text{sinon} \end{cases} \\
\tilde{b}_{i,j} &= \begin{cases} b_{i,j} & \text{si } i \neq 0 \\ 0 & \text{sinon} \end{cases} \\
\tilde{c}_{i,j} &= \begin{cases} c_{i,j} & \text{si } j \neq N \\ 0 & \text{sinon} \end{cases} \\
\tilde{d}_{i,j} &= \begin{cases} d_{i,j} & \text{si } j \neq 0 \\ 0 & \text{sinon.} \end{cases}
\end{aligned} \tag{4.32}$$

On suppose ainsi qu'il n'y a pas de masse en dehors de Ω. Les distinctions de cas précédentes correspondent aux flux entrants aux bords de Ω, qui sont nuls ici. Le problème 3 s'écrit maintenant

$$(a, b, c, d) \in \underset{(e,f,g,h) \in C_\rho^N}{\operatorname{argmin}} J_1(e, f, g, h), \tag{4.33}$$

avec

$$J_1 : \begin{cases} \mathbb{R}^{4N(N+1)} & \longrightarrow & \mathbb{R} \\ (a,b,c,d) & \longmapsto & \sum_{i=0}^{N}\sum_{j=0}^{N-1} |a_{i,j} + \bar{U}^x_{i,j}|^2 + |b_{i,j} - \bar{U}^x_{i,j}|^2 \\ & & \qquad + \sum_{i=0}^{N-1}\sum_{j=0}^{N} |c_{i,j} + \bar{U}^y_{i,j}|^2 + |d_{i,j} - \bar{U}^y_{i,j}|^2. \end{cases} \tag{4.34}$$

Il n'y a pas de contraintes sur les vitesses entrantes aux bords de Ω, on obtient donc pour $i, j = 0, \ldots, N$

$$\begin{aligned} a_{N,j} &= -\min(\bar{U}^x_{N,j}, 0) \\ b_{0,j} &= \max(\bar{U}^x_{0,j}, 0) \\ c_{i,N} &= -\min(\bar{U}^y_{i,N}, 0) \\ d_{i,0} &= \max(\bar{U}^y_{i,0}, 0). \end{aligned} \tag{4.35}$$

En notant

$$J_2 : \begin{cases} \mathbb{L}^2(\Omega)^4 & \longrightarrow & \mathbb{R} \\ (v,w,y,z) & \longmapsto & \|v\|_2^2 + \|w\|_2^2 + \|y\|_2^2 + \|z\|_2^2 - 2\langle \begin{pmatrix} w - v \\ z - y \end{pmatrix}, U\rangle, \end{cases} \tag{4.36}$$

on a la proposition suivante, établie dans la preuve de la proposition 2 dans le cas 1D :

Proposition 4. *(a,b,c,d) minimise J_2 sur l'injection de C_ρ^N dans $\mathbb{L}^2(\Omega)^4$.*

Démonstration. Comme dans la preuve de la proposition 2, on a la relation

$$J_2(a,b,c,d) = \frac{J_1(a,b,c,d) - 2\sum_{i,j}(\bar{U}^x_{i,j})^2 + (\bar{U}^y_{i,j})^2}{N}. \tag{4.37}$$

□

On obtient comme dans le cas monodimensionnel la proposition suivante, qui relie la projection sur l'ensemble des champs à divergence positive à la fonctionnelle J_2.

Proposition 5. *Soit u la projection $\mathbb{L}^2$ de U sur l'ensemble C_ρ des champs à divergence positive. En notant $u^x_-, u^x_+, u^y_-, u^y_+$ les parties négative et positive des composantes de u, $(u^x_-, u^x_+, u^y_-, u^y_+)$ minimise J_2 sur l'ensemble*

$$C_\rho^{\text{dec}} = \left\{ (v,w,y,z) \in \mathbb{L}^2(\Omega), \nabla \cdot \begin{pmatrix} w - v \\ z - y \end{pmatrix} \geqslant 0 \right\}. \tag{4.38}$$

Démonstration. La démonstration est similaire à la preuve de la proposition 2. □

En utilisant le théorème 1 de [29], on obtient une première estimée entre la projection de U et sa version discrète :

Proposition 6. *Soit $v_N^x = b_{i,j} - a_{i,j}$ sur $L_{i,j}$, $v_N^y = d_{i,j} - c_{i,j}$ sur $J_{i,j}$, et $v_N = (v_N^x, v_N^y)$. Il existe une constante $c > 0$ telle que pour tout $N \geqslant 1$, on a*

$$\|u - v_N\|_2^2 \leqslant c \left(d\left((u_-^x, u_+^x, u_-^y, u_+^y), C_\rho^N\right) + d(v_N, C_\rho^{\mathrm{dec}}) \right), \tag{4.39}$$

où C_ρ^N est identifiée à son injection dans $\mathbb{L}^2(\Omega)^4$.

Il est plus difficile d'obtenir une borne explicite en fonction de N que précédemment. En effet, en dimension 1, on devait simplement estimer $d((u_-, u_+), C_\rho^N)$, puisque C_ρ^N s'identifiait à un sous-ensemble de C_ρ^{dec}. De plus, pour estimer la distance de (u_-, u_+) à C_ρ^N, le candidat

$$\begin{aligned} a_{i,j} &= \left(\frac{1}{|L_{i,j}|} \int_{L_{i,j}} u^x \right)_- \\ b_{i,j} &= \left(\frac{1}{|L_{i,j}|} \int_{L_{i,j}} u^x \right)_+ \\ c_{i,j} &= \left(\frac{1}{|J_{i,j}|} \int_{J_{i,j}} u^y \right)_- , \\ d_{i,j} &= \left(\frac{1}{|J_{i,j}|} \int_{J_{i,j}} u^y \right)_+ \end{aligned} \tag{4.40}$$

ne forme plus un quadruplet admissible dans C_ρ^N. De plus, on ne sait pas si le théorème de régularité 3 est vrai en dimension 2. Néanmoins, dans le cas où u appartient à $\mathbb{H}^1(\Omega, \mathbb{R}^2)$, on peut poser

$$\begin{aligned} a_{i,j} &= \left(\frac{1}{\Gamma_l} \int_{\Gamma_l} u \cdot \mathrm{d}n \right)_- \\ b_{i,j} &= \left(\frac{1}{\Gamma_l} \int_{\Gamma_l} u \cdot \mathrm{d}n \right)_+ \\ c_{i,j} &= \left(\frac{1}{\Gamma_d} \int_{\Gamma_d} u \cdot \mathrm{d}n \right)_- \\ d_{i,j} &= \left(\frac{1}{\Gamma_d} \int_{\Gamma_d} u \cdot \mathrm{d}n \right)_+ \end{aligned} \tag{4.41}$$

où Γ_l et Γ_d sont les bords gauche et droit de $M_{i,j}$. Dans ce cas, on a

$$\begin{aligned} a_{i,j} + b_{i+1,j} + c_{i,j} + d_{i,j+1} - a_{i+1,j} - b_{i,j} - c_{i,j+1} - d_{i,j} &= \frac{1}{N} \int_{\partial M_{ij}} u \cdot \mathrm{d}n \\ &= \frac{1}{N} \int_{M_{i,j}} \nabla \cdot u \end{aligned} \tag{4.42}$$

qui est positif car u est dans C_ρ. On peut vérifier que les conditions de positivité du flux sont vérifiées pour les cas particuliers $i = 0, N$, $j = 0, N$. On a donc construit un élément $(a, b, c, d) \in C_\rho^N$.

Proposition 7. *Supposons que u soit dans $\mathbb{H}^1(\Omega, \mathbb{R}^2)$. Soit w_N^x la fonction constante par morceaux valant $b_{i,j} - a_{i,j}$ sur $L_{i,j}$, $w_N^y = d_{i,j} - c_{i,j}$ sur $J_{i,j}$ et $w_N = (w_N^x, w_N^y)$.Alors il existe une constante κ dépendant uniquement de u telle que*

$$\|u - w_N\|_2 \leqslant \frac{\kappa}{N^2} \tag{4.43}$$

La preuve utilise le résultat suivant, dont la preuve est en annexe.

Lemme 11. *Soit $\phi \in C^1([0,1]^2)$. Il existe une constante $c > 0$ telle que pour tous N, i et j*

$$\int_{M_{i,j}} \left| \phi(x,y) - N \int_{\Gamma_l} \phi(t)\, \mathrm{d}t \right|^2 \mathrm{d}x\, \mathrm{d}t \leqslant \frac{c}{N^2} \int_{M_{i,j}} |\nabla \phi|^2. \tag{4.44}$$

Démonstration de la proposition 7. Posons

$$\begin{aligned} M_{i,j}^l &= M_{i,j} \cap L_{i,j} \\ M_{i,j}^r &= M_{i,j} \cap L_{i+1,j} \\ M_{i,j}^d &= M_{i,j} \cap J_{i,j} \\ M_{i,j}^u &= M_{i,j} \cap J_{i,j+1} \end{aligned} \tag{4.45}$$

c'est-à-dire les parties gauche, droite, basse, haute d'une cellule. On peut décomposer

$$\begin{aligned} \|u - w_N\|_2^2 &= \sum_{i,j=0}^{N-1} \int_{M_{i,j}} |u(x,y) - w_N(x,y)|^2 \,\mathrm{d}x\, \mathrm{d}t \\ &= \sum_{i,j=0}^{N-1} \int_{M_{i,j}^l} |u^x(x,y) - (b_{i,j} - a_{i,j})|^2 \,\mathrm{d}x\, \mathrm{d}t \\ &\quad + \int_{M_{i,j}^r} |u^x(x,y) - (b_{i+1,j} - a_{i+1,j})|^2 \,\mathrm{d}x\, \mathrm{d}t \\ &\quad + \int_{M_{i,j}^d} |u^y(x,y) - (d_{i,j} - c_{i,j})|^2 \,\mathrm{d}x\, \mathrm{d}t \\ &\quad + \int_{M_{i,j}^u} |u^y(x,y) - (d_{i,j+1} - c_{i,j+1})|^2 \,\mathrm{d}x\, \mathrm{d}t. \end{aligned} \tag{4.46}$$

Estimons le premier terme (les trois autres sont similaires). Comme u est dans $\mathbb{H}^1(\Omega,\mathbb{R}^2)$, soit $(\phi_p)_p \in C^1([0,1]^2,\mathbb{R}^2)^{\mathbb{N}}$ qui tend vers u dans $\mathbb{H}^1(\Omega,\mathbb{R}^2)$.

$$\begin{aligned}\sum_{i,j=0}^{N-1}\int_{M_{i,j}^l}|u^x(x,y)-(b_{i,j}-a_{i,j})|^2\,\mathrm{d}x\,\mathrm{d}t &= \int_{M_{i,j}^l}|u^x(x,y)-N\int_{\Gamma_l}u^x(t)\,\mathrm{d}t|^2\,\mathrm{d}x\,\mathrm{d}t\\ &\leqslant 4\int_{M_{i,j}^l}|u^x(x,y)-\phi_p^x(x,y)|^2\,\mathrm{d}x\,\mathrm{d}t\\ &\quad+4\int_{M_{i,j}^l}|\phi_p^x(x,y)-N\int_{\Gamma_l}\phi_p^x(t)\,\mathrm{d}t|^2\,\mathrm{d}x\,\mathrm{d}t\\ &\quad+4\int_{M_{i,j}^l}|N\int_{\Gamma_l}\phi_p^x(t)\,\mathrm{d}t-N\int_{\Gamma_l}u^x(t)\,\mathrm{d}t|^2\,\mathrm{d}x\,\mathrm{d}t.\end{aligned} \tag{4.47}$$

Comme l'opérateur trace est continu de $\mathbb{H}^1(\Omega,\mathbb{R}^2)$ vers $\mathbb{H}^{\frac{1}{2}}(\Gamma_l)$, et comme $(\phi_p)_p$ tend vers u dans $\mathbb{H}^1(\Omega,\mathbb{R}^2)$, les premier et dernier termes tendent vers 0 quand p tend vers l'infini. On utilise alors le lemme 11 pour dominer le deuxième terme par $\dfrac{4c}{N^2}\displaystyle\int_{M_{i,j}^l}|\nabla\phi_p(x,y)|^2\,\mathrm{d}x\,\mathrm{d}t$. En faisant tendre p vers l'infini, on obtient l'estimée voulue :

$$\sum_{i,j=0}^{N-1}\int_{M_{i,j}^l}|u^x(x,y)-(b_{i,j}-a_{i,j})|^2\,\mathrm{d}x\,\mathrm{d}t\leqslant\frac{4c}{N^2}\|\nabla u(x,y)\|_2^2. \tag{4.48}$$

En posant $\kappa=16c$, on a l'inégalité désirée. □

Corollaire 2. *Sous l'hypothèse que u soit dans $\mathbb{H}^1(\Omega,\mathbb{R}^2)$,*

$$d(u,C_\rho^N)\leqslant\frac{\kappa}{N^2}. \tag{4.49}$$

Perspectives. Il manque deux ingrédients pour obtenir une estimée similaire à (4.12) et donc la convergence. D'une part, on n'a pas la version 2D du théorème de régularité 3, comme sa démonstration utilise des arguments spécifiques à la dimension 1. Il faut d'autre part montrer une estimée sur $d(v_N,C_\rho^{\mathrm{dec}})$. Un candidat naturel à la projection de v_N sur C_ρ^{dec} serait l'interpolation affine par morceaux

$$\begin{aligned}w^x&=(b_{i,j}-a_{i,j})+N\left(x-\frac{i}{N}\right)(b_{i+1,j}-a_{i+1,j}-b_{i,j}+a_{i,j})\\ w^y&=(d_{i,j}-c_{i,j})+N\left(y-\frac{j}{N}\right)(d_{i+1,j}-c_{i+1,j}-d_{i,j}+c_{i,j})\end{aligned} \tag{4.50}$$

sur $M_{i,j}$. En effet, on a

$$\nabla \cdot w = a_{i,j} + b_{i+1,j} + c_{i,j} + d_{i,j+1} - a_{i+1,j} - b_{i,j} - c_{i,j+1} - d_{i,j} \tag{4.51}$$

qui est positive car v_N est dans C_ρ^N. La distance $\mathbb{L}^2$ entre v_N et w vaut

$$\frac{2}{3N^2} \sum_{i,j=0}^{N-1} \left((b_{i+1,j} - a_{i+1,j} - b_{i,j} + a_{i,j})^2 + (d_{i+1,j} - c_{i+1,j} - d_{i,j} + c_{i,j})^2 \right). \tag{4.52}$$

On obtient une estimée sur $d(v_N, C_\rho^{\mathrm{dec}})$ si les coefficients (a, b, c, d) vérifient une estimée discrète $\mathbb{H}^1$, c'est-à-dire que la somme dans (4.52) reste bornée quand N tend vers l'infini. En effet, si les $(\bar{U}_{ij}^{xy})$ sont admissibles dans C_ρ^N, (4.52) devient

$$\frac{2}{3N^2} \sum_{i,j=0}^{N-1} \left(\left(\bar{U}_{i+1,j}^x - \bar{U}_{i,j}^x \right)^2 + \left(\bar{U}_{i,j+1}^y - \bar{U}_{i,j}^y \right)^2 \right), \tag{4.53}$$

que l'on peut dominer par $\frac{2}{3N^2} \|\nabla U\|_2$ si $U \in \mathbb{H}^1(\Omega, \mathbb{R}^2)$. Les deux résultats manquants pour avoir la convergence du schéma en dimension 2 sont donc de même nature : la question est de savoir si la projection de quantités ayant des comportements $\mathbb{H}^1$ continu ou discret sur les ensembles admissibles C_ρ ou C_ρ^N préserve leur régularité initiale.

4.4 Analyse du cas unilatéral en dimension 1

Analysons à présent l'extension du cas précédent au cas unilatéral en dimension 1, c'est-à-dire au cas où ρ n'est plus saturée sur tout Ω. Étant donné un pas de temps $\tau > 0$, $(a, b) \in \mathbb{R}^{2N+2}$ est admissible si pour tout $i = 0, \ldots, N-1$

$$\begin{aligned} \rho_i + \tau \left(a_{i+1}\rho_{i+1} + b_i\rho_{i-1} - a_i\rho_i - b_{i+1}\rho_i \right) &\leqslant 1, \\ \rho_i + \tau \left(a_{i+1}\rho_{i+1} + b_i\rho_{i-1} - a_i\rho_i - b_{i+1}\rho_i \right) &\geqslant 0. \end{aligned} \tag{4.54}$$

où $\rho_i = \rho(M_i)$ et $M_i = \left[\frac{i}{N}, \frac{i+1}{N}\right]$, avec la convention $\rho_{-1} = \rho_N = 0$. Conformément au formalisme utilisé pour montrer le théorème 2, on pose

$$J_1 : \begin{cases} \mathbb{R}^{2N+2} & \longrightarrow & \mathbb{R} \\ (a, b) & \longmapsto & \displaystyle\sum_{i=0}^{N} \rho_i |a_i + \bar{U}_i|^2 + \rho_{i-1} |b_i - \bar{U}_i|^2 \end{cases} \tag{4.55}$$

où $\bar{U}_i = \dfrac{1}{\rho(L_i)} \displaystyle\int_{L_i} U \, \mathrm{d}\rho$ avec $L_i = M_i - \frac{1}{2N}$ la translation à gauche de M_i, L_0 et L_N étant des demi-cellules aux bords de Ω. On pose également

$$J_2 : \begin{cases} \mathbb{L}^2_\rho(\Omega) \times \mathbb{L}^2_\rho(\Omega) & \longrightarrow & \mathbb{R} \\ (v, w) & \longmapsto & \|v\|^2_{\mathbb{L}^2_\rho(\Omega)} + \|w\|^2_{\mathbb{L}^2_\rho(\Omega)} - 2\langle w - v, U \rangle_{\mathbb{L}^2_\rho(\Omega)} \end{cases} \tag{4.56}$$

où

$$\mathbb{L}^2_\rho = \left\{ v \ : \Omega \longrightarrow \mathbb{R}, \int_\Omega v^2 \, \mathrm{d}\rho < \infty \right\} \tag{4.57}$$

est un espace de Hilbert muni de $\langle v, w \rangle_{\mathbb{L}^2_\rho(\Omega)} = \int_\Omega vw \mathrm{d}\rho$ et $\|v\|^2_{\mathbb{L}^2_\rho(\Omega)} = \langle v, v \rangle_{\mathbb{L}^2_\rho(\Omega)}$. u minimise toujours J_2 sur $C_\rho = \left\{ u \in \mathbb{L}^2(\Omega),\, \nabla \cdot u \geqslant 0 \text{ sur } \rho = 1 \right\}$. Mais en calculant $J_2(a, b)$ pour a, b constantes par morceaux sur chaque L_i, on obtient

$$J_2(a, b) = \sum_{i=0}^{N} \rho(L_i) \left(|b_i - \bar{U}_i|^2 + |a_i + \bar{U}_i|^2 \right) - \frac{2}{N} \sum_{i=0}^{N} \rho(L_i) \bar{U}_i^2. \tag{4.58}$$

Cela indique que la bonne manière d'écrire le problème 3 pourrait être

$$(a, b) \in \underset{(c,d) \in C^N_\rho}{\operatorname{argmin}} \sum_{i=0}^{N} \rho(L_i) \left(|a_i + \bar{U}_i|^2 + |b_i - \bar{U}_i|^2 \right) \tag{4.59}$$

à la place de la minimisation de J_1. Sous ce formalisme, on peut alors estimer la distance $\mathbb{L}^2_\rho$ entre u et v_N la fonction constante par morceaux valant $b_i - a_i$ sur L_i en estimant les distances $d((u_-, u_+), C^N_\rho)$ et $d(v_N, C^{\mathrm{dec}}_\rho)$ dans $\mathbb{L}^2_\rho$. L'enjeu principal est à présent de construire un candidat dans C^N_ρ étant donné $u \in C_\rho$, et réciproquement un candidat dans C^{dec}_ρ étant donné $v_N \in C^N_\rho$. Par exemple, avec une densité $\rho = 1 - \epsilon$ presque saturée partout, il n'y a pas de contrainte encodée dans $C_\rho = \mathbb{L}^2_\rho(\Omega)$, alors que C^N_ρ interdit tout champ discret dont le flux associé à τ et ρ dépasse l'espace disponible restant ϵ, c'est-à-dire qu'on doit avoir sur chaque cellule

$$\tau \left(a_{i+1} \rho_{i+1} + b_i \rho_{i-1} - a_i \rho_i - b_{i+1} \rho_i \right) \leqslant \epsilon. \tag{4.60}$$

Ceci illustre que sans hypothèse sur ρ, les espaces de contrainte C_ρ et C^N_ρ peuvent être très différents, menant à une mauvaise approximation de la projection de U sur C_ρ par les champs de vitesses discrets de C^N_ρ. Néanmoins, dans des cas pratiques, on peut espérer qu'il existe des zones bien définies où $\rho = 1$ (par exemple un ouvert), hors desquelles ρ n'approche pas 1 et où l'on puisse utiliser l'analyse du cas saturé.

4.5 Saturation sur un ouvert

Malgré le résultat négatif dans le cas général non saturé, intéressons nous à un cas particulier en dimension 1. On suppose ici que la zone saturée $\{\rho = 1\}$ est de la forme $\left] \frac{k}{N}, \frac{l}{N} \right[$ avec $0 < k < l < N$ et $\rho = 0$ ailleurs. On se donne un champ de vitesses borné $U \in \mathbb{H}^1(\Omega) \cap \mathbb{L}^\infty(\Omega)$. Le problème 3 s'écrit alors

$$(a, b) \in \underset{(c,d) \in C^N_\rho}{\operatorname{argmin}} \ |c_k + \bar{U}_k|^2 + |d_l - \bar{U}_l|^2 + \sum_{i=k+1}^{l-1} |c_i + \bar{U}_i|^2 + |d_i - \bar{U}_i| \tag{4.61}$$

où $(c,d) \in C_\rho^N$ si

$$\begin{aligned}
&0 \leqslant \tau\left(-c_{i+1} - d_i + c_i + d_{i+1}\right) \leqslant 1, \quad \forall i = k+1, \ldots, l-2 \\
&0 \leqslant \tau\left(-c_{k+1} + c_k + d_{k+1}\right) \leqslant 1, \\
&0 \leqslant \tau\left(d_{l-1} - c_{l-1} - d_l\right) \leqslant 1, \\
&\tau c_k \leqslant 1 \\
&\tau d_l \leqslant 1.
\end{aligned} \tag{4.62}$$

Les trois premières conditions expriment que le flux sortant d'une maille saturée doit être positif mais inférieur à 1 (pour ne pas faire apparaître une densité négative dans la maille). Les deux dernières conditions contraignent la masse à rester positive dans les cellules du bord. Soit $(\tilde{a}, \tilde{b})$ un minimiseur de la fonctionnelle de (4.61) sous les contraintes suivantes définissant l'ensemble $\tilde{C}_\rho^N$:

$$\begin{aligned}
&0 \leqslant -\tilde{a}_{i+1} - \tilde{b}_i + \tilde{a}_i + \tilde{b}_{i+1}, \quad \forall i = k+1, \ldots, l-2 \\
&0 \leqslant -\tilde{a}_{k+1} + \tilde{a}_k + \tilde{b}_{k+1}, \\
&0 \leqslant \tilde{b}_{l-1} - \tilde{a}_{l-1} - \tilde{b}_l.
\end{aligned} \tag{4.63}$$

En d'autres termes, on contraint les flux sortants des mailles saturées à être positifs, mais on ne les contraint plus automatiquement à être inférieurs à 1. Dans ce contexte, on peut utiliser la preuve de la proposition 3 pour obtenir une borne l^∞ sur $(\tilde{a}, \tilde{b})$ qui ne dépend que de U et N. Alors pour τ suffisamment petit, $(\tilde{a}, \tilde{b}) \in C_\rho^N$. Le problème 3 se réduit par conséquent ici à

$$(a,b) \in \underset{(c,d)\in\tilde{C}_\rho^N}{\operatorname{argmin}} \sum_{i=k+1}^{l-1} |c_i + \bar{U}_i|^2 + |d_i - \bar{U}_i| + |c_k + \bar{U}_k|^2 + |d_l - \bar{U}_l|^2. \tag{4.64}$$

Il y a une identification claire entre ce problème et celui étudié en section 4.2. En particulier, on obtient la convergence de ce problème vers la solution du problème de divergence positive sur la zone saturée grâce à la proposition 2. On peut étendre ce procédé pour traiter tous les cas où la zone saturée est un ouvert quelconque "raisonnable" (par exemple une union finie d'intervalles). De même, si le résultat de convergence était établi en dimension 2, on pourrait traiter de manière analogue le cas où la zone saturée est une union (finie) d'ouverts connexes non dégénérés - par exemple convexes, ou avec un bord lisse à rayon de courbure minoré.

4.6 Vers une analyse du schéma dynamique

Dans ce qui précède, on s'est contenté d'étudier le premier pas du schéma de volumes finis "découplés" et de montrer la convergence de la vitesse induite par celui-ci vers la

projection de la vitesse sur l'ensemble des vitesses macroscopiquement admissibles. Une analyse complète du schéma dynamique en temps semble délicate. En effet, le résultat positif de convergence du théorème 2 a été établi dans un cadre simple : en dimension 1, avec un seul type, et dans le cas saturé. La difficulté de relâcher ces hypothèses a été pointée dans les sections 4.3 et 4.4 ; illustrons néanmoins celle-ci en essayant d'estimer l'erreur de prédiction d'un pas du schéma en utilisant l'estimation donnée par le théorème 2. On se donne un pas de temps τ, un maillage de taille $N \geqslant 1$, une densité initiale $\rho^0 \in \mathbb{W}_2(\Omega)$ et un champ de vitesse U. Soit ρ^τ une solution de

$$\left\{ \begin{array}{rcl} \partial_t \rho + \nabla \cdot (\rho u) & = & 0 \\ u & = & P_{C_\rho}(U) \\ \rho(0,\cdot) & = & \rho^0, \end{array} \right. \tag{4.65}$$

prise au temps τ. On note $\mathrm{VF}_v^\tau(\rho)$ la densité constante par morceaux obtenue par application du schéma de volumes finis pendant un temps τ, où v est la projection de la discrétisation de U sur la contrainte (4.8). On cherche à estimer une distance $d(\mathrm{VF}_v^\tau(\rho), \rho^\tau)$ en fonction de N et de τ. On décompose selon

$$d\left(\mathrm{VF}_v^\tau(\rho), \rho^\tau\right) \leqslant d\left(\mathrm{VF}_v^\tau(\rho), \mathrm{VF}_u^\tau(\rho)\right) + d\left(\mathrm{VF}_u^\tau(\rho), \rho^\tau\right), \tag{4.66}$$

où u est la projection de U sur l'ensemble des vitesses à divergence positive. Dans le cas où l'on a une estimée du type $\|u - v\|_2 \leqslant \dfrac{c}{\sqrt{N}}$ comme celle donnée par le le théorème 2, le premier terme est d'ordre $\dfrac{\tau}{\sqrt{N}}$ en norme 2.
Il est plus difficile d'estimer le deuxième terme. En effet, à part dans le cas purement saturé - instable dans le problème dynamique - on peut s'attendre à ce que la vitesse effective $P_{C_{\rho(t)}}(U)$ et devienne rapidement différente de u comme illustré dans l'exemple suivant.

Exemple 4. *Soit $\epsilon > 0$ fixé, et $U(x) = 1 - x$ en dimension 1 avec $\Omega = [0,1]$. Soit $\rho^0 = 1 - \epsilon$ la densité initiale presque saturée. On peut calculer la densité exacte en $t = \tau$, dans l'asymptotique $\epsilon \ll \tau \ll 1$:*

$$\rho^\tau \approx \mathbf{1}_{\left[\frac{\tau}{2}, 1+\frac{\tau}{2}\right]}. \tag{4.67}$$

D'autre part, comme il n'y a pas de contrainte de saturation en $t = 0$, on a $u = U$. Étant donné un maillage de taille N, la densité donnée par le pas de volumes finis vaut

$$\nu(x) = (1-\tau)\mathbf{1}_{x \leqslant \frac{1}{N}} + \left(1 + \frac{\tau}{N}\right)\mathbf{1}_{x \in \left[\frac{1}{N}, 1\right]}. \tag{4.68}$$

Le coût de transport $W_2^2(\rho^\tau, \nu)$ est alors d'ordre τ (il faut répartir l'excédent de masse de 1 à $1 + \frac{\tau}{2}$).

Sans hypothèses supplémentaires, l'erreur commise en un pas de temps τ peut donc être d'ordre $\sqrt{\tau}$, trop grand pour pouvoir espérer montrer la convergence. On peut néanmoins espérer que dans des cas génériques, avec une zone saturée bien identifiée, le schéma approche raisonnablement le problème 2 de transport sous contraintes.

4.7 Implémentation

On décrit l'implémentation en python de l'algorithme de découplage présenté ci-dessus. On se place dans le cas le plus général : deux espèces, en deux dimensions, dans le cas d'une contrainte unilatérale $\rho_1 + \rho_2 \leqslant 1$. On effectue un maillage carré de l'espace $\Omega = [0,1] \times [0,1]$ en N^2 mailles. Comme dans le chapitre 3, on se donne des densités discrétisées $(\rho^{j,n}_{k,l})_{k,l} \in \mathbb{R}^{N^2}$ et des vitesses souhaitées définies sur les arêtes $(U^{j,x}_{k,l})_{k,l}$. On peut également intégrer à la vitesse souhaitée de multiples composantes : un champ extérieur, un potentiel d'interaction, une chemoattraction, etc. On résout à chaque étape le problème de projection (4.5) à l'aide de la librairie *scipy.cvxopt*.
La prise en compte d'une contrainte unilatérale donne lieu à quelques obstacles à éviter. En effet, la fonctionnelle de la formule (4.5) présente une partie quadratique de la forme $X \mapsto X^T M X$, avec M matrice diagonale dont la diagonale est formée des densités en ligne. Dès lors qu'une des deux densités s'annule, la matrice M est dégénérée et le problème de minimisation mal posé. On contourne cette difficulté en éliminant les entrées correspondantes, et en posant $u^+ = U$ ou $u^- = -U$, voire $u^\pm = 0$ (ce qui n'a pas d'importance car ces vitesses transportent une masse nulle). De même, si une densité est faible en une case, la matrice M possède une petite valeur propre, et donc un mauvais conditionnement, ce qui peut faire exploser le calcul numérique de l'optimum. On pallie donc à ce comportement en éliminant toutes les entrées de M inférieures à un seuil donné. On met ensuite à jour les densités par un pas de schéma de volumes finis découplés associé aux vitesses $u^{j,\pm}$.
L'exemple jouet du croisement de deux disques est représenté sur la figure 4.5. À l'instar du schéma développé dans le chapitre 3 (voir figure 3.4), ce schéma prédit d'abord un étalement sur une zone plus vaste pour éviter la sursaturation en vue du croisement. L'implémentation du schéma de volumes découplés est disponible sur GitLab[1].

Ce schéma présente un inconvénient majeur inhérent à sa structure de volumes finis. Analysons l'exemple d'une densité saturée supportée par un intervalle $]a, b[$ en dimension 1, sujette à une vitesse $U = c > 0$ constante. Étant donné un pas de temps dt et un nombre de cases N, à la k-ième itération du schéma les k cases à droite de b supportent de la masse, et ce quelle que soit la vitesse choisie. Le front de la densité avance donc à vitesse $\dfrac{\delta_x}{\delta_t} = \dfrac{1}{N dt}$. Deux cas se présentent alors :

- si $\dfrac{1}{N dt} > c$ le front de la densité avance plus vite que dans le modèle, le schéma est diffusif ;
- si $\dfrac{1}{N dt} < c$ le front de la densité avance moins vite que le modèle, le schéma n'est pas diffusif mais ne reproduit pas le modèle.

[1] https://gitlab.math.u-psud.fr/bourdin/decoupled-finite-volume-scheme-for-macroscopic-crowd-motion-models

On est donc obligé dans le cas général d'imposer une condition CFL de type $\frac{1}{Ndt} < \|U\|_\infty$ pour que le schéma reproduise le modèle, mais il est alors diffusif.

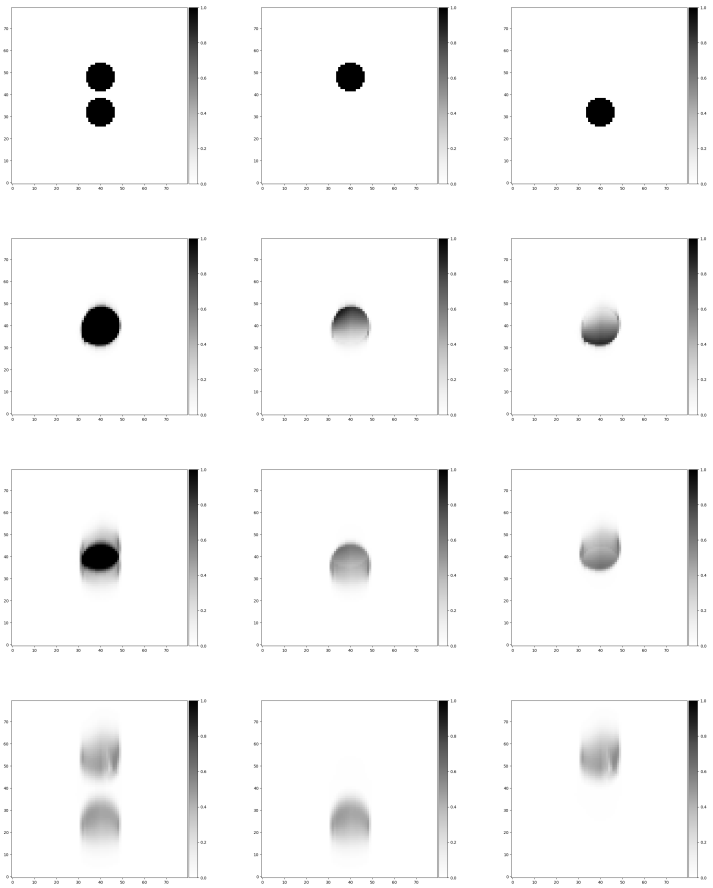


Figure 4.5: Le croisement de deux disques pour le schéma de volumes finis découplés. La première colonne représente la somme des densités, les deux autres les densités séparées.

Chapitre 5

Multibody and macroscopic impact laws: a Convex Analysis Standpoint

On étudie dans ce chapitre la relation entre la projection macroscopique du champ de vitesses souhaitées sur l'ensemble des vitesses à divergence positive et son pendant microscopique. Étant donné N sphères de rayon r, un N-uplet de vitesses $(v_1, \dots, v_N) \in \mathbb{R}^{2N}$ est admissible pour la configuration microscopique si pour toute paire de sphères (i, j) à distance critique $2r$,

$$e_{ij} \cdot (v_j - v_i) \geqslant 0, \tag{5.1}$$

où e_{ij} est le vecteur unitaire pointant depuis le centre de la sphère i vers le centre de la sphère j. On peut alors étudier dans quel formalisme et sous quelles hypothèses sur les configurations de sphères la projection macroscopique peut être identifiée comme limite de projections correspondant à des configurations microscopiques de sphères. Le résultat est a priori négatif : on montre dans cette section que la projection sur C_ρ n'est pas limite d'un processus de projection microscopique, excepté en dimension 1. En dimension 2, la projection macroscopique fait notamment intervenir comme multiplicateur de Lagrange un champ de pression qui présente un principe du maximum violé dans le cadre microscopique. De plus, la notion de saturation n'est pas clairement définie pour des collections de sphères rigides en dimension 2 : certaines configurations sont rigides (comme l'empilement sur un réseau carré) sans pour autant atteindre la saturation maximale possible $\rho_{\max} = \pi/2\sqrt{3}$. On définit néanmoins un formalisme dans lequel un résultat d'homogénéisation est énoncé dans le cas des réseaux structurés hexagonal et carré. On montre sous ces hypothèses une convergence vers un champ de vitesses qui vérifie une contrainte macroscopique anisotrope : le champ doit être croissant le long des directions de contact du réseau microscopique. Dans le cas du réseau carré, on impose au champ u à ce que $\partial_x u_x$ et $\partial_y u_y$ soient positives ; pour le réseau hexagonal on doit avoir $\partial_{e_j}(u.e_j) \geqslant 0$, où (e_0, e_1, e_2) sont les trois directions

d'empilement du réseau hexagonal compact. L'espace naturel dans lequel définir les champs de pressions sous-jacents est alors un espace de Sobolev anisotrope, où l'on demande uniquement la dérivabilité faible dans les directions de saturation. Ce chapitre, issu d'un travail réalisé conjointement à Bertrand Maury, a fait l'objet d'une publication, voir [16]. Les résultats d'homogénéisation sont établis dans le cas de réseaux très structurés et pour les problèmes instantanés. Un résultat dynamique général de convergence d'un modèle microscopique dynamique vers un modèle macroscopique développé dans les sections précédentes apparaît alors impossible. Les modèles macroscopiques de mouvement de foules ne sont donc pas la limite de modèles à échelle individuelle de description des individus, mais sont construits *ad hoc* sur des heuristiques provenant des modèles microscopiques.

Abstract These lecture notes address mathematical issues related to the modeling of impact laws for systems of rigid spheres and their macroscopic counterpart. We analyze the so-called Moreau's approach to define multibody impact laws at the mircroscopic level, and we analyze the formal macroscopic extensions of these laws, where the non-overlapping constraint is replaced by a barrier-type constraint on the local density. We detail the formal analogies between the two settings, and also their deep discrepancies, detailing how the macroscopic impact laws, natural ingredient in the so-called *Pressureless Euler Equations with a Maximal Density Constraint*, are in some way irrelevant to describe the global motion of a collection of inertial hard spheres. We propose some preliminary steps in the direction of designing macroscopic impact models more respectful of the underlying microscopic structure, in particular we establish micro-macro convergence results under strong assumptions on the microscopic structure.

5.1 Introduction

The modeling of particle systems spreads over a wide range of approaches, which rely on various levels of description of the particles. At one end of this range, the microscopic / Lagrangian setting is based on an individual description of particles, which "simply" obey Newton's Laws. At the other end, macroscopic models rely on a description of the collection of particles by a local density, and designing models amounts to elaborating equations verified by the velocity fields, under the implicit assumption that such a velocity is indeed well-defined. Between those extreme levels of descriptions, Boltzmann-like models are based on a kinetic description of the particle collection, namely a function $f(x, v, t)$ which quantifies at time t the number of particles around x at velocity v. Note that this setting makes it possible to handle a diffuse limit (smooth f representing an infinite number of infinitely small particles), as well as finite collection (a single particle at a given velocity is represented as a Dirac mass in the (x, v) space). From this standpoint, the kinetic description can be considered as microscopic in a generalized sense. This setting is particularly relevant to describe the limit of a low-density gas with the underlying

hypothesis of elastic binary collisions between particles, and it is a natural bridge between Lagrangian models, considered as untractable for many-body systems, and macroscopic models which can be used to investigate the global behavior of these systems, by means of theoretical analysis or numerical computations. Considerable energy has been, and is still, deployed to rigorously obtain macroscopic models from the Boltzmann equation, like Euler or Navier-Stokes equations (see e.g. [33, 71]).

We are interested here in *dense* collections of finite size particles (more commonly called *grains* in this context), subject to possibly non-elastic collisions. The non-dilute character of the collections together with the non-elastic character of the collision is likely to rule out the hypothesis of sole binary collisions which prevails in the Boltzmann context: multiple (or quasi-simultaneous) collisions together with persistent contacts can be expected to be generic in this situation. As a consequence, Boltzmann-like equations can no longer be considered as a natural step between microscopic and macroscopic models, and most macroscopic models which have been proposed to describe the behavior of dense (up to jammed) granular media have indeed been built independently from any homogenization procedure. We propose here to investigate the possibility to identify some ingredients that might appear in relevant macroscopic models considered as limits of microscopic ones. Let us make it clear that we are far from proposing a full and rigorous construction of a macroscopic model from the microscopic one, which is clearly out of reach. We shall rather focus on a crucial part of microscopic models, namely collision laws, and investigate the possibility to infer collision laws at the macroscopic level which would be respectful of the microscopic structure.

If one restricts to local interactions due to direct contact between entities, microscopic models based on finite size grains essentially rely on impact laws. Different strategies have been carried out to formalize this type of direct interactions. In the Molecular Dynamic approach (MD), see e.g. [2], one considers that grains are slightly deformable by implementing a short range force of the repulsive type. Note that this force is commonly taken as a computational trick to handle the non-overlapping constraint. This makes it possible to circumvent the very question of collisions, for it leads to classical Ordinary Differential Equations which fit in a classical theoretical framework (Cauchy-Lipschitz theory), and which can be solved by standard numerical schemes to perform actual computations. Note that, in spite of the natively elastic character of such interactions, some ingredients can be included to account for inelastic collision, as well as shear forces (see [60]).

In order to directly address micro-macro issues, we shall restrict ourselves here to alternative approaches, called Contact Dynamics (CD), based on a hard-sphere setting. As detailed in a recent review ([54]), several strategies can still be carried out to formalize the behavior of the system whenever the non-overlapping constraint is up to be violated. A popular, sometimes called *event-driven*, strategy, consists in handling binary collisions only. In this setting multiple collisions are considered as so rare that they can be disregarded, which makes it possible to use explicit expressions of post-collisional velocities. Another strategy is based on extending the so-called Darboux-Keller shock dynamics to multibody

collisions. It consists in changing the time scale in the neighborhood of a collision event, to set it at the impulse scale. The dynamics is then described as a sequence of compression and extension phases (see [38] for a detailed description of this method). We also refer to [31] for a very detailed account of thermodynamical aspects of collision problems.

The developments we present here are based on an alternative approach, called Moreau's approach in [54], which considers instantaneous impacts involving an arbitrary large number of grains, treated in a global way (see [63, 52, 5]). As detailed below, it relies on basic concepts of Convex Analysis, the principal of which being the cone of feasible direction associated to the set of admissible configurations (configurations with no overlapping), and the associated polar cone (set of vector which have a nonpositive scalar product with all feasible directions) which is the outward normal cone. Given a restitution coefficient $e \in [0, 1]$, the post-collisional velocity is determined from the projection of the pre-collisional velocity on the outward normal cone. Since everything can be written as a simple expression of the projection on the cone of feasible directions, we shall actually focus on this very notion in the largest part of these notes.

We shall end this introductory section by a few considerations on microscopic impact law following Moreau's approach, and what appears to be the canonical extension of this approach to the macroscopic setting. Section 5.2 is then dedicated to a detailed analysis on these impact laws. Identifying similarities and discrepancies between these formally similar laws is the object of Section 5.3. We describe in particular Laplace-like operators which are canonically associated to the collision laws in both settings, We introduce a notion of Abstract Maximum Principle (detailed in the appendix), which is verified in the macroscopic setting but not in the microscopic one, which deeply differentiates both models, and enlighten in some way the poorness of the macroscopic law.

In Section 5.4, we investigate the possibility to elaborate macroscopic impact models which are more respectful of the underlying microscopic structure. As detailed in Section 5.5, a rigorous homogenization procedure makes it possible to build such macroscopic models under very strong assumptions on the structure.

Although the resulting evolution problems are out of the scope of our work, we dedicate Section 5.6 to some remarks on this aspect of the problem. In the microscopic setting, the question is delicate but well understood: the problem is well-posed for analytic data, but might admit multiple solution otherwise, even for infinitely smooth data. In the macroscopic setting, under oversimplifying assumptions, the expected model takes the form of the so called *pressureless equations with maximal density constraint*, which essentially fits into a sound framework in the one-dimensional setting only ([12, 6]). For higher dimension, little is known on this equation. Let us add that the system is commonly written without any collision law, the actual choice being usually made in an implicit way, depending on the approach which is followed. For instance, in [12], particular solutions are built by means of *sticky blocks* with a purely inelastic collision law, whereas in [25, 26], the approach is based on compressible Euler equation with a barrier-like pressure with respect to the density,

natively leading to a purely elastic behavior.

The largest part of this text is meant to be accessible to graduate students, so we tried to preserve self-consistency as far as possible, writing at some points full proofs of elementary results, in particular in the appendix.

From single collision to multibody impact laws

We introduce here Moreau's approach of impact laws, which fits in the general class of Contact Dynamics Methods (see [52, 58]). Let us start with a point particle subject to remain in the upper half plane $\mathbb{R} \times \mathbb{R}_+$, with a purely inelastic collision law on the boundary. We denote by $r = (x, y)$ its position, and by u its velocity. If this particle is not subject to any force, its motion follows

$$u^+ = P_{C_r} u^-, \text{ with } C_r = \left| \begin{array}{ll} \mathbb{R}^2 & \text{if } y > 0 \\ \mathbb{R} \times \mathbb{R}_+ & \text{if } y = 0 \end{array} \right. \tag{5.2}$$

where u^- (resp. u^+) is the pre- (resp. post-) collisional velocity, and P_{C_r} is the euclidian projection on C_r. When the particle does not touch the wall, the velocity is constant. When a collision occurs, with pre-collisional velocity $u^- = (u_x, u_y)$ (with $u_y < 0$), the post-collisional velocity is $u^+ = (u_x, 0)$.

In the case of an elastic collision, we introduce a restitution coefficient $e \in (0, 1]$. The post-collisional velocity is now $u^+ = (u_x, -eu_y)$. This behavior can be written in a way which can generalized to the multi-collisional situation. We introduce the outward normal cone to K, defined as

$$N_r = C_r^\circ = \left\{ v \in \mathbb{R}^2,\, v \cdot w \leqslant 0 \quad \forall w \in C_r \right\} = \{0\} \times \mathbb{R}_-.$$

The collision law can be written

$$u^+ = u^- - (1 + e) P_{N_r} u^-.$$

In the multi-collisional situation, the Moreau's approach consists in straightforwardly write the previous collision law, with the appropriate notion of cone of feasible velocities and outward normal cone. Consider a many-body system of hard spheres in $\mathbb{R}^d$, centered at $r_1, \ldots, r_n$, with common radius R. The feasible set writes

$$K = \left\{ r \in \mathbb{R}^{dn}\,,\; D_{ij} = |r_j - r_i| - 2R \geqslant 0 \quad \forall i \neq j \right\}.$$

Denoting $e_{ij} = \dfrac{r_j - r_i}{|r_j - r_i|}$, the set of admissible velocities is

$$C_r = \{ v\,,\; D_{ij}(r) = |r_j - r_i| - 2R = 0 \Rightarrow e_{ij} \cdot (v_j - v_i) \geqslant 0 \}\,. \tag{5.3}$$

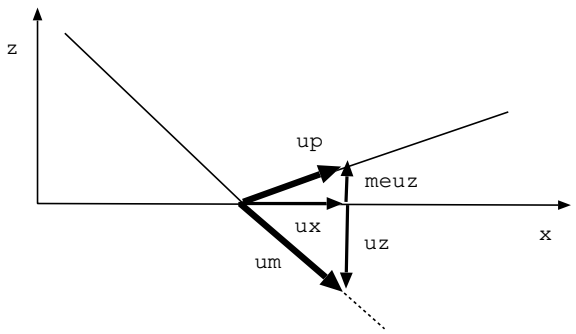


Figure 5.1: Collision against a wall. Depending on the restitution coefficient $e \in [0,1]$, the post collisional velocity can take any value between $(u_x, 0)$ (purely non-elastic) and $(u_x, -u_y)$ (elastic).

(N.B.: we use the notation $a \cdot b$ to denote the dot product of vectors in the physical space $\mathbb{R}^d$, while $\langle \cdot \mid \cdot \rangle$ shall be used for generalized velocity vectors in $\mathbb{R}^{nd}$, or elements in abstract Hilbert spaces.)

Let $r = (r_1, \ldots, r_n) \in K$ be given. As previously, the outward normal cone to K at r is defined as the polar set to the cone of feasible velocities:

$$N_r = C_r^\circ = \left\{ v \in \mathbb{R}^{dn},\ \langle v \mid w \rangle \leqslant 0 \quad \forall w \in C_r \right\}.$$

To alleviate notation, we shall now denote by U the pre-collisional velocity, and by u the post-collisional velocity. With these new notation, the collision model writes

$$u = U - (1+e) P_{N_r} U, \tag{5.4}$$

where $e \in [0,1]$ is the restitution coefficient. Since N_r and C_r are mutually polar, it holds that

$$I = P_{N_r} + P_{C_r},$$

where I is the identity operator in $\mathbb{R}^{dn}$ (see [50], or the proof of Proposition 13 in the appendix). As a consequence, the post-collisional velocity can be expressed in terms of $P_{C_r} U$, for any $e \in [0,1]$,

$$u = U - (1+e)(U - P_{C_r} U), \tag{5.5}$$

which simply reduces to $u = P_{C_r} U$ for $e = 0$. For the sake of simplicity, we shall therefore focus on this purely inelastic situation, keeping in mind that the knowledge of $P_{C_r} U$ makes it possible to recover the whole range of elastic collision laws through (5.5).

Macroscopic impact laws

We describe here informally how the Moreau's approach described above can be developed at the macroscopic scale. More details will be given in the next section. We consider an

infinite collection of inertial grains described by a macroscopic density ρ, which is subject to remain below a prescribed value, which we set at 1. We denote by $\widehat{K}$ the corresponding set of densities of a given mass, which are assumed to be supported in some domain Ω.

We denote by U the velocity field at some instant, defined on the support of ρ, and we aim at defining a collision law which would give us the post-collisional velocity from this pre-collisional velocity U. In the purely inelastic setting, a natural candidate for this law amounts to define the post-collisional velocity u_+ as the projection u of U on the set of all those vector fields which have a nonnegative divergence on the saturated zone, that is the macroscopic counterpart of C_r (defined by (5.3)). Indeed, having $\nabla \cdot u < 0$ in the neighborhood of some point in the saturated zone would lead to an increase of ρ, thereby a violation of the constraint. As will be detailed below, this cone $\widehat{C}_\rho$ can be described as the set of all those velocity fields which have a nonnegative divergence (in a weak sense) over the saturated zone.

5.2 A closer look to micro and macro impact laws

In this Section, we give some details on the mathematical formulation of the impact laws presented above, in the microscopic and macroscopic settings, and we investigate their similarities and discrepancies.

5.2.1 Saddle point formulation of the microscopic impact law

We consider as previously a system of hard spheres in $\mathbb{R}^d$, centered at $r_1, \ldots, r_n$, with common radius R. The feasible set writes

$$K = \left\{ r \in \mathbb{R}^{dn}\,,\ D_{ij} = |r_j - r_i| - 2R \geqslant 0 \quad \forall i \neq j \right\}. \tag{5.6}$$

The set of feasible velocities C_r is defined by (5.3). Let us denote by $m \in \mathbb{N}$ the number of contacts, i.e. the number of pairs $\{i,j\}$ such that $D_{ij} = |r_j - r_i| - 2R = 0$. We introduce $B \in \mathcal{M}_{m,n}(\mathbb{R})$ the matrix which expresses the constraints, each row of which is

$$G_{ij} = (0, \ldots, 0, -e_{ij}, 0, \ldots, 0, e_{ij}, 0, \ldots, 0) \in \mathbb{R}^{dn}, \tag{5.7}$$

where $e_{ij} = (r_j - r_i)/|r_j - r_i|$. The feasible set can be written

$$C_r = \{v\,,\ Bv \leqslant 0\} = B^{-1}\Lambda_+^o,\ \Lambda_+ = \mathbb{R}_+^m, \tag{5.8}$$

where Λ_+° is the polar cone to Λ_+, that is $\mathbb{R}_-^m$, and $B^{-1}\Lambda_+^\circ$ its preimage by B.

The problem which consists in projecting $U \in \mathbb{R}^{nd}$ on C_r fits into the abstract setting of Proposition 14 in the appendix, and it can be put in a saddle point form:

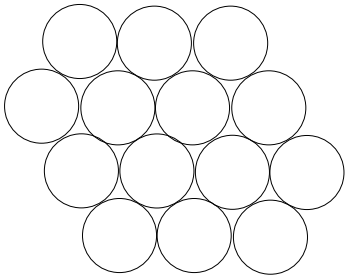


Figure 5.2: In the configuration represented here, the number of primal degrees of freedom is $2 \times n = 28$, whereas the number of contacts is $m = 29$.

Proposition 3. *Let $C_r \in \mathbb{R}^{dn}$ be defined by (5.3) (or equivalently by (5.8)). Denote by $B^\star$ the transpose of the matrix B. If $u = P_{C_r} U$ then there exists $p \in \Lambda_+ = \mathbb{R}^m_+$ such that*

$$\left| \begin{array}{rcl} u + B^\star p & = & U \\ Bu & \leqslant & 0 \\ \langle Bu \,|\, p \rangle & = & 0. \end{array} \right. \tag{5.9}$$

Conversely, if $(u, p) \in \mathbb{R}^{dn} \times \mathbb{R}^m_+$ is a solution to (5.9), then $u = P_{C_r} U$.

Proof. Let us start by a preliminary remark: the fact that the image of $B^\star$ is closed (it is a finite dimensional linear space) is not sufficient to ensure that $B^\star(\Lambda_+)$ is closed (see Remark 22 in the appendix). This property is nevertheless true here, because $B^\star(\Lambda_+)$ is spanned by a finite number of vectors

$$B^\star(\Lambda_+) = \left\{ -\sum_{ij} p_{ij} G_{ij},\, p_{ij} \geqslant 0 \right\} \tag{5.10}$$

where G_{ij} is defined by (5.7), which implies closedness by Lemma 14. Proposition 14 then ensures existence of $p \in \Lambda_+ = \mathbb{R}^m_+$ such that $u + B^\star p = U$, with the complementarity condition $\langle Bu \,|\, p \rangle = 0$. □

If we furthermore assume that $B^\star$ is one-to-one, i.e. B is onto, then p is unique. The one-to-one character of $B^\star$ is lost as soon as the number of constraints is larger than the number of degrees of freedom (hyperstatic situation). For identical disks in 2d, it can appear as soon as $n = 14$ discs are involved (see Figure 5.2). For many-body triangular lattices, the number of primal degrees of freedom is $2n$, while the number of constraints is asymptotically $3n$, which mean that the dimension of the kernel of $B^\star$ is asymptotically n.

The problem nevertheless presents some sort of uniqueness property, restricted to the homogeneous problem ($U = 0$). The following proposition essentially states a very intuitive

fact: if one considers any static configuration of a finite number of hard spheres in the open space (i.e. with no walls), under the assumption that interaction contact forces are only repulsive, then all forces are actually zero. This property will be used to show that the solution set for the pressure field (Proposition 3) is *bounded*.

Lemme 12. *We consider an admissible configuration $r \in K$, and the associated matrix $B \in \mathcal{M}_{m,n}(\mathbb{R})$ (the raws of which are given by (5.7)). The set*

$$S = \left\{ q \in \mathbb{R}_+^m,\, B^\star q = 0 \right\} = \ker B^\star \cap \Lambda_+$$

is reduced to $\{0\}$.

Proof. Let us first establish the uniqueness for the homogeneous problem. We consider $q = (q_{ij}) \in \mathbb{R}_+^m$ such that

$$B^\star q = \sum_{i \sim j} q_{ij} G_{ij} = 0.$$

where $i \sim j$ means that the particles i and j are in contact. Let i_0 denote the index of an extremal vertex of the convex hull $\mathrm{conv}(q_i,\, 1 \leqslant i \leqslant n)$. By Hahn-Banach's theorem, the compact $\{q_{i_0}\}$ and the set $\mathrm{conv}\{q_i,\, i \neq i_0\}$, which is closed and convex, can be separated in a strict sense by a plane in $\mathbb{R}^d$. We denote by x an element of this plane, and by v a normal vector to it. One has

$$(q_{i_0} - x) \cdot v > 0,\, (q_j - x) \cdot v < 0 \quad \forall j = 1, \ldots, n,\, j \neq i_0,$$

so that $(q_{i_0} - q_j) \cdot v > 0$ for $j \neq i_0$. Now the balance of contact forces exerted upon sphere i_0 in the direction v reads

$$\sum_{j \neq i_0} q_{ji_0} e_{ji_0} \cdot v$$

where $e_{ji_0} \cdot v > 0$, and $q_{ji_0} \geqslant 0$ for all j. This quantity is positive unless $q_{ji_0} = 0$ for all $j \neq i_0$. Therefore all multipliers associated to a contact with sphere i_0 are equal to 0, and this approach can be iterated for the reduced family $(q_j,\, j \neq i_0)$. By downward induction on the number of active spheres, we prove that S is reduced to $\{0\}$. □

An important consequence of this expected property is the boundedness of the solution set for (5.9).

Proposition 4. *Under the assumptions of Proposition 3, the solution set for the dual component p*

$$S = \left\{ q \in \mathbb{R}_+^m,\, B^\star q = U - u \right\} = (p + \ker B^\star) \cap \Lambda_+ \ ,$$

where (u, p) is a solution to (5.9), is bounded.

Proof. This is a direct consequence of Proposition 17 (in the appendix) and Lemma 12. □

5.2.2 Saddle point formulation of the macroscopic impact law

In the macroscopic setting, we consider that the solid phase is represented by a density supported in a domain $\Omega \in \mathbb{R}^d$, subject to remain below the value 1. We denote by $\widehat{K}$ the set of all those measures, which is the macroscopic counterpart of the set K of n-sphere configurations with no overlapping. We shall disregard here issues possibly related to wall conditions or mass at infinity: we assume that Ω is bounded, and that that the support of ρ is strongly included in Ω (i.e. the support of ρ is at a positive distance from $\partial\Omega$, which we denote by $\omega \subset\subset \Omega$).

We define a pre-collisional configuration as a density in $\widehat{K}$, together with a pre-collisional velocity field

$$U \in L^2_\rho(\Omega)^d = \left\{ v : \Omega \to \mathbb{R}^d,\ \rho - \text{measurable},\ \int_\Omega |v|^2 \,\mathrm{d}\rho < \infty \right\}.$$

We describe here a natural way to define a post collisional velocity u, natural in the sense that it directly follows the same principles as the microscopic law. This strategy to define a post-collisional velocity in the purely inelastic setting follows the framework proposed in [45, 46] for macroscopic crowd motion models. Feasible velocities are those which are non-concentrating in the saturated zone $[\rho = 1]$. For smooth velocities, it amounts to prescribe a nonnegative divergence in this zone. Such a set can be properly defined by duality as

$$\widehat{C}_\rho = \left\{ v \in L^2_\rho(\Omega)^d,\ \int_\Omega v \cdot \nabla q \,\mathrm{d}\rho \leqslant 0 \quad \forall q \in \Lambda_\rho,\ q \geqslant 0 \quad a.e. \right\} \tag{5.11}$$

where the space Λ_ρ of pressure test functions is defined as

$$\Lambda_\rho = \left\{ p \in H^1(\Omega),\ p(1-\rho) = 0 \ a.e. \right\}. \tag{5.12}$$

Note that, since we assumed that the support of ρ is strongly included in Ω, it holds that $\Lambda_\rho \subset H^1_0(\Omega)$.

It can be easily checked that, for a smooth velocity field v and a regular saturated zone, belonging to $\widehat{C}_\rho$ is equivalent to verifying $\nabla \cdot v \geqslant 0$ on $[\rho = 1]$.

The non-elastic collision law writes

$$u = P_{\widehat{C}_\rho} U,$$

where the projection $P_{\widehat{C}_\rho}$ is with respect to the L^2_ρ norm.

Let us check that it fits into the abstract setting of Proposition 14. with $V = L^2_\rho(\Omega)^d$, and $\Lambda = \Lambda_\rho$. We define $\Lambda_+ \subset \Lambda_\rho$ as the set of all those functions in Λ_ρ which are nonnegative almost everywhere:

$$\Lambda_+ = \{ q \in \Lambda_\rho,\ q(x) \geqslant 0 \ a.e. \text{ in } \ \Omega \}. \tag{5.13}$$

We introduce

$$\widehat{B} \ : \ v \in V = L^2_\rho(\omega)^d \longmapsto \widehat{B}v \in \Lambda',$$

where $\widehat{B}v$ is defined by

$$\left\langle \widehat{B}v, p \right\rangle = \int v \cdot \nabla p \, \mathrm{d}\rho. \tag{5.14}$$

Note that Λ and Λ' are not identified here, and that $\widehat{B}$ maps V onto Λ', so that the adjoint operator $\widehat{B}^\star$ is defined in $\mathcal{L}(\Lambda, V)$, the set of continuous linear mappings from Λ into V.

The saddle-point formulation of the problem can be written

$$\left| \begin{array}{rcll} u + \nabla p & = & U & \rho\text{-a.e. in } \Omega, \\ \text{“} -\nabla \cdot u & \leqslant & 0 & \text{in } [\rho = 1]\text{”}, \\ p & \geqslant & 0 & \rho\text{-a.e. in } \Omega, \\ \displaystyle\int_\Omega u \cdot \nabla p \, \mathrm{d}\rho & = & 0, & \end{array} \right. \tag{5.15}$$

where the second equation (between quotation marks) is meant in a weak sense, i.e.

$$\int_\Omega u \cdot \nabla q \, \mathrm{d}\rho \leqslant 0 \quad \forall q \in \Lambda_+.$$

This condition can also be written in an abstract way: $\widehat{B}u \in \Lambda_-$, where Λ_- is the polar cone to Λ_+, i.e. the cone of all those linear forms in Λ' which are nonpositive over Λ_+.

We may now state the well-posedness result.

Proposition 5. *Let $\rho \in \widehat{K}$ be given as a density defined over a bounded domain Ω, with* $\mathrm{supp}(\rho)$ *strongly included in Ω. Problem* (5.15) *admits a unique solution $(u, p) \in V \times \Lambda_+$, and the primal component u of this solution is the projection of U on $\widehat{C}_\rho$.*

Proof. From Proposition (14) (more precisely Corollary 1), it is sufficient to prove that $\widehat{B}$ (defined by (5.14)) is onto. Let us prove that there exists a constant $\beta > 0$ such that for every $q \in \Lambda$

$$\left\| \widehat{B}^\star q \right\|_{L^2_\rho} \geqslant \beta \|q\|_{H^1},$$

which writes $\|\nabla q\|_{L^2_\rho} \geqslant \beta \|q\|_{H^1}$ in the present context. Due to Poincaré Inequality, which holds true because $\Lambda \subset H^1_0(\Omega)$, it is sufficient to establish that the inequality $\|\nabla q\|_{L^2_\rho} \geqslant \beta \|\nabla q\|_{L^2}$ holds for any $q \in \Lambda$.

For $q \in \Lambda$, by Theorem 1.56 in [69] one has $(1 - \rho)\nabla q = (1 - \rho)\,\mathbf{1}_{q \neq 0} \nabla q = 0$, so that $\|\nabla q\|_{L^2_\rho} = \|\nabla q\|_{L^2}$. As a consequence, $\widehat{B}^\star$ has a closed range, and so does $\widehat{B}$ by Banach-Steinhaus Theorem. The range of $\widehat{B}$ is also dense thanks to the same inequality, thus $\widehat{B}$ is onto. □

5.3 Micro-macro issues

We detailed in the previous section impact laws for a collection of rigid spheres, in the Moreau's spirit, and we proposed a natural instantiation of the same principles at the macroscopic level. The macroscopic version may appear as a natural candidate to handle collision between clusters of infinitely many hard spheres represented by a diffuse density. We shall see here that some considerations may comfort this standpoint in the one-dimensional setting. Yet, for dimensions $d \geqslant 2$, we shall prove that the macroscopic law presented in the previous section is *not* a relevant model for describing the impact between large collections of hard spheres.

One dimensional setting

In the one-dimensional setting (hard spheres move on a fixed line) the two approaches are mutually consistent, as we shall see here.

First, the notion of maximal density is well defined at the microscopic level: a cluster of spheres (represented by segments in 1d) is saturated if the solid phase covers some zone of the real line, which corresponds to $\rho = 1$ in the macroscopic setting.

Now consider such a cluster of n segments covering an interval $I \in \mathbb{R}$, and the corresponding macroscopic density $\rho = \mathbf{1}_I$ (characteristic function of I). We consider a pre-collisional velocity field U that pushes the configuration against the boundary of the feasible set, i.e. such that $\partial_x U \leqslant 0$. In this case the constraint will be saturated overall the cluster so that, at the macroscopic scale, $-\nabla \cdot u = -\partial_x u = 0$, and Problem (5.15) is a classical Darcy problem

$$\left| \begin{array}{rcll} u + \partial_x p & = & U, & \\ -\partial_x u & = & 0 & \text{in } I. \end{array} \right. \tag{5.16}$$

Eliminating the velocity yields a Poisson problem on the pressure

$$-\partial_{xx} p = -\partial_x U, \tag{5.17}$$

with Dirichlet boundary conditions on the boundary of I.

At the microscopic level, we simply consider pre-collisional velocities $U_1, \ldots, U_n$, with $U_i = U(q_i)$, and we make a slight abuse of notation by keeping U to denote the vector of velocities. Since the velocities $U_1, \ldots, U_n$, are non-increasing, the constraint will also be saturated, which leads to a Darcy-like problem

$$\left| \begin{array}{rcl} u + B^\star p & = & U \\ Bu & = & 0. \end{array} \right. \tag{5.18}$$

Eliminating the velocity yields a Poisson-like problem

$$BB^\star p = BU, \tag{5.19}$$

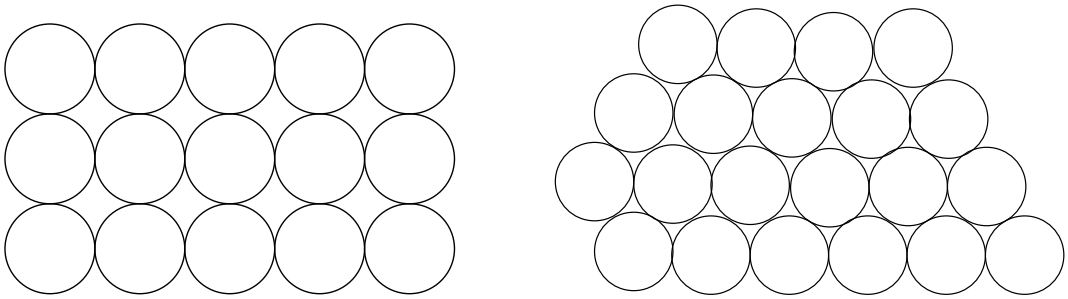


Figure 5.3: Square (left) and triangular (right) lattices.

with

$$B = \begin{pmatrix} 1 & -1 & 0 & \dots & \dots \\ 0 & 1 & -1 & \dots & \dots \\ 0 & 0 & \ddots & \ddots & \dots \\ 0 & 0 & \dots & 1 & -1 \end{pmatrix} \in \mathcal{M}_{n-1,n}(\mathbb{R}),$$

and

$$BB^\star = \begin{pmatrix} 2 & -1 & 0 & \cdot & \cdot & 0 \\ -1 & 2 & -1 & 0 & \cdot & \cdot \\ 0 & -1 & \cdot & \cdot & & \cdot \\ \cdot & & \cdot & \cdot & \cdot & \cdot \\ \cdot & & & \cdot & 2 & -1 \\ 0 & \cdot & \cdot & 0 & -1 & 2 \end{pmatrix} \in \mathcal{M}_{n-1}(\mathbb{R}),$$

that is the discrete Laplacian matrix. The two formulations are mutually consistent in the sense that the linear system (5.19) is a standard finite difference discretization of the Poisson problem (5.17), which is covered by rigorous convergence results (see e.g. [4]).

Case $d \geqslant 2$

In higher dimensions the situation is fully different. First, the notion of maximal density is not clearly defined at the microscopic level. Let us consider collections of identical discs. The maximal packing density $\rho_{max} = \pi/2\sqrt{3} \approx 0.9069\ldots$, and corresponds to the triangular lattice (see Fig. 5.3, right). Yet the actual density of moving collections of rigid disks is generally strictly less than this maximal value, which does not mean that the flow is unconstrained (as the macroscopic setting would suggest). These considerations call for a clear identification of configurations which saturate the constraint. It is tempting to consider as maximal in some sense any density corresponding to such configurations, for

which there are no free disks, so that constraints are activated everywhere. The triangular lattice is clearly jammed, but so is the cartesian lattice ($\rho = \pi/4 \approx 0.79$), and it is possible to build looser jammed configurations, for example by removing some non neighboring discs from the triangular lattice. We refer to [67] for a general review on the notion of maximal random packing.

Beyond this difficulty to properly define the notion of maximal density, the microscopic and macroscopic projections exhibit deep discrepancies in dimensions higher than 1. We propose here to enlighten these discrepancies by considering the underlying Poisson problems for the pressure in both settings. Like in the one-dimensional setting, we first consider the macroscopic setting, which is in some manner *simpler* than the microscopic one, in spite of its infinite dimensional character.

The pressure can be shown, under some assumptions, to verify a Poisson like problem in the saturated zone, The first step consists in proving that the problem verifies the abstract maximum principle (see Definition 5), that is

$$\widehat{B}U \in -\Lambda_- = -\Lambda_+^\circ \Longrightarrow \exists p \in \Lambda_+ \; s.t. \; \widehat{B}\widehat{B}^\star p = \widehat{B}U.$$

Proposition 6. *We assume that* $\mathrm{supp}(\rho)$ *is strongly included in* Ω*, and that* Ω *is connected. The couple* $(\widehat{B}, \Lambda_+)$ *verifies the maximal principle (Definition 5).*

Proof. Since $\widehat{B}^\star$ is one to one, it amounts to check that, if $\widehat{B}U \in -\Lambda_-$, i.e. if U is such that

$$\int_\Omega U \cdot \nabla q \geqslant 0 \quad \forall q \in \Lambda, \, q \geqslant 0 \; a.e., \tag{5.20}$$

then the (unique) solution $p \in \Lambda$ to

$$\int_\Omega \nabla p \cdot \nabla q = \int_\Omega U \cdot \nabla q \quad \forall q \in \Lambda$$

takes nonnegative values almost everywhere, i.e. it lies in Λ_+. This property takes the form of a maximum principle for the Laplace operator, in an extended sense: the saturated zone $[\rho = 1]$ may be not be the closure of an open domain, it may in particular have an empty interior, while having a positive measure (see Remark 12). This property is obtained by a standard procedure, which consists in taking a test function equal to the negative part of p, i.e. $q = p^- = -\min(0, p)$. We have that $\nabla q = -\nabla p \mathbf{1}_{p \leqslant 0}$ (see Theorem 1.56 in [69]), and $q \geqslant 0$, so that

$$-\int_\Omega |\nabla p^-|^{\,2} = \int_\Omega U \cdot \nabla p^- \geqslant 0,$$

which implies that ∇p^- vanishes almost everywhere, i.e. p^- is constant on Ω. Since it is 0 in the neighborhood of the boundary, it vanishes on Ω i.e. $p \geqslant 0$ a.e. in Ω. □

In other words, if the pre-collisional field is non-expansive, i.e. $\nabla \cdot U \leqslant 0$, then the pressure field p is a weak solution to the Poisson problem

$$-\Delta p = -\nabla \cdot U, \tag{5.21}$$

on the saturated zone $[\rho = 1]$, with homogeneous boundary conditions.

Let us add that the PDE above can only we legitimately written under certain conditions on the saturated zone. If the latter presents some pathologies, for example if it has an empty interior (like the complement of a dense open set) and yet a positive measure, then p might be non-trivial (Λ is not reduced to $\{0\}$), whereas (5.21) is *not* even verified in the sense of distributions. Indeed, since Λ does not contain any smooth function (except 0), the formulation (5.20) is much weaker than (5.21) considered in the sense of distribution.

A proper Poisson problem can be recovered under some additional assumptions, for instance if the saturated zone is "regular" in the sense that $[\rho = 1] = \overline{\omega}$ where ω is a smooth domain. The condition that is actually needed is actually the following: ω is such that Λ (defined by (5.12)) is equal to the closure of $C_c^\infty(\omega)$ in $H^1(\Omega)$ (the functions of $C_c^\infty(\omega)$ being extended by 0 outside ω). Under these conditions, (5.20) implies that p is a weak solution (in a standard sense) to the Poisson problem (5.21).

Remarque 12. *We mentioned the fact that the space Λ might contain no smooth function, while being non-trivial. We prove in this remark that it is indeed the case. We propose to investigate the case where the saturated zone $[\rho = 1]$, which contains the support of all those functions in Λ, is the complement of a dense open set ω, which excludes any nontrivial smooth function. We describe below how to build a nontrivial function in Λ. We assume $d = 2$, but the approach can be straightforwardly extended to higher dimensions. We consider a sequence $(c_n) \in \Omega^{\mathbb{N}}$ that is dense in Ω, and a sequence $(R_n) \in (0, +\infty]^{\mathbb{N}}$ such that $\sum \pi R_n^2 \leqslant |\Omega|/2$. For a given $r_n < R_n$, we denote by γ_n the circle of radius r_n, centered at c_n, by Γ_n the cocentric circle of radius R_n, and by Ω_n the ring domain between these circles. We denote by g_n the solution to the following Dirichlet problem in Ω_n*

$$\left| \begin{array}{rcll} -\Delta g & = & 0 & \text{in } \Omega_n, \\ g & = & 0 & \text{on } \gamma_n, \\ g & = & 1 & \text{on } \Gamma_n, \end{array} \right. \tag{5.22}$$

extended by 0 inside the small disc, and by 1 outside the large one. Since the capacity of a point is 0 in $\mathbb{R}^2$ (see e.g. [47]), one can choose r_n, with $0 < r_n < R_n$, sufficiently small to ensure that

$$\int_\Omega |\nabla g_n|^2 \leqslant \frac{1}{2^{2n}}.$$

We denote by ω the union of the small discs (centered at c_n, with radius r_n), which is open and dense by construction. Now consider the function $G_n = g_1\, g_2 \dots g_n$. It holds that

$$\nabla G_n = \sum_{k=0}^{n} \nabla g_n \sum_{j \neq k} g_j,$$

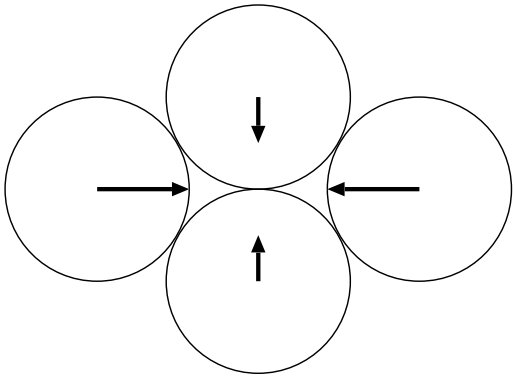


Figure 5.4: In the configuration represented here, considering that the distances are subject to remain 0 (constraint $Bu = 0$), the pre-collisional velocity tends to push any two grains in contact toward overlapping, and yet the pressure between the two grains in the center will be negative.

so that (all the g_j take values between 0 and 1 by construction), by the triangular inequality in $L^2(\Omega)$,

$$\|\nabla G_n\|_{L^2(\Omega)} \leqslant \sum_{k=0}^{n} \|\nabla g_k\|_{L^2(\Omega)} \leqslant \sum_{k=1}^{n} \frac{1}{2^k} \leqslant 2.$$

The sequence (G_n) is therefore bounded in $H^1(\Omega)$ (the gradient is bounded in L^2, and they all vanish in the first disc centered at c_0, with radius r_0). One can extract a sub-sequence which weakly converges in $H^1(\Omega)$ to some function $G \in H^1(\Omega)$. Since the convergence is strong in L^2 by Rellich's Theorem, the convergence (up to a subsequence) holds almost everywhere, so that G is by construction equal to 1 almost everywhere in the complement of the union of the large discs (centered at c_n, with radius R_n). By assumption on R_n, the measure of this set is positive (larger than $|\Omega|/2$), so that G is different from 0, while vanishing by construction in the dense union ω of the small discs.

At the microscopic level the picture is different in general. In particular, the approach carried out in the proof on the previous proposition is no longer valid. The difficulty comes from the fact that $BB^\star$, which is the straight analog of $-\Delta$ in the one-dimensional setting, does *not* verify any maximum principle in general.

Consider the simple situation represented in Figure 5.4, with a pre-collisional velocity directed toward the center. If we consider (like in the proof of the previous proposition) the problem with an equality constraint ($Bu = 0$, which means that the hard grains are glued together), eliminating the velocity leads to a discrete Poisson problem

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

Since the horizontal velocities have a much larger magnitude, in spite of the fact that the pre-collisional velocity pushes the grains against each other (i.e. $BU > 0$), it is clear that the pressure associated with the contact between the two central grains will be *negative*, which rules out the maximum principle (in the sense of Definition 5). As a consequence, the solution to the impact problem, with a unilateral constraint, will not be the same: the grains at the center will be pushed apart during the collision, which implies (thanks to the complementarity constraint $\langle Bu \,|\, p \rangle = 0$) that the corresponding pressure is 0. Note that some sort of Poisson Problem can be recovered for the pressure associated with the impact law, by *removing* the raws of B which correspond to non activated contacts, i.e. with $-G_{ij} \cdot u < 0$. If one denote by $\overline{B}$ the corresponding matrix, and by $\overline{p}$ the corresponding pressure, it holds that

$$\overline{B}\,\overline{B}^{\star}\overline{p} = \overline{B}\,\overline{U},$$

with a reduced matrix $\overline{B}$ which may also not verify the maximum principle.

This violation of the maximum principle for $BB^{\star}$ is generic in the hard-sphere, microscopic, setting, as soon a dense collections of grains are concerned. It can be checked for simple situation that the matrix $BB^{\star}$, unlike in the one-dimensional setting, has *positive* off-diagonal entries.

Square and triangular lattices

As an illustration of the previous considerations, and as an introduction to the next section, let us make some remarks on very specific situations, where the overall behavior of a collection of rigid discs can be seen to significantly differ from the behavior given by the macroscopic impact law.

Consider at first a jammed configuration structured according to a square lattice (see Figure 5.3 (left)). On each row, the non-overlapping constraints impose horizontal velocities to be non-decreasing. Similarly, on each column, the vertical velocities must be non-decreasing also. Two fields of Lagrange multipliers can therefore be associated to the constraints in the main directions x and y, which act on the system independently from each other. As a consequence, two constraints must be verified, to be compared to the single scalar constraint of the macroscopic constraint $\nabla \cdot u \geqslant 0$.

In the case of a triangular lattice (see Figure 5.3 (right)), the monotonicity of the velocity is imposed in each of the 3 principal directions.

In both cases, the microscopic constraints are much stronger than the macroscopic one, which is therefore obviously irrelevant to model at the macroscopic scale the collections of hard discs. The next section is dedicated to designing macroscopic models more respectful of the underlying microscopic structure, in the case of crystal-like configurations.

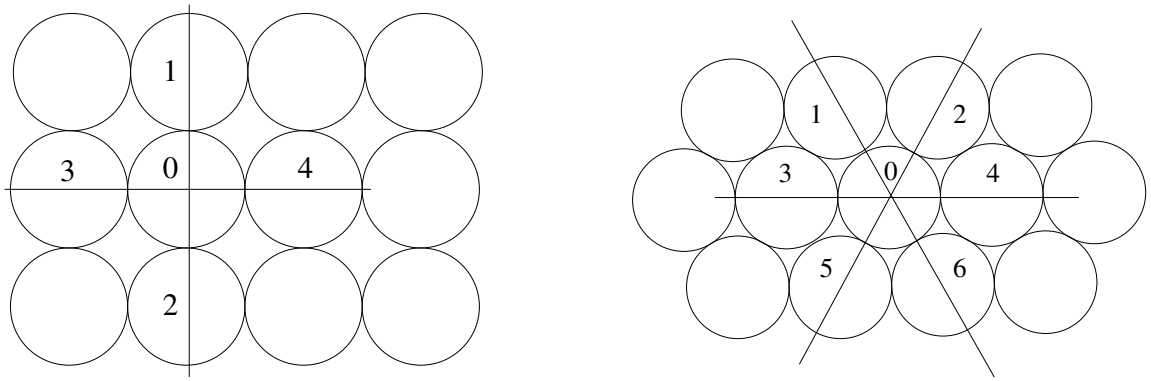


Figure 5.5: The two structured jammed configuration. On both side, the spheres in contact with the sphere 0 are labelled from 1 to 4 or 6.

5.4 Anisotropic macroscopic collision laws

We develop here some macroscopic models intended to represent configuration of jammed grains introduced in section 5.3, namely configurations that are structured in a periodic way. The approach is the following: starting from the constraints at the microscopic level, we extend them to the macroscopic level.

Square lattice

We first propose a macroscopic model adapted from the microscopic configuration of spheres jammed on a cartesian lattice, as depicted on Figure 5.5. Let us study the constraints on the velocity of the central sphere denoted by 0 in the Figure: there are four, each one corresponding to a contact with an adjacent sphere. The microscopic constraint for spheres in contact described by (5.3) writes here

$$\left| \begin{array}{l} (u_1 - u_0) \cdot e_y \geqslant 0 \\ (u_0 - u_2) \cdot e_y \geqslant 0 \\ (u_4 - u_0) \cdot e_x \geqslant 0 \\ (u_0 - u_3) \cdot e_x \geqslant 0 \end{array} \right. \tag{5.23}$$

System (5.23) can be reformulated in a more concise way: the quantities $u_x = u \cdot e_x$ and $u_y = u \cdot e_y$ must be non decreasing along each axis. In a macroscopic setting, we want thus to translate this constraint by subjecting $\partial u_x/\partial x$ and $\partial u_y/\partial y$ to be nonnegative in some sense, considering that u has L^2 regularity. To that purpose, we define the directional derivatives of the two components in a dual way, imposing

$$-\int_\Omega u_x \frac{\partial p}{\partial x} \geqslant 0 \tag{5.24}$$

for every nonnegative test function p such that its weak partial derivative in x can be defined. In order to clarify this last condition, we will introduce anisotropic Sobolev spaces, naturally defined to formalize the notion of "weakly derivable along one direction". The following description of these spaces is extracted from [35]. In what follows, Ω is a strictly convex bounded open set, with regular boundary. We refer to [35] for the study of more general domains.

Définition 3. *The anisotropic Sobolev space in the direction x on Ω is defined by:*

$$H^1_x(\Omega) = \left\{ f \in L^2(\Omega),\ \frac{\partial f}{\partial x} \text{ weakly exists in } L^2(\Omega) \right\} \tag{5.25}$$

where "$\frac{\partial f}{\partial x}$ weakly exists in $L^2(\Omega)$" means that

$$\forall g \in C^1(\Omega),\ \int_\Omega \frac{\partial f}{\partial x} g = -\int_\Omega \frac{\partial g}{\partial x} f. \tag{5.26}$$

This space is endowed with the norm $\|f\|^2_{H^1_x} = \|f\|^2_2 + \left\| \frac{\partial f}{\partial x} \right\|^2_2$

We then define $H^1_{0,x}(\Omega)$ as the closure of $C^\infty_c(\Omega)$ in $H^1_x(\Omega)$, and $H^{-1}_x(\Omega)$ as the dual of $H^1_{0,x}(\Omega)$. Since integrating along a single direction is sufficient to prove Poincaré inequality in the usual Sobolev space $H^1_0(\Omega)$ (see Proposition 9.18 and Corollary 9.19 in [9]), the anisotropic counterpart of Poincaré inequality holds :

Proposition 7. *There exists $c > 0$ such that for every $f \in H^1_{0,x}(\Omega)$,*

$$\|f\|_2 \leqslant c \left\| \frac{\partial f}{\partial x} \right\|_2 \tag{5.27}$$

We may now introduce the macroscopic model corresponding to a square microscopic structure. Let $\rho \in L^\infty(\Omega)$, such that $\rho \leqslant 1$ a.e., we still denote by ρ the measure of density ρ with respect to the Lebesgue measure. In order to avoid boundary issues, we assume that all the measures we consider are supported on a set strongly included in Ω (that is to say, as previously, at positive distance to $\partial\Omega$). In order to alleviate notation (and to deal with realistic situations when the grains configuration is structured), we assume ρ to be saturated over all its support, so that there is no need in mentioning the dependence in ρ in the notation.

Problem 1. *Given $\rho \leqslant 1$ and $U \in L^2(\Omega)^2$, find $u = (u_x, u_y) \in L^2(\Omega)^2$ that realizes the projection*

$$\min_{u \in C^x \cap C^y} \int_\Omega \|u - U\|^2_2 \, \mathrm{d}\rho \tag{5.28}$$

where the constraints set C^x and C^y are defined by duality:

$$C^\alpha = \left\{ u \in L^2(\Omega), \int_\Omega \frac{\partial q}{\partial e_\alpha} u_\alpha \leqslant 0 \quad \forall q \in \Lambda_\alpha,\, q \geqslant 0 \quad a.e \right\} \tag{5.29}$$

and $\Lambda_\alpha = \left\{ q \in H^1_{0,e_\alpha}(\Omega), q(1-\rho) = 0 \quad a.e. \right\}$ *for* $\alpha = x, y$.

Since C^x and C^y are closed convex cones, the projection problem 1 admits a unique solution. We can write the saddle-point formulation of the problem, that is the instantiation of the abstract formulation (5.51) to the present situation.

Proposition 8. *There exists a unique pair of nonnegative Lagrange multipliers (or pressures)* $(p_x, p_y) \in \Lambda_x \times \Lambda_y$ *such that*

$$u + \frac{\partial p_x}{\partial x} e_x + \frac{\partial p_y}{\partial y} e_y = U \quad \rho\text{-}a.e. \tag{5.30}$$

Proof. The constraint set $C = C^x \cap C^y$ can be written

$$C = \left\{ u \in L^2(\Omega)^2, \langle Bu \mid (q_x, q_y) \rangle \leqslant 0, \forall (q_x, q_y) \in \Lambda_+ \right\}$$

where

$$B : \left\{ \begin{array}{ccc} L^2(\Omega)^2 & \longrightarrow & H_x^{-1}(\Omega) \times H_y^{-1}(\Omega) \\ u & \longmapsto & \left(-\frac{\partial u_x}{\partial x}, -\frac{\partial u_y}{\partial y} \right) \end{array} \right.$$

and $\Lambda_+ = \{(q_x, q_y) \in \Lambda_x \times \Lambda_y,\, q_x \geqslant 0,\, q_y \geqslant 0 \quad a.e.\}$. By Poincaré inequality, there exists a constant $\beta > 0$ such that $|B^\star \mu| \geqslant \beta |\mu|$, with

$$B^\star : \left\{ \begin{array}{ccc} H^1_{0,x}(\Omega) \times H^1_{0,y}(\Omega) & \longrightarrow & L^2(\Omega)^2 \\ (q_x, q_y) & \longmapsto & \frac{\partial q_x}{\partial x} e_x + \frac{\partial q_y}{\partial y} e_y. \end{array} \right.$$

Corollary 1 in the appendix guarantees the existence of a pair of Lagrange multipliers, the uniqueness comes from the one-to-one character of $B^\star$ given by the same inequality. □

The saddle-point formulation (5.30) can be projected onto the two axes, leading to two independant systems. The problem then reduces to finding

$$(u_x, p_x) \in L^2(\Omega) \times H^1_{0,x}(\Omega) \text{ and } (u_y, p_y) \in L^2(\Omega) \times H^1_{0,y}(\Omega)$$

solutions to

$$\left| \begin{array}{ll} \frac{\partial p_x}{\partial_x} + u_x = U_x & \rho\text{-a.e.} \\ -\frac{\partial u_x}{\partial x} \leqslant 0 & \text{where } \rho = 1, \end{array} \right. \tag{5.31}$$

$$\left|\begin{array}{lll} \dfrac{\partial p_y}{\partial_y} + u_y & = U_y & \rho\text{-a.e.} \\ -\dfrac{\partial u_y}{\partial y} & \leqslant 0 & \text{where } \rho = 1. \end{array}\right. \tag{5.32}$$

This is the macroscopic counterpart of what we had seen at the end of Section 5.3: two independant pressure fields appear, acting separately on each component of the velocity in order to correct the compressions in x and y.

Remarque 13. *This model introduces anisotropy, so that the collision is no longer rotationally invariant: Figure 5.6 shows the situation of two colliding blocks, under three angles of impact. In the case of an impact along one of the two principal directions, no perturbation occurs in the tangential direction whereas in the case of an impact involving both directions (second case in Figure 5.6), a transverse velocity appear. Note that in the third case of two blocks colliding along a direction that is very close to one of the two axes, the post-impact velocity is mainly directed along the transverse direction.*

We shall now build a more pathological macroscopic model derived from the microscopic configuration of a triangular (or hexagonal) stack of particles (see Figure 5.5, right). The well-posedness of the saddle-point formulation (i.e. existence and uniqueness of pressure fields) is more delicate than before: uniqueness is lost, as we shall see later on, and the existence is still an open problem. We introduce unit vectors along the principal directions:

$$e_0 = (1,0),\ e_1 = \left(-\frac{1}{2}, \frac{\sqrt{3}}{2}\right),\ e_2 = \left(-\frac{1}{2}, -\frac{\sqrt{3}}{2}\right).$$

In this case, any sphere has 6 neighbors, two along each axis directed by the e_i. As for the previous configuration, one can write the microscopic constraints on the sphere 0

$$\left|\begin{array}{l} (u_4 - u_0)\cdot e_0 \geqslant 0,\ (u_0 - u_3)\cdot e_0 \geqslant 0,\ (u_1 - u_0)\cdot e_1 \geqslant 0 \\ (u_0 - u_6)\cdot e_1 \geqslant 0,\ (u_5 - u_0)\cdot e_2 \geqslant 0,\ (u_0 - u_2)\cdot e_2 \geqslant 0 \end{array}\right. \tag{5.33}$$

that can be reformulated by saying that $u_i = u \cdot e_i$ has to be increasing along each axis of constraint. We can thus write the macroscopic model reformulating this monotonicity exactly as above: the non-overlapping constraint becomes $Bu \leqslant 0$, with

$$B : \left\{\begin{array}{ccc} L^2(\Omega)^2 & \longrightarrow & H^{-1}_{e_0}(\Omega) \times H^{-1}_{e_1} \times H^{-1}_{e_2}(\Omega) \\ u & \longmapsto & (-\partial_{e_0} u_0, -\partial_{e_1} u_1, \partial_{e_2} u_2) \end{array}\right. \tag{5.34}$$

where u_i is the projection of u on the vector e_i. The triangular macroscopic model then writes

Problem 2. *Given $\rho \leqslant 1$ and $U \in L^2(\Omega)^2$, find $u \in L^2(\Omega)^2$ that realizes the projection*

$$\min_{Bu \leqslant 0} \int_\Omega \|u - U\|_2^2 \, \mathrm{d}\rho. \tag{5.35}$$

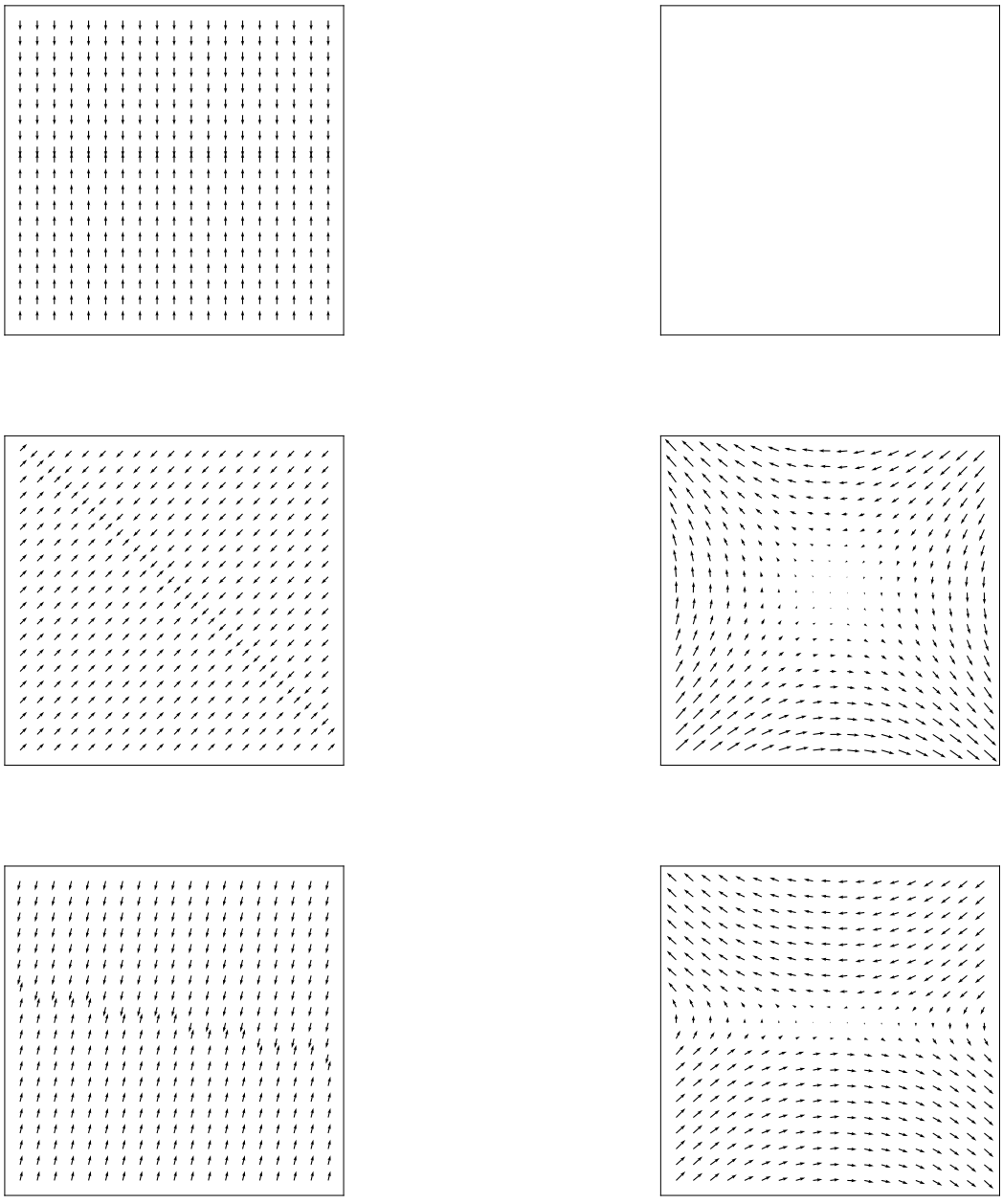


Figure 5.6: Three impacts between opposing blocks, varying the angle of incidence. On the left, the velocity fields before the impact; on the right, the velocity fields after the impacts.

The constraint $Bu \leqslant 0$ is to be interpreted in a dual way: we then require $\langle Bu \,|\, q \rangle \leqslant 0$ for every nonnegative pressures $q \in \Lambda_0 \times \Lambda_1 \times \Lambda_2$, with

$$\Lambda_i = \left\{ q_i \in H^1_{0,e_i}(\Omega)\,,\; q\,(1-\rho) = 0 \quad a.e. \right\},\; i = 0, 1, 2.$$

Since Problem 2 consists in projecting on a closed convex cone, it admits a unique solution. The saddle-point formulation reads:

$$\text{Find } (u, p_0, p_1, p_2) \in \left(L^2(\Omega)^2 \times \Lambda_0 \times \Lambda_1 \times \Lambda_2 \right) \text{ such that}$$

$$\left| \begin{array}{c} u + \displaystyle\sum_{i=0}^{2} \partial_{e_i} p_i e_i = U, \\ p_i \geqslant 0, \\ \displaystyle\sum_{i=0}^{2} \int_\Omega u_i \partial_{e_i} p_i = 0. \end{array} \right. \tag{5.36}$$

Remarque 14. *The operator $B^\star$ is not onto, as we show here by exhibiting a non-trivial element in $\ker(B^\star)$. Consider an hexagon H included in Ω, and f a piecewise constant function equal to 0 out of H, and alternatively 1 and -1 depending on the position in H, as represented on Figure 5.7. Define now p_i as the solution of*

$$\left| \begin{array}{lll} \partial_i p_i & = f & \text{in } H \\ p_i & = 0 & \text{in } \Omega \setminus H \end{array} \right. \tag{5.37}$$

Due to the symmetries of f, this equation is compatible with the limit condition $p = 0$ on ∂H: every line directed by any e_i has an intersection of the same length with zones labelled by 1 or -1. Moreover, we have $\displaystyle\sum_{i=0}^{2} \partial_i p_i e_i = f \sum_{i=0}^{2} e_i = 0$, so p lies in $\ker(B^\star)$.

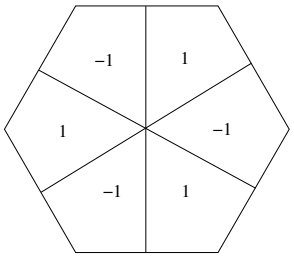


Figure 5.7: Counterexample to the injectivity of $B^\star$

Remarque 15. *It is an interesting counterpart to the microscopic counterexample presented on Figure 5.2. When the number of spheres increases, there is* $3/2$ *times more constraints than degrees of freedom; thus the dimension of the kernel of* $B^\star$ *tends to infinity in the micro case. Accordingly, the macroscopic example above provides an infinite family of independent vectors in* $\ker(B^\star)$.

5.5 Homogenization issues

This section deals with issues pertaining to the convergence of microscopic models towards macroscopic ones. Let us make it clear that such convergence is out of reach in general. We shall rather describe a general framework to address these issues, and establish some convergence results in very particular situations, in the case when the microscopic situation is structured. The idea is the following: we start from a macroscopic velocity field, and we span the domain with a sequence of saturated configurations of spheres of radius tending to 0. At each scale, we project the field on the feasible set, which contains all those fields which comply with the non-overlapping constraint.

5.5.1 General procedure

We describe here a general procedure to formalize questions concerning micro-macro convergence. First, we need to define a way to compare microscopic velocities to macroscopic fields. Given a field $U \in L^2(\Omega)^2$, and n non overlapping touching spheres in Ω, denote $D_1, \dots, D_n \subset \Omega$ the Voronoï cells associated to the spheres. Define

$$\tilde{U}_i = \frac{1}{|D_i|} \int_{D_i} U(x) \, \mathrm{d}\rho(x) \quad \forall 1 \leqslant i \leqslant n. \tag{5.38}$$

Let $\tilde{u}$ be the solution of the microscopic problem associated to $\tilde{U}$. Finally, let v be the piecewise constant function equal to $\tilde{u}$ on the cell D_i. This mapping is depicted in Figure 5.8. We have built an operator

$$\phi_n : \left\{ \begin{array}{ccc} L^2(\Omega)^2 & \longrightarrow & L^2(\Omega)^2 \\ U & \longmapsto & v \end{array} \right. \tag{5.39}$$

which maps a pre-collisional macroscopic velocity U to a post-collisional velocity, computed through projection at the microscopic level.

We are now able to formulate the homogenization problem statement, in terms of two general questions:

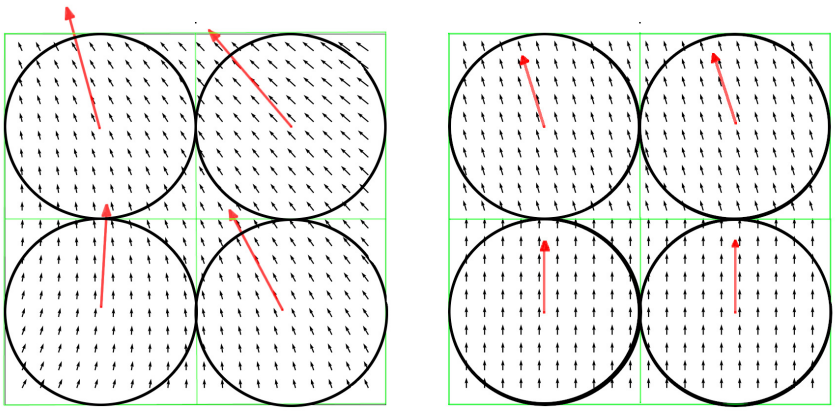


Figure 5.8: On the left, the construction of a microscopic vector field $\tilde{U}$ from the macroscopic one U, in the square configuration. The green lines delimit the Voronoï cells associated to the spheres, the red arrows are the mean of the vector field on each cell (expanded for a sake of clarity). On the right, construction of a macroscopic vector field v (black) from a microscopic field $\tilde{u}$ (red).

Homogenization of impact laws

Given a velocity field $U \in L^2(\Omega)^2$ and a sequence $(x^n) = (x_i^n)_{1 \leqslant i \leqslant n}$ of collections of hard sphere configurations, with a common radius δ_n (with $\delta_n \to 0$), such that

$$\sum_{i=1}^{n} \mathbf{1}_{B(x_i^n, \delta_n)} \tag{5.40}$$

weakly converges to some limit density when n goes to $+\infty$, what are the possible limits of $\phi_n(U)$? Is it possible to prescribe constraints on the microscopic structure so that ϕ_n converges to some projection operator at the macroscopic level, which would encode the characteristics of the microscopic structure ?

These questions should be seen as a wide research program which is way beyond the scope of these notes. We shall restrict ourselves to some short comments, and to providing a detailed answer in very specific situations (see Section 5.5.2).

First, various sorts of constraints can be expected: isotropic ones like in Section 5.2, anisotropic ones according to some principal directions like in Section 5.4, or possibly not linked to any underlying regular structure in the grain configuration. The notion of local maximal value, already discussed in Section 5.2, is also an issue: consider e.g. a configuration where a part of the saturated domain is spanned by a square lattice, and another part is

spanned by a triangular mesh. As we shall see below, the projection operators will actually converge towards an operator activated respectively when $\rho = (1 - \pi/4)$ and $\rho = \pi/2\sqrt{3}$, depending on the local microscopic structure, so that the maximal density is not defined uniformly over the saturated zone.

In all generality, when there is no reason to assume any regularity / periodicity in the microscopic structure, one may expect some sort of averaging in the direction of contacts, with a local constraint on the density based on the so-called Random Maximal Packing, that is around 0.64 for three-dimensional collections of identical hard spheres ([66]). This may legitimate an isotropic approach like the one presented in Section 5.2.2, based on a uniform maximal density, and an isotropic constraint on the velocities. Yet, as extensively described in the literature on granular media (see e.g. [59]), complex force networks are observed within collections of grains, over scales that go way beyond the microscopic size of the grains. Such observations advocate for the need to develop macroscopic models which would reflect some anisotropy at the mesoscopic scale, in the spirit of what is done in the next section for highly structured configurations.

Remarque 16. *One could question the choice of using Voronoï Cells instead of defining the field v to be constant on every sphere, and null elsewhere. The reason is that we aim at showing strong L^2 convergence results, which will not hold for velocities supported on spheres. For instance, consider the constant field $U = e_x$ for the squared configuration with radius tending to 0. As no constraint is activated, v is automatically equal to U everywhere, and similarly, $\tilde{u}_i^n = e_x$ for every $1 \leqslant i \leqslant k_n$ (k_n being the number of spheres needed to span the domain). If we define w^n piecewise constant on every sphere equal to 0 elsewhere, there subsists an irreducible gap*

$$\|U - w^n\|_2^2 = (1 - \pi/4)\,\lambda(\Omega) + o(1)$$

$\pi/4$ being the proportion of Ω spanned by the spheres for a square lattice.

5.5.2 Homogenization for structured configurations

We detail here micro-macro convergence results in very specific situations, namely when the microscopic spheres on which we interpolate in the previous subsection are organized on square or triangular lattices. More precisely, under the framework presented at the beginning of Section 5.5.1, we establish that, given a pre-collisional velocity field U, the velocities obtained by projection at the microscopic level (operators ϕ_n defined by (5.39)) converge to a velocity obtained by projection at the macroscopic level, according to the projection operators detailed in Section 5.4. Here, for the sake of simplicity we consider the whole set $\omega = [0, 1] \times [0, 1]$ to be spanned by spheres disposed on a square or triangular lattice, as in Figure 5.5. We can thus disregard the issues of maximal density raised in the previous section, as we know that the microscopic density measure will weakly converge to a constant, that we can set to 1.

First, we need to fix some common notation for the two structures considered. Let $V = L^2(\omega)^2$ be the set of macroscopic velocities, and for $n \in \mathbb{N}$, k_n the number of spheres of radius $1/n$ needed to span ω, either for the square or the triangular configuration. Denote then $\tilde{V}_n = \left(\mathbb{R}^2\right)^n$ the set of microscopic velocities, $V_n \subset V$ the set of functions constant on each Voronoï cell $D_i \subset \omega$ associated to the sphere configuration.
We are going to use a classical theorem in Numerical Analysis, designed to estimate errors for approximated problems of optimisation. Here V_n is seen as an approximation space for V. The main idea here is to approximate not only the space of functions V, but also the space of constraints. We denote by $C_\rho \subset V$ the set of velocities satisfying the macroscopic anisotropic constraint (defined in section 5.4), and $C_n \subset V_n$ the set of velocities satisfying the microscopic constraint once the velocity of a cell is attributed to the sphere in the cell: C_n is seen as an approximation of the constraint space C_ρ. We then have the following error estimate between the exact solution u and the approximate solution v:

Theorem 1 (Adapted from Falk, 74' [29])**.** *There exists $a, b > 0$ such that for every $f \in C_n$ and $g \in C_\rho$,*

$$\|u - v\|_2 \leqslant \left(a\|u - f\|_2^2 + b\|u - U\|_2 \left(\|u - f\|_2 + \|g - v\|_2\right)\right)^{1/2} \tag{5.41}$$

Given $U \in V$, denote by $\tilde{U}^n$ the approximation defined by (5.38) and $v_n = \phi_n(U)$. Denote as before u and $\tilde{u}^n$ the solutions of the macroscopic and the microscopic problems associated to U and $\tilde{U}^n$, with respect to the constraint sets C_ρ and C_n. Figure 5.8 illustrates this construction.

Proposition 9. *The sequence v_n converges towards u in $L^2(\omega)^2$.*

Proof. Let u_n^{int} be the piecewise constant function equal to $\dfrac{1}{|D_i|}\displaystyle\int_{D_i} u(x)\,\mathrm{d}x$ on every D_i. Since u is in C_ρ, one can verify that $u_n^{\text{int}} \in C_n$. On the other hand, since $\tilde{u}^n$ is the solution to the microscopic problem, v_n is in C_ρ. Therefore, we can use Theorem 1, that guarantees the existence of a, b some positive constants such that for every $n \in \mathbb{N}$

$$\|u - v_n\|_2^2 \leqslant a\left\|u - u_n^{\text{int}}\right\|_2^2 + b\,\|u - U\|_2\left\|u - u_n^{\text{int}}\right\|_2 \tag{5.42}$$

It is then sufficient to show that the piecewise constant approximation u_n^{int} tends to u in $L^2(\omega)^2$. Let $\epsilon > 0$ and $f \in C^0(\omega)^2$ be such that $\|u - f\|_2 \leqslant \epsilon$; denote f_n^{int} the piecewise constant approximation of f. We have

$$\left\|u - u_n^{\text{int}}\right\|_2 \leqslant \|u - f\|_2 + \left\|f - f_n^{\text{int}}\right\|_2 + \left\|f_n^{\text{int}} - u_n^{\text{int}}\right\|_2 \tag{5.43}$$

Using that f is uniformly continuous on ω, define $n_0 \in \mathbb{N}$ such that for every $x, y \in \omega$

satisfying $|x-y| \leqslant \dfrac{2}{n_0}$, $|f(x)-f(y)| \leqslant \epsilon$. Thus for $n \geqslant n_0$

$$\begin{aligned}\left\|f - f_n^{\text{int}}\right\|_2^2 &\leqslant \sum_{i=1}^{k_n} \int_{D_i} \left|f(x) - f_n^{\text{int}}(x)\right|^2 \mathrm{d}x \\ &\leqslant \sum_{i\in I} \int_{D_i} \epsilon^2 = \epsilon^2.\end{aligned}$$

On the other hand, using Jensen inequality

$$\begin{aligned}\left\|u_n^{\text{int}} - f_n^{\text{int}}\right\|_2^2 &= \sum_{i=1}^{k_n} \int_{D_i} \left|u_n^{\text{int}}(x) - f_n^{\text{int}}(x)\right|^2 \mathrm{d}x \\ &= \sum_{i=1}^{k_n} \lambda(D_i) \left|\frac{1}{\lambda(D_i)} \int_{D_i} (u(y) - f(y))\,\mathrm{d}y\right|^2 \\ &\leqslant \sum_{i=1}^{k_n} \int_{D_i} |u(y) - f(y)|^2 \,\mathrm{d}y \\ &= \|u - f\|_2^2\end{aligned}$$

Thus u_n^{int} converges in $L^2(\omega)^2$ towards u, and so does v_n. □

Remarque 17. *In the previous proof, two ingredients can be identified as essential in the process of elaborating general homogenization results:*

- $u_n^{\text{int}} \in C_n$*: a field that respects the macroscopic constraint must check the microscopic constraints once integrated on the Voronoï cells; and reciprocally the piecewise constant approximation* v_n *of the corrected microscopic field must satisfy the macroscopic constraint. Thus the macro/micro constraints must be compatible under the mapping that we defined above.*
- u_n^{int} *needs to converge for the* L^2 *norm toward* u*: this is in particular true if the spheres span the whole saturation area, in the sense that the diameter of the Voronoï cells tends to 0.*

5.6 Evolution models

We describe here the evolution problems which are associated to the impact laws that have been described in the first sections of these notes. Let us first make it clear that writing an evolution problem associated to the impact laws studied in Sections 5.4 and 5.5 is irrelevant a priori. Indeed, the assumptions which can be made on the microscopic structure of

a granular medium are instantaneously ruled out as soon as the medium undergoes any deformation. A macroscopic model respectful of the current state of the medium in terms of microscopic structure should rely on some parameters to reflect the local organization of grains, which strongly conditions the impact law as we detailed in the previous sections. We shall rather present evolution problems for the microscopic setting, which takes the form of a second-order in time differential inclusion, and for the macroscopic scale we shall consider the isotropic setting only (the divergence is nonnegative on the saturated zone).

Microscopic evolution problem

Like in Section 5.2.1, we consider n moving rigid spheres centered at $r_1, \ldots, r_n$, with common radius R, subject to forces $f = (f_1, \ldots, f_n)$. We denote by $m_1, \ldots, m_n$, the masses of the grains, and by $M \in \mathcal{M}_{nd}$ the associated mass matrix. We denote by K the feasible set (defined by (5.6)), by C_r the cone of feasible direction (defined by (5.8)), and by $N_r = C_r^\circ$ the outward normal cone. Note that N_r is $\{0\}$ as soon as r lies in the interior of K, i.e. when there is no contact. We shall consider[1] that $N_r = \emptyset$ for $r \notin K$. The most concise way to write a class of evolution problems for this system, considering that impacts are frictionless, is the following (see e.g.[63]):

$$M\frac{d^2r}{dt^2} + N_r \ni f. \tag{5.44}$$

When there is no contact, $N_r = \{0\}$, and we recover n independent ODE's in $\mathbb{R}^d$. The contraint $r \in K$ is implicitly prescribed because $N_r = \emptyset$ as soon as $r \notin K$. Note also that we have (see the proof of Proposition 3)

$$N_r = \left\{ -\sum_{ij} p_{ij} G_{ij},\, p_{ij} \geqslant 0, \right\} \tag{5.45}$$

where G_{ij} is defined by (5.7). It guarantees that contact forces verify the Law of Action-Reaction, and that only repulsive forces are exerted (grains do not glue to each other).

Yet, Inclusion (5.44) is essentially compatible with all impact laws which do not violate the Law of Action-Reaction, including some laws which would lead to an increase of kinetic energy. An impact law of the type (5.4) has to be prescribed. We shall now write the full evolution system, in the purely inelastic setting, and with an explicit involvement of interaction forces. In the dynamic setting, these forces are generically singular in order to instantaneously change the velocities of the grains, and we shall represent them by positive

[1] This convention is consistent with the definition of N_r as the Fréchet subdifferential of the indicatrix function I_K of K, which is indeed $\emptyset$ outside of K.

measures in time, denoted by $\mathcal{M}_+(0,T)$. In the purely inelastic setting, the system writes

$$\left|\begin{array}{lcl} M\dfrac{d^2r}{dt^2} & = & f+\displaystyle\sum_{ij} p_{ij}G_{ij} \\ p_{ij} & \in & \mathcal{M}_+(0,T) \\ \mathrm{supp}(p_{ij}) & \subset & \{\ t,\ D_{ij}(r(t))=0\} \\ u^+ & = & P_{C_r}u^-. \end{array}\right. \tag{5.46}$$

More general impact laws can be considered, by setting $u^+ = u^- - (1+e)P_{N_r}u^-$ for $e \in [0,1]$. As detailed in [5, 41], the relevance of the impact law is ensured by the fact that the velocity has bounded variations in time, so that it admits at each time left and right limits.

The system is formally well-posed in the sense that it fits in classical Cauchy-Lipschitz theory when there is no contact, and whenever a contact occurs the impact law univocally expresses the post-collisional velocity with respect to the pre-collisional one. It can be checked that kinetic energy is preserved for $e = 1$, and part of it is lost during each collision for $e < 1$.

There is indeed a well-posedness results for this system, under the condition that the forcing term f is *analytic* (see e.g. [5]). Counter-examples to uniqueness exist for the case of a single grain and a wall, in the elastic setting ($e = 1$), with a forcing term which is infinitely differentiable (see [63]).

A similar counter-example can be built in the purely non-elastic case ($e = 0$), we again refer to [63, 5] for the analytic expression of the forcing term. In order to illustrate the principle of these counter-examples, we plot in Figure 5.9 a numerical computation of two distinct solutions associated to the same forcing term, for a single particle forced toward a wall. As detailed in [41], the plotted numerical solutions correspond to two different sequences of time steps.

The macroscopic counterpart of (5.46) is the so-called Pressureless Euler equations with maximal density constraint, which describes the motion of a granular fluid made of particles which do not interact unless saturation (set at $\rho = 1$) is reached. In the purely inelastic setting, the system writes

$$\begin{cases} \partial_t\rho + \nabla\cdot(\rho u) = 0, \\ \partial_t(\rho u) + \nabla\cdot(\rho u\otimes u) + \nabla p = 0, \\ \rho \leqslant 1, \\ (1-\rho)p = 0, \\ p \geqslant 0, \\ u^+ = P_{\widehat{C_\rho}}(u^-). \end{cases} \tag{5.47}$$

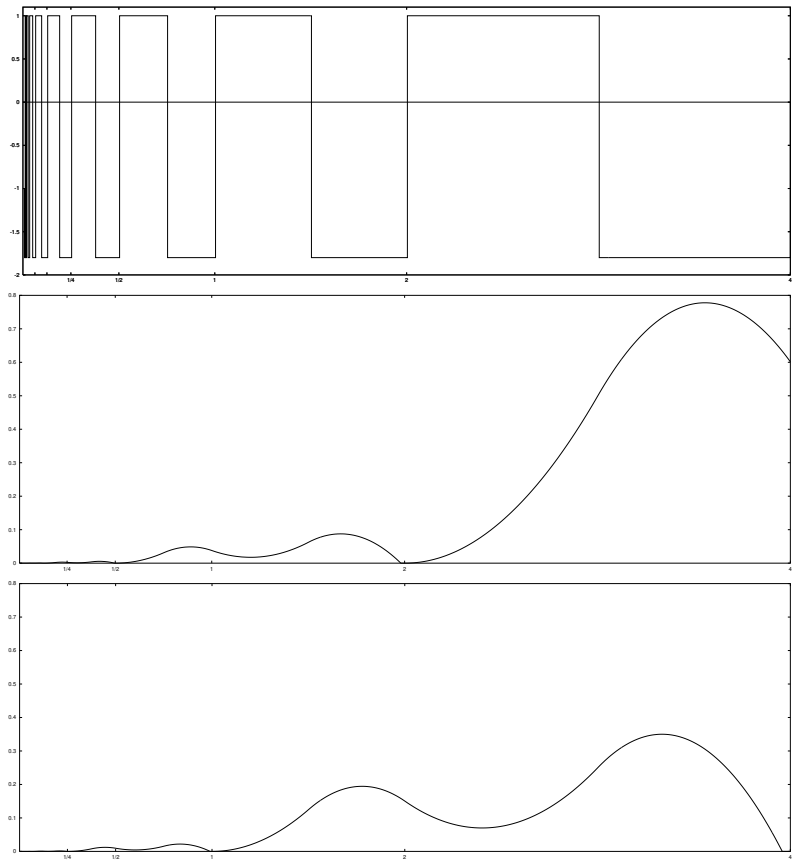


Figure 5.9: Two distinct solutions associated to the same forcing term

where $\widehat{C}_\rho$ is the cone of feasible velocities defined by (5.11). These equation must be understood in a weak sense. In particular the pressure p is likely to be very singular in time, like in the microscopic setting, and the momentum equation is meant in a distributional sense. Little is known concerning this system, which is usually written without the impact law (last equation of the system). Note that this law can be replaced, at least formally, by any law of the type

$$u^+ = u^- - (1+e)\left(u^- - P_{\widehat{C}_\rho} u^-\right),$$

with $e = 1$ for the elastic case. This equation is well-understood in the one-dimensional setting, see e.g.[12] where particular "sticky-blocks" solutions are built, and can be used to build solutions of the system. This class of solutions corresponds to the situation where the initial density is the sum of characteristic functions of segments, each one initially moving at a uniform velocity. Since no forcing term is involved, segments remains segments, possibly

merging to form larger segments, and the model can be treated exactly according to the microscopic model (5.46). Note that this approach, presented in the purely inelastic setting, could be extended to various impact laws ($e \in (0,1]$). Note also that, since sticky blocks reproduce the microscopic setting, the non-uniqueness result which we mentioned obviously extends to the macroscopic problem, if one accounts for a time-dependent forcing term.

This constructive approach does not straightforwardly extend to higher dimensions for obvious reasons: the saturated zone is likely to deform in a complex way, which makes the macroscopic model fully different from the microscopic one. An extension has nevertheless been proposed recently in [7] to build solutions to (5.47) (without the impact law), again in a purely non-elastic spirit. A numerical approach is proposed in [25] to approximate candidate solutions to (5.47). It is based on barotropic Euler equations, i.e. compressible Euler equations where the pressure is assumed to be a function of the local density of *barrier* type: it is taken in the form $p = \varepsilon p(\rho)$, where p is smooth on $[0,1)$ and blows up to $+\infty$ at 1^-. When ε goes to 0, the action of the pressure disappears in non saturated zones, whereas the blow-up at 1 prevents the density to pass the maximal value. Again, the impact law is not integrated in the global limit system, but this approach natively recovers the elastic setting ($e = 1$). In [26], a similar approach is carried out in the case of a variable congestion (the constraint $\rho \leqslant 1$ is replaced by $\rho \leqslant \rho^\star$, where $\rho^\star$ is a given, non-uniform, barrier density). We also refer to [56] for an analysis of a similar system with additional memory effects induced by the presence of an underlying viscous fluid. An alternative approach, also of the constructive type, is proposed in [44], it is based on a time discretization scheme of the splitting type: at each time step the density is transported according to the pre-collisional velocity (the congestion constraint is disregarded), possibly leading to a violation of the constraint. The density is then projected on $\widehat{K}$ according to the Wasserstein distance (like in the crowd motion model presented in [45, 46]), and the post-collisional velocity is then a posteriori built from the projected density. This approach natively restricts to the purely non-elastic setting. Let us add that these exploratory approaches do not provide a full theoretical framework to the full system (5.47) (including the impact law).

Let us add a few comments on the difficulty to handle the collision law, in the process of building solutions to the full system. The impact law (last equation of (5.47)) implicitly assumes that left and right limits exist for the velocity field, which is far from being obvious. In the microscopic setting, it is linked to the BV regularity in time of the velocity field, which makes clear sense in this purely Lagrangian setting. In the purely non-elastic setting the velocity of a given particle may undergo jumps, but each of these jumps also corresponds to a decrease of the kinetic energy. If the forcing term is controlled, the total variation due to these jumps can be shown to be bounded. In the macroscopic setting, the velocity field is defined in a Eulerian way, i.e. $u(x,t)$ corresponds to the current velocity of the medium at x, and BV character of velocities for Lagrangian particles has no clear counterpart in this Eulerian description.

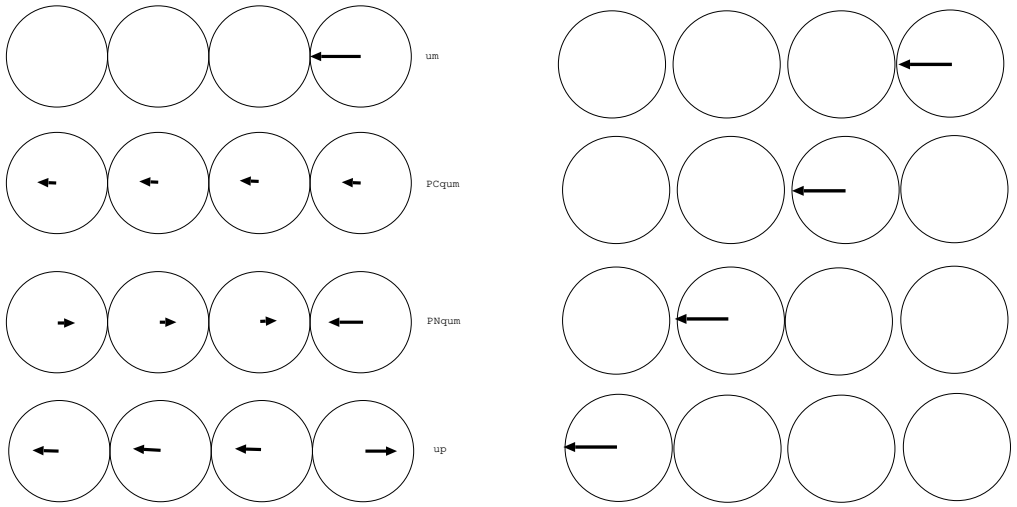


Figure 5.10: Newton's craddle: computation of u^+ with Moreau's approach, with initially touching discs (left), and slightly pulled apart discs (right).

Stability issues

As suggested by the non-uniqueness result for the evolution problems, the problem is unstable with respect to data, and in particular to grain positions. A striking illustration of this instability is given by the so-called *Newton's craddle*, which can be described as follows: a straight raw of touching identical hard spheres is hit on one of its end by a hard sphere. Actual experiments on this setting show that the apparent post-collisional velocity affects the sphere on the opposite side only, while the other spheres (including the hitting one) stay still. A straight application of the approach we presented (Moreau's approach) in the elastic setting leads to a fully different picture, presented in Figure 5.10 (left): the hitting sphere is pushed backward (i.e. rightward), almost as if it had hit a wall (the speed is slightly reduced), while the rest of the spheres are pushed leftward at a small velocity, in such a way that total momentum and kinetic energy are conserved. Yet, by considering an initial situation where grains are slightly pulled apart (initial distances set at an arbitrary small value), the experimentally observed behavior is recovered, after a series of quasi-simultaneous binary collisions as illustrated again in Figure 5.10 (right). Similar examples of the high sensitivity of the impact law to the configuration, possibly inducing significant changes in the future behavior of the system, can straightforwardly be built for the macroscopic one-dimensional problem, in the elastic setting.

Appendix

We gather here some well-known theoretical results, and some less classical ones, on the saddle-point formulation of cone-constrained minimization problems.

Let V be a Hilbert space, and $J : V \longrightarrow \mathbb{R}$ a continuously differentiable functional. We denote by $DJ(u) \in V'$ its differential at u, and by $\nabla J(u)$ its gradient:

$$J(u+h) = J(u) + \mathrm{d} < DJ(u), h > +o(h) = J(u) + \langle \nabla J(u) \mid h \rangle + o(h).$$

Linear constraints

Proposition 10. *Let K be a linear subspace of V, and u a local minimizer of J over K. Then $\nabla J(u)$ is orthogonal to K, which we can write*

$$\nabla J(u) + \xi = 0,\ \xi \in K^{\perp}.$$

Proof. Fix any $h \in K$. For $t \in \mathbb{R}$ in a neighborhood of 0,

$$J(u+th) = J(u) + t\langle \nabla J(u) \mid h \rangle + o(t) \geqslant J(u),$$

which yields $\langle \nabla J(u) \mid h \rangle = 0$. □

We now assume that $K = \ker B$, where $B \in \mathcal{L}(V, \Lambda)$, and Λ is a Hilbert space, identified to its dual space. We furthermore restrict ourselves to the case of a quadratic functional

$$v \longmapsto J(v) = \frac{1}{2}|v - U|^2, \tag{5.48}$$

for a given $U \in V$.

Proposition 11. *Let $K = \ker B$ be a linear subspace of V, and u a local minimizer of J (defined by (5.48)) over K, the linear functional ξ defined in the previous proposition lies in $\overline{B^{\star}(\Lambda)}$.*

If we assume that B has a closed range, then $\xi \in B^{\star}(\Lambda)$. If we identify V with its dual space, considering accordingly that $B^{\star}$ maps Λ to V, it means that there exists $p \in \Lambda$ such that

$$\left\{ \begin{array}{lcl} u + B^{\star}p & = & U \\ Bu & = & 0 \end{array} \right. \tag{5.49}$$

Conversely, without any assumption on B, if $(u, p) \in V \times \Lambda$ verifies (5.49), then u is the projection of U on $K = \ker B$.

Proof. We have $\xi \in K^{\perp} = \overline{B^{\star}(\Lambda)}$ (see e.g. [9]). Now if B has a closed range, $B^{\star}$ has also a closed range, so that $\overline{B^{\star}(\Lambda)} = B^{\star}(\Lambda)$, which yields (5.49).

Conversely, if (5.49) is verified, then $U - u \in K^{\perp}$, which implies that u is the projection of U on $K = \ker B$. □

Proposition 12. *Under the assumption of Proposition 11, if we furthermore assume that B is onto, then λ is unique.*

Proof. This is a straightforward consequence of $\ker(B^\star) = B(V)^\perp = \{0\}$. □

Remarque 18. *Problem* (5.49) *is commonly called saddle-point formulation of the constrained minimization problem. Indeed, if we define the* Lagrangian *of the problem as*

$$L \,:\, (v,q) \in V \times \Lambda \longmapsto L(v,q) = J(v) + \langle Bv \,|\, q \rangle,$$

then (u,p) verifies (5.49) *if and only if it is a saddle point for L in $V \times \Lambda$, i.e.*

$$L(u,q) \leqslant L(u,p) \leqslant L(v,p) \quad \forall v \in V,\, q \in \Lambda.$$

Unilateral constraints

We now consider the projection of an element on a closed convex cone C. This cone, like all the cones we shall consider in this section, admits the origin as a pole, i.e. $\mathbb{R}_+ C \subset C$. More precisely, $U \in V$ being given, we aim at minimizing

$$v \longmapsto J(v) = \frac{1}{2}|v - U|^2$$

over C. We denote by N the polar cone to C:

$$N = C^\circ = \{v \in V,\, \langle v \,|\, w \rangle \leqslant 0 \quad \forall w \in C\}.$$

Proposition 13. *(Moreau 1962)*
Let C be a closed convex cone in V and $N = C^\circ$ its polar cone. Then the identity in V decomposes as the orthogonal sum of the projections on C and N. In other words, for any $U \in V$, it holds that

$$u + \xi = U,\, u = P_C U,\, \xi = P_N U,\, \langle u \,|\, \xi \rangle = 0.$$

Conversely, if $U = u + \xi$ with $u \in C$, $\xi \in N$, and $\langle u \,|\, \xi \rangle = 0$, then $u = P_C U$ and $\xi = P_N U$.

Proof. For the sake of completeness, we give here a short proof of this standard result established in [50]. Let us first recall that, for any closed convex set, u is the projection of U on K if and only if $u \in K$ and

$$\langle U - u \,|\, w - u \rangle \leqslant 0 \quad \forall w \in K.$$

Applying this to the closed convex set C, and using the fact that C is a cone, we have that

$$\langle U - u \,|\, tw - u \rangle \leqslant 0 \quad \forall w \in C,\, t \in]0, +\infty[.$$

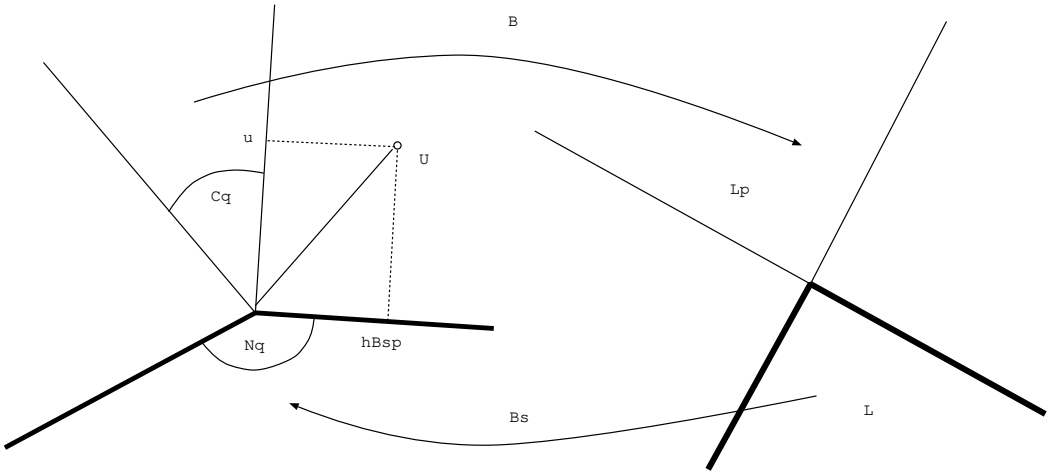


Figure 5.11: Mutually polar cones

By dividing by t, and having t go to $+\infty$, we obtain $\langle U-u \,|\, w\rangle \leqslant 0$, i.e. $U-u \in N = C^\circ$. Let us now prove that $\xi = U-u$ is the projection of U on N. For any $w \in N$

$$\langle U-(U-u) \,|\, w-(U-u)\rangle = \langle u \,|\, w\rangle + \langle U-u \,|\, 0-u\rangle.$$

The first term is nonnegative by polarity, and so is the second one because u is the projection of U on C.

Finally, $\langle u \,|\, \xi\rangle \leqslant 0$ and, since $0 \in N$,

$$0 \geqslant \langle U-\xi \,|\, 0-\xi\rangle = -\langle u \,|\, \xi\rangle,$$

so that $\langle u \,|\, \xi\rangle = 0$.

Conversely, if $U = u+\xi$ with $u \in C$, $\xi \in N$, $\langle u \,|\, \xi\rangle = 0$, then for any $w \in C$

$$\langle U-u \,|\, w-u\rangle = \langle \xi \,|\, w-u\rangle = \langle \xi \,|\, w\rangle \leqslant 0,$$

so that $u = P_C U$. The proof is similar for $\xi = P_N U$. □

We assume now that C is defined by duality as

$$C = \{v \in V,\, \langle Bv \,|\, \mu\rangle \leqslant 0,\, \forall \mu \in \Lambda_+\}\,, \tag{5.50}$$

where $B \in L(V,\Lambda)$, Λ is a Hilbert space identified to its dual, and Λ_+ is a closed convex cone in Λ (see Figure 5.11).

Remarque 19. *One may interrogate the motivation for defining a convex cone by means of another convex cone. This approach will be proven fruitful in many situations where*

K is natively described in an implicit way, i.e. as the collection of elements which verify certain unilateral constraints, whereas Λ_+ is defined in a explicit way, like $\mathbb{R}^d_+$ in the finite dimensional setting, or as a subset of real functions taking nonnegative values, so that projecting on Λ_+ can be computed straightforwardly.

Lemme 13. *Let C be a closed convex cone in V, defined by* (5.50)*. It holds that*

$$N = C^\circ = \{w \in V,\ \langle w \mid v\rangle \leqslant 0,\ \forall v \in C\} = \overline{B^\star \Lambda_+}.$$

Proof. For any $\mu \in \Lambda_+$, any $v \in C$, it holds $\langle B^\star \mu \mid v\rangle = \langle Bv \mid \mu\rangle \leqslant 0$, so that $B^\star(\Lambda_+) \subset N$, so that $\overline{B^\star(\Lambda_+)} \subset N$. Now assume that the inclusion is strict: there exists $w \in N$, $w \notin \overline{B^\star(\Lambda_+)}$. By Hahn-Banach separation Theorem, there exists $h \in V$, $\alpha \in \mathbb{R}$ such that

$$\langle h \mid B^\star \mu\rangle \leqslant \alpha < \langle h \mid w\rangle \quad \forall \mu \in \Lambda_+.$$

Since μ goes over a cone, the left hand side inequality implies that $\langle h \mid B^\star \mu\rangle \leqslant 0$ for all $\mu \in \Lambda_+$, so that $\alpha \geqslant 0$ and $h \in C$ by definition of C. We then have $\langle h \mid w\rangle > 0$, which contradicts the fact that $h \in C$, $w \in N = C^o$. □

Let us now introduce the so-called saddle-point formulation of the projection problem

$$\left|\begin{array}{lcl} u + B^\star p & = & U \\ Bu & \in & \Lambda_- \\ p & \in & \Lambda_+ \\ \langle Bu \mid p\rangle & = & 0. \end{array}\right. \tag{5.51}$$

Remarque 20. *The condition $p \in \Lambda_+$ will correspond in actual applications (impact laws) to $p \geqslant 0$. It can be written the same way in the abstract setting, if one considers the partial order associated to the closed convex cone Λ_+ (see e.g. [28]).*

Remarque 21. *The term saddle-point formulation comes from the fact that there is an equivalence between* (5.51) *and $(u, p) \in V \times \Lambda_+$ being a saddle point for the Lagrangian*

$$L(v, q) = \frac{1}{2}|\ v - U|^2 + \langle Bv \mid q\rangle,$$

i.e.

$$L(u, q) \leqslant L(u, p) \leqslant L(v, p) \quad \forall v \in V,\ q \in \Lambda_+.$$

Proposition 14. *Let C be a closed convex cone in V, defined by* (5.50)*, and $U \in V$.*

Let u be the projection of U on C. If the cone $B^\star(\Lambda_+)$ is closed, then there exists $p \in \Lambda_+$ such that (u, p) is a solution to System (5.51)*,*

Conversely, if there exists (u, p) solution to System (5.51)*, then u is the projection of U on C.*

Proof. If $B^\star(\Lambda_+)$ is closed, it identifies with $N = C^\circ$ by Lemma 13, so that there exists $p \in \Lambda_+$ such that $U = u + B^\star p$ by Proposition 13. Since $U = u + B^\star p$ is the decomposition of U over two mutually convex cones (see again Proposition 13), the two terms are orthogonal, i.e. $\langle B^\star p \mid u \rangle = 0$

Conversely, if (u, p) solution to System (5.51), then $u = P_C U$ (and $B^\star p = P_N U$), thanks to Proposition 13. □

Corollaire 1. *Under the assumptions of the previous proposition, if B is* onto*, then Problem* (5.51) *admits a solution (u, p), and it is unique.*

Proof. Uniqueness is straightforward: if B is onto, then $\ker B^\star = B(V)^\perp = \{0\}$, i.e. $B^\star$ is one-to-one, and there exists at most one $p \in \Lambda_+$ such that $U = u + B^\star p$. Since B has a closed range, so does $B^\star$ by the Banach-Steinhaus theorem. As a consequence, there exists $\beta > 0$ such that $|B^\star \mu| \geqslant \beta |\mu|$. Now if a sequence $(B^\star \mu_n)$, with $\mu_n \in \Lambda_+$, converges to $w \in V$, then (μ_n) is a Cauchy sequence by the previous inequality, thus it converges to $\mu \in \Lambda_+$, so that $w = B^\star \mu \in B^\star(\Lambda_+)$. □

Remarque 22. *In the case of a linear space, assuming B has a closed range is enough to ensure that $B^\star(\Lambda)$ is closed (see Proposition 11). In the case of unilateral constraints, a stronger assumption is needed: B has to be assumed* onto *for ensuring the closed character of $B^\star(\Lambda_+)$. Indeed, the image of a closed convex cone by a linear mapping with closed range* is not necessarily closed*, even in the finite dimensional setting. Consider e.g. $\Lambda = \mathbb{R}^3$, and the parabola*

$$\mathcal{P} = \left\{(x, y, z),\, z = 1,\, y = x^2\right\}.$$

Now consider the closed convex cone spanned by this parabola, i.e.

$$\Lambda_+ = \mathrm{conv}\left(\mathbb{R}_+ \mathcal{P} \bigcup \mathbb{R}_+ e_y\right),$$

where e_y is the unit vector in the direction y. The projection of Λ_+ on the (x, y) plane is $\mathbb{R} \times]0, +\infty[\cup \{(0, 0)\}$, which is not closed.

Yet, an important family of cones enjoys the property of being linearly mapped onto a closed set, those are the cones spanned by a finite number of vectors.

Lemme 14. *Let V be a Hilbert space, and N a convex cone spanned by a finite number of vectors:*

$$N = \left\{ \sum_{i=1}^{n} \alpha_i G_i,\, (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n_+ \right\}.$$

Then N is closed, as is its image by any linear mapping.

Proof. We give a full proof of this classical result to enlighten the importance of the fact that N is spanned by a *finite* number of vectors. We prove the result by induction on the number of vectors. For $n = 1$, the result is obvious. Assume that the property is true for $n \geqslant 1$, and consider the cone N associated to $n+1$ vectors. If the $G_i's$ are independent, we call W the space spanned by these vectors, and we introduce

$$G \ : \ \alpha \in \mathbb{R}^{n+1} \longmapsto \sum_{i=1}^{n+1} \alpha_i G_i \in W.$$

This map is invertible, and its reciprocal G^{-1} is linear and continuous from W to $\mathbb{R}^{n+1}$. Now consider $v^k = \sum \alpha^k G_i$ converging to $v \in W$. then $G^{-1}v^k$ converges to $G^{-1}v$, i.e. α^k converges toward α in $\mathbb{R}^{n+1}_+$ by continuity (G^{-1} is a linear mapping between finite dimensional spaces).

Now if the family is linearly dependent, there exists $\mu_1, \ldots, \mu_{n+1}$, not all equal to 0, such that

$$\sum_{i=1}^{n+1} \mu_i G_i = 0. \tag{5.52}$$

We consider a sequence (α^k) in $\mathbb{R}^{n+1}_+$ such that

$$\sum_{i=1}^{n+1} \alpha^k G_i \longrightarrow v.$$

We assume (without loss of generality) that one of the coefficient of (5.52) is negative. We now consider, for any k, the largest $\beta^k \geqslant 0$ such that $\alpha^k + \beta^k \mu_i \geqslant 0$ for $1 \leqslant i \leqslant n+1$. Equality holds for at least one of the indices. Since at least one index i_0 realizes equality an infinite number of times, we extract the corresponding subsequence (without changing the notation). The limit v writes

$$v = \lim \sum_{i \neq i_0} (\alpha_i^k + \beta^k \mu_i) G_i$$

which lies in the cone spanned by the n vectors $(G_i)_{i \neq i_0}$ (by the induction hypothesis), so it is in N. □

We now address some theoretical issues related to the description of solution sets for the pressure $p \in \Lambda$ for equations of the type (5.51). Like in the case of equality constraints (Proposition 12), the solution p is unique a soon as B is onto, and uniqueness is lost whenever the range of B is not dense in Λ. Yet, in the finite dimensional setting, the solution set can be proven to be bounded under some conditions which are typically met for impact laws in granular media. The approach is based on the notion of *asymptotic cone* (see e.g. [14]):

Définition 4. *Let V be a Hibert space, $K \subset V$ a closed convex subset, and $u \in K$. The set*

$$\overrightarrow{K} = \bigcap_{t>0} t(K-u),$$

which does not depend on the choice of $u \in K$, is called the asymptotic cone of K *(see e.g. [42]).*

Proposition 15. *Let V be a Hibert space, and $K \subset V$ a closed convex subset. For any $u \in K$, the asymptotic cone $\overrightarrow{K}$ is the set of directions h such that the half line $u + \mathbb{R}_+ h$ is contained in K.*

Proof. If $u + \mathbb{R}_+ h \subset K$, then h is in $\overrightarrow{K}$ by definition. Conversely, if $h \in \overrightarrow{K}$, h writes $t(v-u)$ for some $t > 0$, with $v = u + h/t \in K$, so that $u + \mathbb{R}_+ h \subset K$. □

This notion provides a criterium to identify bounded convex sets (in the finite dimensional setting).

Proposition 16. *Let V be a finite dimensional Hibert space, and $K \subset V$ a closed convex subset which contains* 0*. Then*

$$K \text{ is bounded} \Longleftrightarrow \overrightarrow{K} = \{0\}.$$

Proof. If K is bounded by M, then tK is bounded by tM, so $\overrightarrow{K}$ contains only 0. Conversely, if K is not bounded, there exists a sequence (u_n) in K, with $|u_n| \to +\infty$. Let u be any element of K. Since V is finite-dimensional the unit sphere is compact, and one can extract a subsequence from $(u_n - u)/|u_n - u|$, which converges to some $v \in H$, with $|v| = 1$. Now consider $t > 0$, and $\theta_n = t/|u_n - u|$. By convexity of K, it holds that

$$(1-\mu_n)u + \mu_n u_n = u + \mu_n(u_n - u) = u + t\frac{u_n - u}{|u_n - u|} \in K.$$

Since K is closed, having n go to infinity yields $u + tv \in K$. As a consequence $\overrightarrow{K}$ contains the nonzero vector v. □

Note that the finite dimension is crucial in the previous proposition. Consider for example the case where $V = \ell^2$ and K is the hypercube $\{x = (x_n) \in V,\, 0 \leqslant x_n \leqslant 1\}$. The closed convex set K does not contain any half-line, while being not bounded.

We may now establish the main property

Proposition 17. *Let V be a finite dimensional Hilbert space, $C \subset V$ a closed convex cone defined by* (5.50)*, $U \in V$, and u the projection of U on C. We assume that $B^\star(\Lambda_+)$ is closed, so that (by Proposition 14) there exists $p \in \Lambda_+$ such that $u + B^\star p = U$. If $\ker B^\star \cap \Lambda_+ = \{0\}$, then the solution set*

$$S = \{q \in \Lambda_+,\, B^\star q = B^\star p = U - u\}$$

is bounded.

Proof. The solution set can be written

$$S = (p + \ker B^\star) \cap \Lambda_+,$$

it is a closed convex set. Consider $h \in \overrightarrow{S}$. By Proposition 15, the half line $p + \mathbb{R}_+ h$ is contained in $S \subset p + \ker B^\star$, which implies $h \in \ker B^\star$. Since S is also contained in the cone Λ_+ is a cone, it also implies that $p/t + h \in \Lambda_+$, for any $t > 0$, which yields, by having t go to 0, $h \in \Lambda_+$. To sum up, $h \in \ker B^\star \cap \Lambda_+ = \{0\}$. We proved that $\overrightarrow{S} = \{0\}$, therefore (by Proposition 16), S is bounded. □

We end this appendix by defining a notion which is relevant to classify problems according to some sort of abstract maximum principle. In the context of collisions, the issue can be formulated as follows: if the pre-collisional velocity fields tends to violate all the constraints, it can be expected that all contacts will be active, i.e. that all interaction forces will be positive, and the unilateral constraints turn out to be equalities. It is an essential tool to exhibit a Poisson like problem for the pressure in impact laws (see the end of Section 5.3). We shall see that this intuitive fact is sometimes ruled out, when a general property is not verified.

Définition 5. *(Abstract Maximum Principle)*
Let C be a closed convex cone in V, associated to $B \in \mathcal{L}(V, \Lambda)$ through Equation (5.50)*. Like in proposition 14, we assume that $B^\star(\Lambda)$ and $B^\star(\Lambda_+)$ are closed, so that, for any $U \in V$, the system* (5.51) *admits at least a solution $(u, p) \in V \times \Lambda_+$, where u is the projection of U on C. We say that the couple (B, Λ_+) (which encodes the structure of the projection problem) verifies the maximum principle if*

$$BU \in -\Lambda_- = -\Lambda_+^\circ \Longrightarrow \exists p \in \Lambda_+ \; s.t. \; BB^\star p = BU.$$

Proposition 18. *If B verifies the abstract maximum principle defined above then, for any U such that $BU \in -\Lambda_-$, there exists a solution (u, p) to* (5.51) *such that*

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

Proof. Let us consider the problem with an equality constraint, i.e. $u \in \ker B$. We denote by u the projection of U on C. From the maximum principle there exists $p \in \lambda_+$ such that $BB^\star p = BU$, which implies that $u = U - B^\star p$ is in $K = \ker B$, so that $u = P_K U$ by Proposition 11. Since $Bu = 0 \in \Lambda_-$ and $p \in \Lambda_+$, the couple (u, p) is *also a solution to the problem with unilateral constraints* (5.51), which ends the proof. □

Chapitre 6

Problème inverse : identification de paramètres

On étudie dans ce chapitre l'identification des paramètres d'un modèle de mouvement de foules afin qu'il prédise des résultats les plus proches possibles d'une vidéo d'un mouvement réel - ou à défaut, d'une vidéo synthétique générée par un modèle dont les paramètres sont cachés. Les objectifs de cette démarche sont multiples. D'une part, la reproduction d'un mouvement réel par un modèle synthétique permet une validation de ce modèle comme pouvant prédire des trajectoires réalistes. D'autre part, en incluant dans le jeu de paramètres l'intensité de certains effets intervenant dans le mouvement (chemoattraction, gradients, adhésion ...), le jeu de paramètre optimal permettrait d'identifier la prédominance d'un phénomène sur un autre - notamment dans le contexte de mouvement de cellules - répondant ainsi à une problématique de sélection de modèle.
Plus précisément, on dispose d'une vidéo que l'on représente par une donnée Z_{obs}, et un modèle que l'on considère comme une boîte noire associant à un jeu de paramètres θ une vidéo Z_θ. On se donne ensuite un coût $\mathcal{L}$ permettant de comparer Z_{obs} et Z_θ. On veut minimiser $\mathcal{L}(Z_{\text{obs}}, Z_\theta)$; l'idée est alors de différencier numériquement cette fonction de θ (qui peut être très compliquée) et d'effectuer une descente de gradient.

Dans ce qui suit, on étudie ce problème d'identification de paramètres pour deux types de modèles de mouvement de foules. On considère d'une part un modèle microscopique de mouvements de cellules où la vitesse dérive d'un potentiel d'interaction entre cellules dit de "Morse" . Ce potentiel est composé d'un terme d'attraction à distance l_c et d'un terme de répulsion à distance l_r. La modèle prend en compte la contrainte de congestion de manière molle, c'est-à-dire que les configurations de superposition sont pénalisées et non interdites. En effet, seul le potentiel répulsif empêche le croisement entre cellules. On veut alors retrouver les longueurs caractéristiques du problème ainsi que l'intensité de l'attraction/répulsion. Le jeu de paramètre θ est un élément de $\mathbb{R}^4$; la faible dimension du

problème permet la convergence d'un schéma de descente de gradient classique.
On considère par ailleurs les modèles macroscopiques de mouvement de foule sous contrainte de congestion à un type tels qu'introduits dans les chapitres 2 et 3. Le paramètre à retrouver est alors le champ de vitesse souhaitée U. Comme U est a priori un élément de $\mathbb{L}^2(\Omega, \mathbb{R}^2)$, on cherche un optimum dans un espace de dimension infinie. Il se pose donc la question de l'approximation de cet espace par un espace de faible dimension pour pouvoir envisager une reconstruction approchée du champ U.
Le travail effectué dans ce chapitre a été réalisé en collaboration avec Gabriel Peyré. L'ensemble du code générant les simulations de ce chapitre est disponible en ligne[1].

6.1 Identification sur un modèle microscopique

On s'intéresse dans cette section au problème d'identification de paramètres pour un modèle microscopique de mouvement de foule, avec une prise en compte de la congestion de manière molle. On suppose que les particules sont soumises à un potentiel d'interaction de Morse, de la forme

$$V(x,y) = -A_c e^{-\frac{\|x-y\|}{l_c}} + A_r e^{-\frac{\|x-y\|}{l_r}}. \tag{6.1}$$

En notant $(X_1, \dots, X_n)$ les positions des particules, le vecteur vitesse de la particule i est donné par

$$\frac{dX_i}{\mathrm{d}t} = -\sum_{j \neq i} \nabla_x V(X_i, X_j). \tag{6.2}$$

Il n'y a ainsi a priori pas d'interdiction forte de la superposition des particules ; en supposant que les particules ont un rayon $r > 0$, certains choix des paramètres A_c, A_r, l_c, l_r peuvent amener à des configurations où des disques se retrouvent à moins de $2r$. Un potentiel de Morse avec des paramètres reproduisant les effets voulus (attraction à longue distance, répulsion à courte distance) est représenté sur la figure 6.1. En particulier, on doit assurer $l_r > l_c$ et choisir $A_r > A_c$ pour avoir répulsion à courte distance. On se donne un jeu de paramètres $\overline{\theta} = (A_r, A_c, l_r, l_c)$ que l'on considère caché et on simule le mouvement en itérant (6.2) par un schéma de différences finies avec le jeu $\overline{\theta}$. On note $Z_{\overline{\theta}}$ le mouvement ainsi obtenu, élément de $(\mathbb{R}^2)^{n \times (\mathrm{n_{it}}+1)}$, en notant $\mathrm{n_{it}}$ le nombre d'itérations du modèle. On veut alors résoudre le problème inverse : étant donné $Z_{\overline{\theta}}$, on cherche à retrouver le vecteur $\overline{\theta}$. Pour ce faire, on se donne une fonction de coût $\mathcal{L}$ qui mesure l'écart entre deux jeux de trajectoires donnés. On cherche alors à minimiser la fonctionnelle f définie par

$$f : \left\{ \begin{array}{ccc} \mathbb{R}^4 & \longrightarrow & \mathbb{R} \\ \theta & \longmapsto & \mathcal{L}(Z_{\overline{\theta}}, Z_\theta). \end{array} \right. \tag{6.3}$$

[1] https://gitlab.math.u-psud.fr/bourdin/crowd-motion-parameter-identification

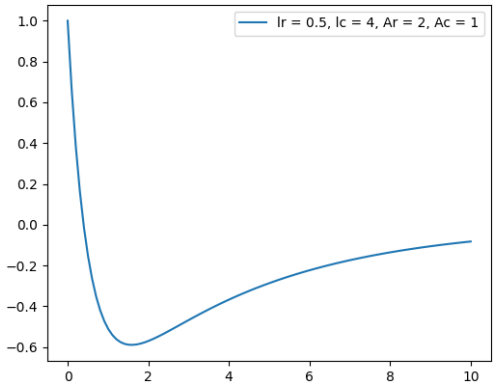


Figure 6.1: Un potentiel de Morse avec des paramètres vérifiant $l_r < l_c$, $A_r > A_c$.

Choix du loss

Le choix du loss $\mathcal{L}$ est primordial dans la résolution du problème. Un choix naturel est

$$\mathcal{L}(Z, W) = \sum_{t=0}^{\mathrm{n_{it}}} \sum_{i=1}^{n} \|Z_i^t - W_i^t\|^2, \tag{6.4}$$

où $\|Z_i^t - W_i^t\|$ est la distance euclidienne entre les positions de la particule i au temps t prédites par les modèles Z et W. Un tel choix est convexe et très facile à calculer. Néanmoins, dans l'optique de traiter des données réelles où l'on n'a pas forcément accès à un suivi des trajectoires et où les particules peuvent être indiscernables, on utilise une distance entre nuages de points, c'est-à-dire invariante par permutation des particules. On choisit alors le loss MMD (Maximum Mean Discrepancies), défini par

$$\mathcal{L}(Z, W) = -\sum_{t=0}^{\mathrm{n_{it}}} \frac{1}{n^2} \sum_{i,j=1}^{n} \|Z_i^t - Z_j^t\| - \frac{1}{n^2} \sum_{i,j=1}^{n} \|W_i^t - W_j^t\| + \frac{2}{n^2} \sum_{i,j=1}^{n} \|Z_i^t - W_j^t\|. \tag{6.5}$$

On vérifie que ce loss est positif et minimal uniquement dans le cas $Z = W$. On renvoie à [11] pour les détails sur le loss MMD ; son avantage est notamment sa grande rapidité de calcul.

Résolution

Afin de minimiser la fonctionnelle f définie par (6.3), on doit calculer son gradient. Un calcul explicite apparaît alors très difficile, étant donné que la dépendance de cette fonction en θ passe par la composition d'un schéma itératif de différences finies. On contourne cette

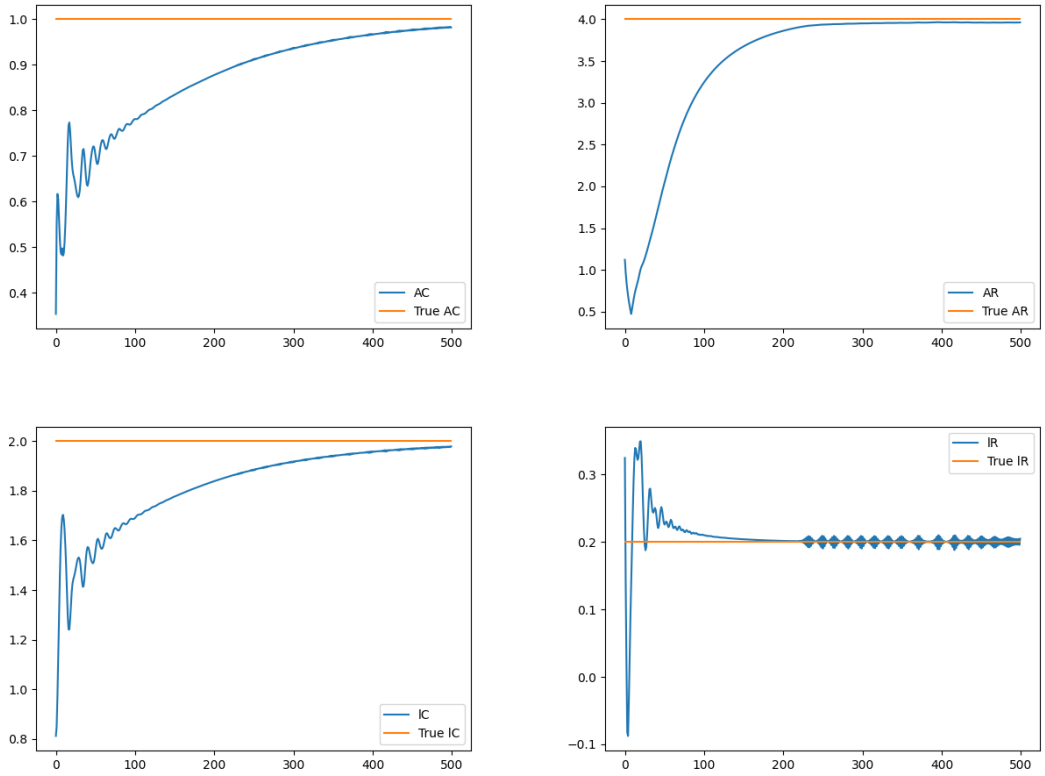


Figure 6.2: L'évolution des paramètres (A_r, A_c, l_r, l_c) en fonction de l'itération lors de la descente de gradient optimale.

difficulté en implémentant le schéma avec la librairie *pytorch* de Python. Cette librairie permet de calculer par backpropagation le gradient d'une quantité dépendant de n'importe quel paramètre en calculant rétrospectivement le gradient de chaque étape du calcul, pourvu que chaque opération soit compatible avec *pytorch*. Le gradient peut de plus être calculé sur GPU, ce qui accélère considérablement le processus.
On représente sur la figure 6.2 une descente de gradient pour des paramètres initiaux $(A_r, A_c, l_r, l_c) = (2, 1, 0.5, 4)$. Le jeu de paramètre converge bien vers $\overline{\theta}$.

6.2 Identification sur un modèle macroscopique

On souhaite à présent identifier les paramètres d'un modèle macroscopique. À l'instar de la section 6.1, on se donne dans un premier temps des données synthétiques générées par le modèle et on cherche à identifier les paramètres du modèle à partir de ces données.

On utilise le modèle macroscopique de mouvement de foule à un type introduit dans le chapitre 2. On se place dans le cas où la vitesse souhaitée est de la forme $U = -\nabla D$. Les paramètres à estimer sont donc le champ D et la densité de saturation $\rho_{\max}$; on suppose cependant la densité de saturation connue dans ce qui suit (on peut par exemple prendre la densité maximale des données observées). On suppose dans la suite le problème discrétisé en temps et en espace, pour $\mathrm{n_{it}}$ itérations et sur un maillage carré de taille N^2. Z_θ est alors un élément de $\mathbb{R}^{N^2 \times (\mathrm{n_{it}}+1)}$ et représente la densité de population dans l'espace au cours du temps.

Choix du loss

Un choix de loss intuitif est de choisir un loss MMD comme dans la section précédente. Cependant, ce loss ne rend pas bien compte des translations. En particulier, le loss entre une mesure et sa translation d'un facteur τ est génériquement inférieur à τ. De même la norme $\mathbb{L}^2$ n'est pas sensible aux translations et aux grandes variations de densité sur des petites échelles d'espace, comme illustré dans l'exemple représenté sur la figure 6.3. Dans cet exemple, on part d'une densité initiale présentant de grandes variations en espace que l'on transporte par une vitesse constante $u = 0.5$. On peut calculer explicitement son image en $t > 0$ - il s'agit de la densité translatée d'un facteur tu. On se donne ensuite une erreur $\epsilon \in [-0.5, 0.5]$ sur la vitesse. On compare la densité exacte ρ^t transportée par u avec une densité ρ_ϵ^t transportée par $u_\epsilon = u + \epsilon$. Étant donné un pas de temps $dt > 0$, on définit

$$\begin{aligned} \mathcal{L}_{\mathbb{W}}(\epsilon) &= \sum_{k=0}^{\lfloor \frac{T}{dt} \rfloor} W_2^2(\rho_\epsilon^{kdt}, \rho^{kdt}) \\ \mathcal{L}_{\mathbb{L}_2}(\epsilon) &= \sum_{k=0}^{\lfloor \frac{T}{dt} \rfloor} \left\| \rho_\epsilon^{kdt} - \rho^{kdt} \right\|_2^2. \end{aligned} \tag{6.6}$$

On voit en particulier que le loss $\mathbb{L}_2$ n'est pas convexe, rendant possible la convergence d'une descente de gradient vers un mauvais paramètre lorsqu'on cherche à retrouver $\epsilon = 0$ à partir de la trajectoire ρ^t.

Cet exemple encourage donc à choisir un loss Wasserstein. En particulier, la distance de Wasserstein entre une mesure et sa translation d'un facteur τ est exactement τ. Néanmoins, une distance de Wasserstein entre deux nuages de points est un problème linéaire d'optimisation et donc coûteux en calcul. Une solution est alors de trouver un compromis entre ces possibilités. Comme démontré dans [30], le transport optimal régularisé constitue une interpolation entre le transport optimal (quand le paramètre de régularisation tend vers 0) et le loss MMD (quand le paramètre de régularisation tend vers $+\infty$). De plus, il se calcule très rapidement à l'aide de l'algorithme de Sinkhorn.

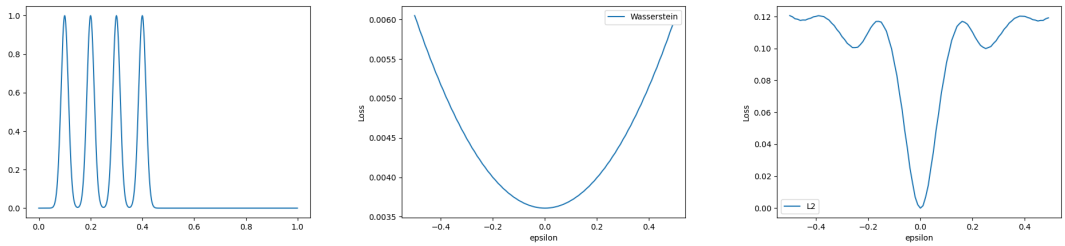


Figure 6.3: À gauche, la densité initiale que l'on transporte selon une vitesse $u = 0.5$ pendant un temps total $T = 0.5$. Au milieu, le loss Wasserstein entre la vraie trajectoire et celle obtenue lorsqu'on perturbe la vitesse en appliquant $u_\epsilon = 0.5 + \epsilon$ pour $\epsilon \in [-0.5, 0.5]$. À droite, le loss L2 qui n'est en particulier pas convexe.

Schéma de splitting

On choisit dans un premier temps d'utiliser le schéma développé dans le chapitre 3. Comme ce chapitre fait intervenir une étape stochastique qui comprend une marche aléatoire que l'on moyenne sur un grand nombre d'itérations et que le calcul du gradient avec *pytorch* nécessite la backpropagation à travers chaque étape de calcul, il apparaît difficile de garder en mémoire chaque pas de ces marches aléatoires afin de différencier à travers l'étape de projection. Rappelons l'heuristique introduite dans [46] pour obtenir cet algorithme de projection stochastique. Lorsque la densité projetée est de la forme $\mu = 1 + \epsilon\rho$, le plan de transport entre μ et sa projection sur l'ensemble des densités admissibles est au premier ordre en ϵ de la forme $T = \mathrm{Id} + \epsilon u$ où

$$u = -\nabla(\Delta^{-1}\rho). \tag{6.7}$$

On modifie donc le schéma en résolvant à chaque étape l'équation (6.7), c'est-à-dire en inversant l'opérateur de Laplacien sur le maillage. La densité obtenue n'est pas nécessairement admissible à chaque étape, mais en prenant des petits pas de temps les trajectoires sont satisfaisantes sur des cas jouets.

On teste la procédure d'identification de paramètres sur un modèle simple en dimension 1. On se donne une densité initiale sur $\Omega = [0, 1]$

$$\rho^0(x) = \sin(6\pi x)^2, \tag{6.8}$$

et un champ de vitesses $U(x) = x(1 - x)$, sous contrainte de densité maximale $\rho_{\max} = 1$. Le paramètre à retrouver est donc $\overline{\theta} = U$. On discrétise l'espace en un maillage de taille $N = 200$, et on simule le transport sous contraintes de congestion pendant 50 pas de temps $dt = 0.02$. Le résultat est alors négatif : la descente de gradient échoue à retrouver $\overline{\theta}$, même en introduisant une régularisation $\mathbb{H}^1$ du champ U afin d'abaisser la dimension du problème.

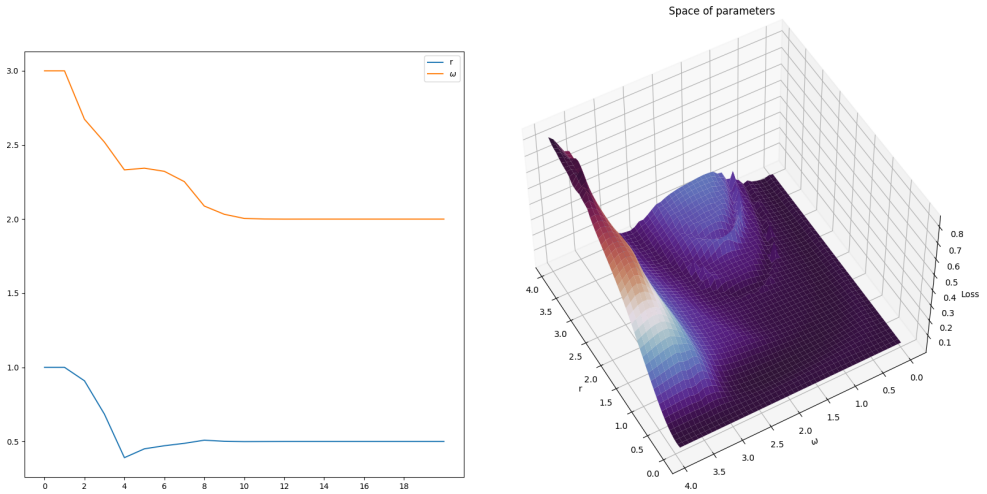


Figure 6.4: À gauche, une descente de gradient partant de $(1, 3)$ et convergeant vers $\overline{\theta}$. À droite, l'espace des paramètres du problème d'identification de paramètres pour le modèle macroscopique $1D$ avec une vitesse paramétrique de la forme (6.9). Les paramètres exacts à retrouver sont $(r, \omega) = (0.5, 2)$. On vérifie que le loss est bien minimal en $\overline{\theta}$.

La descente de gradient est effectuée dans un espace de paramètres de dimension trop élevée. Finissons cette section avec un exemple d'une descente de gradient qui converge dans un espace de dimension 2, lorsqu'on a paramétrisé la vitesse.

Exemple 5. *On se donne un champ dépendant de deux paramètres*

$$U_{r,\omega}(x) = r\sin(\omega x). \tag{6.9}$$

Les paramètres à retrouver sont alors $\overline{\theta} = (r, \omega)$*. Dans ce formalisme, la descente de gradient converge, comme représenté sur la figure 6.4 où l'on a calculé le loss pour tous les couples* $(r, \omega) \in [0, 4]^2$*. La fonction à minimiser présente un minimum global en* $\overline{\theta}$*. En réalisant en parallèle plusieurs descentes de gradient partant de conditions initiales aléatoires et en sélectionnant celle qui converge vers le loss le plus petit, on parvient à retrouver* $\overline{\theta}$*.*

JKO, paramétrisation par un réseau de neurones

On souhaite s'inspirer de l'exemple 5 pour pouvoir approcher un champ de vitesses quelconque par un champ paramétrique. Si l'exemple traite un cas trop simple pour pouvoir approcher n'importe quel champ, on a vu qu'une méthode non paramétrique échouait à converger pour le modèle macroscopique. Une solution est alors d'utiliser un réseau de neurone, vu comme une fonction dont les paramètres sont les poids du réseau. On augmente alors la dimension des l'espace des paramètres sans que cette dimension dépende de la

taille du maillage. Conformément à ce qui est fait dans [15], on se donne une vitesse U inconnue que l'on approche par un champ U_θ de la forme $U_\theta = -\nabla D_\theta$, où D_θ est un réseau de neurones "ICNN" (pour Input Convex Neural Network), c'est-à-dire ici

$$D_\theta(x) = \sum_{k=1}^{M} c_k h(a_k x + b_k), \tag{6.10}$$

où h est la fonction partie positive (ou ReLU dans le contexte du machine learning), M est le nombre de neurones de la couche et $\theta = (a_k, b_k, c_k)_{k=1,\dots,M}$ le jeu de paramètres. À l'instar du modèle présent dans [15], on utilise le modèle JKO à un type développé dans le chapitre 2. On se place dans le cas bidimensionnel avec $\Omega = [0,1]^2$, maillé en N^2 cases. On se donne un champ scalaire convexe D, un pas de temps $dt > 0$ et un nombre d'itérations $\mathrm{n_{it}}$. On simule alors le schéma JKO régularisé présenté dans la section 2.4 : on obtient une observation $(\rho^{kdt})_{k=0,\dots,\mathrm{n_{it}}} \in \mathbb{R}^{(\mathrm{n_{it}}+1)\times N^2}$. D'autre part, étant donné un jeu de paramètres θ, on simule le schéma JKO associé à D_θ, pour obtenir $(\rho_\theta^{kdt})_{k=0,\dots,\mathrm{n_{it}}}$. On définit alors le loss

$$\mathcal{L}_{\mathbb{W},\epsilon}(\theta) = \sum_{k=0}^{M} W_2^\epsilon(\rho^{kdt}, \rho_\theta^{kdt}), \tag{6.11}$$

où W_2^ϵ est la distance de Wasserstein régularisée introduite dans la section 2.4. On implémente le modèle et le calcul du loss avec la librairie *pytorch* de Python. En particulier, on peut calculer par backpropagation le gradient de $\mathcal{L}_{\mathbb{W},\epsilon}$ et effectuer une descente de gradient afin de chercher le θ_{opt} optimal.

Remarque 23. *Il n'y a pas forcément de paramètre optimal $\overline{\theta}$ ici. En particulier, le loss optimal n'est pas forcément nul, ni forcément atteint.*

On représente sur la figure 6.5 l'apprentissage du potentiel

$$D(x,y) = 10\left(\left(x - \frac{1}{2}\right)^2 + \left(y - \frac{1}{2}\right)^2\right). \tag{6.12}$$

On part d'une densité initiale $\rho^0(x,y) = (1-x)(1-y)$ et on simule son mouvement sous le flot gradient du potentiel D en appliquant un schéma JKO régularisé, comme défini en section 2.4. On voit sur la figure 6.5 que le réseau de neurones approche fidèlement le potentiel D et que les densités soumises à D et à $D_{\theta_{\mathrm{opt}}}$ ont un comportement similaire. En particulier, le potentiel estimé a la même forme que le potentiel exact (même s'il diffère car il est défini à une constante près).

6.3 Perspectives : vers un apprentissage sur données réelles

On présente comme perspective d'application une tentative d'apprentissage sur données réelles. On utilise les données de Pedestrian Dynamics Data Archive[2], site qui met en ligne

[2]Voir https://ped.fz-juelich.de/da/doku.php

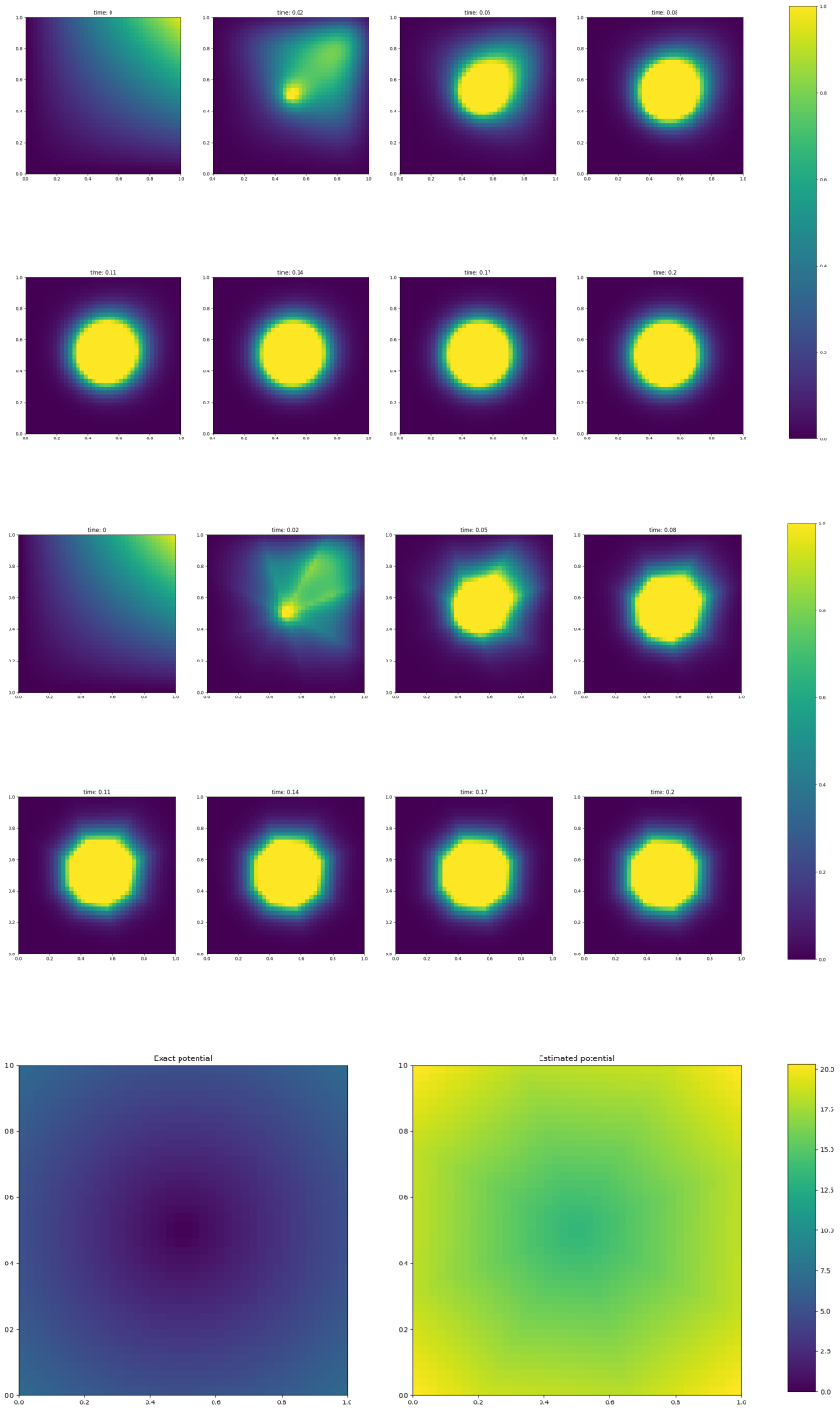


Figure 6.5: En haut, le mouvement de la densité soumise au champ $-\nabla D$, au milieu la densité soumise au champ estimé $-\nabla D_{\theta_{\text{opt}}}$ après descente de gradient sur θ. En bas à gauche, le potentiel exact, en bas à droite, le potentiel estimé par la descente de gradient.

un volume conséquent d'expériences de croisements de foules, d'évacuation, de parcours de piétons ... Pour chaque expérience, on a accès à une vidéo du mouvement mais également à l'ensemble des trajectoires dans la population. On commence par transformer ces trajectoires en densités, en remplaçant chaque individu (ponctuel) par un disque de densité fixé. On obtient ainsi la vidéo de l'évolution d'une densité dans le temps.
On utilise l'expérience "Crowds in front of bottlenecks from the perspective of physics and social psychology" représentée sur la figure 6.6 : une foule dense répartie sur une zone bidimensionnelle tente de sortir par une porte qui force les gens à passer chacun·e son tour. On souhaite donc lancer l'identification de paramètres sur le modèle JKO de la partie 6.2, où l'on a paramétré le champ de potentiel par un réseau de neurones ICNN. Dans le cas où l'algorithme converge, on s'attend à observer un potentiel minimal sur la porte (typiquement la distance à cette porte).

Remarque 24. *En pratique les vidéos de mouvement de foules ne sont pas à effectif constant : des individus peuvent sortir du champ de la caméra. On contourne ici cette difficulté ajoutant la possibilité de perdre de la masse aux bords du domaine. Une autre solution pourrait être d'utiliser un transport optimal "unbalanced", outil qui permet de transporter des mesures à masse variable.*

On n'obtient malheureusement pas la convergence du potentiel estimé vers un potentiel de cette forme-là. En outre, lors de l'expérience le loss ne tend pas vers 0. Plusieurs facteurs peuvent expliquer cela et sont à étudier dans le but d'estimer des champs de vecteurs gouvernant des mouvements réels :

- Le réseau de neurone (ici à 20 neurones) n'est pas assez complexe pour décrire le phénomène en œuvre.
- Le mouvement de foule ne peut être décrit par une attraction de la forme $U = -\nabla D$ où D est un potentiel.
- On n'a pas trouvé le loss optimal ; une autre méthode de descente de gradient peut être envisagée pour trouver le potentiel optimal.
- L'effectif non constant rend le mouvement plus difficile à analyser.

Figure 6.6: Une expérience d'évacuation. La foule sort par la porte de droite au compte-goutte. Source : https://ped.fz-juelich.de/da/doku.php.

Partie II

Modèles SIR condensés

Chapitre 7

Modèle SIR sur un graphe de communautés

On développe dans ce chapitre un modèle déterministe de propagation d'épidémie dans un établissement scolaire. Le modèle présente une granularité variable ; les quantités épidémiologiques (S, I et R) représentent les fractions saine, infectée ou remise d'un groupe de personnes. On compare le modèle à sa version probabiliste et on introduit des notions d'analyse qualitative permettant de repérer les populations à risque au sein de l'établissement.

7.1 Modèle SIR par communautés

On se donne une population V, organisée en un ensemble de particules $\mathcal{P}$ (des classes, des sous-ensembles de classes ou des singletons pour les professeurs). Plus précisément, $\mathcal{P}$ est une partition de V en parties typiques de cardinal 1 pour les professeurs, ou d'ordre 20 pour les classes, voire d'ordre plus petit pour les parties de classes. En effet, l'hypothèse sous-jacente à l'élaboration du modèle est qu'un ensemble d'individus restant toujours groupé peut être décrit par ses ratios de susceptibles et d'infectés. Conformément à la littérature en épidémiologie, on considère que chaque individu peut se trouver dans l'état S (sain·e), I (infecté·e) ou R (remis·e). On suppose donc qu'on peut attribuer à chaque particule X des proportions S_X, I_X et R_X entre 0 et 1.

Remarque 25. *Deux points de vue sont possibles ici, l'un déterministe et l'autre probabiliste. On peut considérer que I_X décrit la proportion d'infecté·e·s dans X, donc que le nombre d'infecté·e·s est $I_X N_X$ où N_X est l'effectif de la particule X. D'autre part, on peut supposer que chaque élément de X a indépendamment une probabilité I_X d'être infecté, et que l'on modélise alors l'évolution de cette loi de probabilité. On retrouve en particulier dans ce cas en moyenne $I_X N_X$ infectés, mais en fonction de la réalisation le nombre réel d'infecté·e·s peut varier. On reviendra plus précisément sur ces nuances dans la section 7.2.*

Contacts directs

On suppose qu'on connaît chaque jour pour chaque paire de particule K_{XY} le nombre total de contacts entre éléments de X et de Y. On suppose alors que chaque élément de X a $\overline{K_{XY}} := \frac{K_{XY}}{N_X}$ contacts avec des éléments de Y.

Remarque 26. *On peut considérer que $\overline{K_{XY}}$ encode le nombre moyen d'interactions entre éléments de X et Y au sein de X, et qu'on lisse les particularités en remplaçant pour chaque paire de $X \times Y$ le vrai nombre de contacts par la valeur moyenne. Cette procédure peut être particulièrement pertinente dans le cas d'une grande population où l'on ne peut lister l'ensemble des contacts, mais où l'on peut uniquement inférer ces nombres de contacts moyens entre groupes de personnes.*

Pour un contact entre les particules X et Y, la probabilité d'infection pour un individu de X dans l'état S est donc, en supposant que les K personnes avec lesquelles il sera en contact dans Y sont tirées uniformément et qu'il y a une proportion I_Y d'infecté·e·s dans Y :

$$1 - (1 - pI_Y)^{\overline{K_{XY}}}. \tag{7.1}$$

Remarque 27. *On a fait l'hypothèse implicite que chaque individu de X tirait uniformément $\overline{K_{XY}}$ personnes avec lesquelles il était en contact. Si on examine le graphe biparti obtenu, le degré de chaque nœud de X est constant, mais celui de Y peut varier : les classes X et Y ne jouent pas un rôle symétrique. Une manière de contourner ce problème est de tirer uniformément un graphe parmi l'ensemble des graphes bipartis K-réguliers, c'est-à-dire les graphes dont chaque arête relie un nœud de X et de Y, et de degré constant K. On retrouve pour chaque nœud de X la formule* (7.1).

Contacts indirects

On considère possible la transmission de l'infection par voie indirecte, par l'émission d'un dépôt (surfacique, ou aérien). On appelle alors contact indirect entre deux personnes la succession de ces deux personnes au même endroit. On suppose qu'on dispose d'une liste des événements de contact indirect entre particules. Plus précisément, on a pour chaque événement :

- Les deux classes en contact indirect,
- La probabilité p de transmission de l'épidémie à une personne de la deuxième particule due au dépôt d'une personne infectée dans la première particule.

Cette probabilité p est d'autant plus élevée que le contact indirect est rapproché dans le temps, et que les particules sont restées longtemps dans le lieu du contact. On peut ainsi

écrire la probabilité d'infection d'un individu de la classe X lors d'un événement de contact indirect avec Y :

$$1 - (1-p)^{I_Y N_Y} . \tag{7.2}$$

Remarque 28. *On a fait l'hypothèse implicite qu'une fraction I_Y de la classe Y était infectée, ce qui correspond au point de vue déterministe de la remarque 25. En effet, si l'on adopte le point de vue probabiliste, on considère qu'il y a un nombre variable d'infecté suivant une binomiale de paramètres N_Y, I_Y. La probabilité d'infection lors du contact est alors :*

$$1 - \sum_{k=0}^{N_Y} \binom{N_Y}{k} I_Y^k (1-I_Y)^{N_Y - k} (1-p)^k. \tag{7.3}$$

Les deux expressions sont néanmoins équivalentes pour $p \ll \dfrac{1}{N_Y}$.

Pour deux particules $X, Y \in \mathcal{P}$, on note e_{XY} leur nombre d'événements de contacts indirects, de probabilités $p_1, \ldots, p_{e_{XY}}$. On note

$$p_{XY} = 1 - \prod_{k=1}^{e_{XY}} (1 - p_k) \tag{7.4}$$

la probabilité pour un individu de X de se faire infecter s'il y a un·e unique infecté·e dans la particule Y au cours des e_{XY} événements de contact indirect. S'il y a $N_Y I_Y$ infecté·e·s dans Y, la probabilité d'infection au cours de la journée est donc

$$1 - (1 - p_{XY})^{N_Y I_Y} \tag{7.5}$$

pour chaque individu de X.

Modèle SIR

On décrit à présent la dynamique épidémiologique du modèle. On fait l'hypothèse que les changements de catégorie S, I, R se font à l'issue de la journée. Pour écrire simplement le modèle, on fait les hypothèses suivantes :

- La probabilité d'infection lors d'un contact direct est constante au sein de toute la population,
- La matrice de contact $(K_{XY})_{X,Y \in \mathcal{P}}$ est la même chaque jour,
- Les événements indirects sont les mêmes chaque jour.

En faisant le bilan des expressions (7.1) et (7.2), et en supposant que le passage de I à R se fait à taux γ, on obtient le système suivant :

$$\begin{cases} S_X^{n+1} &= S_X^n \prod_{Y\in\mathcal{P}} (1-pI_Y^n)^{\overline{K_{XY}}} \prod_{Y\in\mathcal{P}} (1-p_{XY})^{I_Y N_Y} \\ I_X^{n+1} &= S_X^n \left(1 - \prod_{Y\in\mathcal{P}} (1-pI_Y^n)^{\overline{K_{XY}}} \prod_{Y\in\mathcal{P}} (1-p_{XY})^{I_Y N_Y}\right) + (1-\gamma)\, I_X^n \\ R_X^{n+1} &= R_X^n + \gamma I_X^n. \end{cases} \tag{7.6}$$

Le modèle s'étend directement à des contacts directs non homogènes en remplaçant le terme $(1-I_Y^n p)^{\overline{K_{XY}}}$ par un produit de $\overline{K_{XY}}$ termes de la forme $(1-p_j^{XY} I_Y^n)_{j=1,\dots,\overline{K_{XY}}}$, avec des probabilités p_j^{XY} variant en fonction du contact et des particules considérées. En revanche, chaque individu de la particule X doit avoir une liste de contacts ayant les mêmes propriétés ($\overline{K_{XY}}$ contacts avec des éléments de Y, de probabilités p_j^{XY}). De même, on peut considérer un modèle dont la transition varie en fonction du jour, en définissant des probabilités et une matrice de contacts qui dépendent du jour.

Remarque 29. *On a fait l'hypothèse implicite que chaque individu de X avait lors d'un contact une probabilité pI_Y^n de se faire infecter, puis on a remplacé à chaque pas de temps la loi d'infection des individus d'une particule par un seul nombre qui correspond à nouveau à la proportion d'infecté·e·s. On fait ensuite pour chaque pas de temps repartir le modèle SIR d'une proportion d'infecté·e·s correspondante. Or cette loi peut varier en fonction de l'hypothèse sous-jacente que l'on a formulée sur la distribution initiale au sein des classes lors de la remarque 25 (déterministe, ou aléatoire indépendante). On remplace donc à chaque pas les probabilités d'infections au sein d'une classe par une version lissée et uniformisée. Il faut cependant garder en tête que le modèle décrit en* (7.6) *ne correspond pas à l'évolution des probabilités moyennes d'infection d'un modèle purement probabiliste à l'échelle microscopique, la non-linéarité du modèle entraînant une divergence à chaque étape lorsqu'on remplace la loi de l'infection dans un groupe par sa moyenne.*

7.2 Comparaison avec des modèles microscopiques probabilistes

On va à présent étudier les conditions pour que le modèle SIR (7.6) décrive l'évolution de la loi de probabilité de l'infection dans un modèle aléatoire défini à l'échelle de l'individu. On cherche donc à définir des variables aléatoires $(E_x)_{x\in V}$ prenant pour valeur S, I ou R dont la loi ne dépende pas du représentant $x \in X$ choisi au sein d'une particule X.
On veut définir une application

$$\psi \ : \mathbb{P}\left(\{\mathrm{S},\mathrm{I},\mathrm{R}\}^Y\right) \longrightarrow \mathbb{P}\left(\{\mathrm{S},\mathrm{I},\mathrm{R}\}^X\right) \tag{7.7}$$

respectueuse du modèle décrit par (7.6) qui à une mesure de probabilité d'infection sur la particule Y associe une mesure de probabilité d'infection sur la particule X après contacts. On veut également partir d'une loi sur $\{\mathrm{S},\mathrm{I},\mathrm{R}\}^Y$ d'une forme donnée et retrouver la même forme après application de ψ (par exemple, partir d'un N_Y-uplet de variables de Bernouilli indépendantes et obtenir un N_X-uplet de variables de Bernouilli indépendantes).
Pour simplifier, on analyse le cas d'un unique événement de contact entre deux classes X et Y, et on suppose $\gamma = 0$. On n'a donc pas de compartiment R.

Tentative 1 : contacts indépendants par personne.

On part de la remarque suivante : l'équation (7.6) est vérifiée dès que chaque individu a $\overline{K_{XY}}$ contacts avec des individus infectés de manière indépendante avec probabilité I_Y^n. On cherche donc à construire un modèle de contacts tel que le nombre de contacts est constant, et que la probabilité d'infection en fonction des variables S^n, I^n est donnée par la formule (7.6).
On suppose donc les individus infectés au jour n selon des lois de Bernouilli indépendantes de paramètre I_X^n ou I_Y^n, et que lors de l'événement de contact entre X et Y, chaque individu de X tire uniformément et indépendamment $\overline{K_{XY}}$ liens de contacts dans Y.

Points positifs. On vérifie que les marginales de la loi d'infection de la particule X au jour $n+1$ sont toutes des Bernouilli de paramètre $S_X^n\left(1-(1-I_Y^n p)^{\overline{K_{XY}}}\right)$. Même si pour Y la formule (7.6) n'est pas exctement vérifiée, l'infection de chaque individu de la classe Y suit une loi de Bernouilli de paramètre

$$S_Y^n\left(1-\left(1-\frac{I_X^n p}{N_Y}\right)^{N_X\overline{K_{XY}}}\right) \tag{7.8}$$

où N_X et N_Y sont les effectifs des classes. Dans l'asymptotique $p \ll 1$, on retrouve bien la bonne formule avec $K_{YX} = \overline{K_{XY}}\dfrac{N_X}{N_Y}$ si les classes ont un effectif différent.

Points négatifs. Le type de loi pour l'infection initiale étant un N_Y-uplet de Bernouilli indépendantes, l'infection des individus de X après contact n'est un N_X-uplet de Bernouilli indépendantes qu'uniquement conditionnellement à la réalisation des $(I_y)_{y\in Y}$ (ou seulement conditionnellement à leur somme). Dans le cas général, les S_x ne sont pas des variables indépendantes : l'hypothèse de Bernouilli indépendantes avant contact ne se propage pas.

Tentative 2 : contacts indépendants par personne, distribution déterministe.

On a vu lors de la première tentative que lorsque le nombre total d'infecté·e·s variait, la distribution des infections de sortie ne se décomposait pas selon un N_Y-uplet de variables

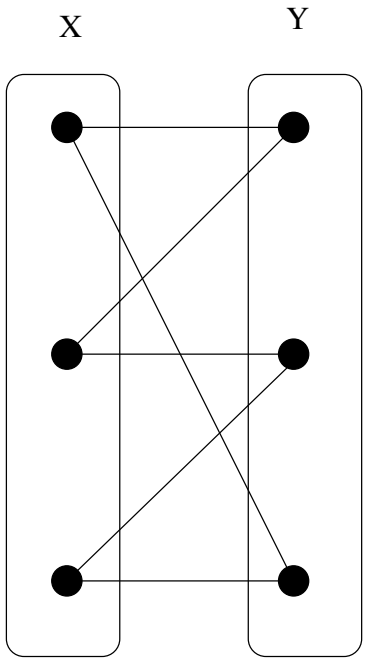


Figure 7.1: Un graphe biparti 2-régulier entre deux particules de 3 éléments : chaque nœud est de degré 2, et chaque arête relie un élément de X à un élément de Y.

indépendantes. On suppose donc à présent qu'il y a $N_X I_X, N_Y I_Y$ infecté·e·s tiré·e·s au hasard dans les classes X et Y (en supposant que ces nombres sont entiers). On tire ensuite comme précédemment le graphe de contacts.

Points positifs. La formule (7.6) est à nouveau valable dans la limite $p \ll 1$ lorsque les effectifs sont différents. On a bien à présent des infections indépendantes au sein d'une classe, avec probabilité I_X, I_Y.

Points négatifs. La loi d'infection après contact est un N_X, N_Y-uplet de variables de Bernouilli indépendantes et non $N_X I_X, N_Y I_Y$ infecté·e·s tiré·e·s aléatoirement dans X et Y. On est donc dans le cadre de la tentative 1, et on se trouvera face aux mêmes problèmes que précédemment au deuxième pas du schéma.

Tentative 3 : graphe K-uniforme

On analyse ici en détail le cas où chaque personne a $\overline{K_{XY}}$ contacts lors de l'événement de contact, qu'elle soit dans la classe X ou la classe Y. Cela n'est possible que dans le cas où les classes ont le même effectif. On suppose donc que les classes sont d'effectif N et l'on note $(Y_1, \dots, Y_N)$, l'état des N personnes du groupe Y qui suit une loi $\mu \in \mathbb{P}\left(\{0,1\}^N\right)$, avec la convention "0 = sain·e", "1 = infecté·e".

Définition 1. *On définit $\mathbb{G}_{\overline{K_{XY}}}(X,Y)$ l'ensemble des graphes bipartis entre X et Y, et $\overline{K_{XY}}$-réguliers, c'est-à-dire les graphes dont chaque arête relie un nœud de X et de Y, et dont chaque nœud est de degré $\overline{K_{XY}}$.*

Un exemple d'un tel graphe est représenté sur la figure 7.1. On tire ainsi un graphe G uniformément dans $\mathbb{G}_{\overline{K_{XY}}}(X,Y)$, puis on tire conditionnellement à $(Y_1,\dots,Y_N)$ et à G les variables $(X_1,\dots,X_N)$, qui suivent indépendamment des lois de Bernouilli définies par

$$\mathbb{P}(X_i=1) = 1-\prod_{i\sim j}(1-pY_j) \tag{7.9}$$

où $i\sim j$ signifie que i et j sont reliés par une arête de G.

Exemple 6. *Si l'on suppose qu'il y a un·e infecté·e introduit dans la classe Y, on aura $\mu=\delta_{(1,0,\dots,0)}$. Cela correspond au point de vue déterministe de la remarque 25. En revanche, si l'on suppose que chaque individu de la classe Y a une probabilité indépendante $\frac{1}{N}$ d'être infecté, on a $\mu=\mu_1\otimes\cdots\otimes\mu_N$, où μ_i est une loi de Bernouilli de paramètre $\frac{1}{N}$.*

On a donc construit une application $\psi : \mathbb{P}\left(\{0,1\}^Y\right) \longrightarrow \mathbb{P}\left(\{0,1\}^X\right)$, qui a une mesure de probabilité d'infection sur la classe Y associe une mesure de probabilité d'infection sur la classe X. On peut exprimer précisément la loi de $\psi(\mu)$:

$$\mathbb{P}\left(X_1=\alpha_1,\dots,X_N=\alpha_N\right) = \mathbb{E}\left[\prod_{i=1}^{N}\left(1-\prod_{i\sim j}(1-pY_j)\right)^{\alpha_i}\left(\prod_{i\sim j}(1-pY_j)\right)^{1-\alpha_i}\right], \tag{7.10}$$

avec pour convention $0^0=1$ pour inclure le cas d'une Bernouilli de paramètre 0.
Comparons les images des deux descriptions possibles de la condition initiale "un·e infecté·e dans la classe Y", à savoir

$$\begin{aligned}\mu &= \delta_{(1,0,\dots,0)}\\ \nu &= \mu_1\otimes\cdots\otimes\mu_N,\end{aligned} \tag{7.11}$$

où les $(\mu_j)_j$ suivent des lois de Bernouilli indépendantes de paramètre $\frac{1}{N}$.

Proposition 8. *$\psi(\mu) = p\left(\delta_{a_1}+\cdots+\delta_{a_{\overline{K_{XY}}}}\right)$, où les a_i sont tirés uniformément sans remise dans X.*

Démonstration. On reprend la formule (7.10), on a :

$$\begin{aligned}\mathbb{P}\left(X_1=\alpha_1,\dots,X_N=\alpha_N\right) &= \mathbb{E}\left[\prod_{i=1}^{N}\left(1-(1-p)\mathbf{1}_{i\sim 1}\right)^{\alpha_i}\left((1-p)\mathbf{1}_{i\sim 1}\right)\right]\\ &= \mathbb{E}\left[\prod_{i=1}^{N} p^{\alpha_i}(1-p)^{1-\alpha_i}\mathbf{1}_{i\sim 1}+\mathbf{1}_{\neg i\sim 1}\mathbf{1}_{\alpha_i=0}\right]\end{aligned} \tag{7.12}$$

Par invariance de $\mathbb{G}_{\overline{K_{XY}}}$ par permutation des éléments de X, on remarque que les $\overline{K_{XY}}$ éléments de X auxquels l'élément 1 de Y sera apparié suivent une loi uniforme sur l'ensemble des parties à $\overline{K_{XY}}$ éléments de X. Le produit précédent vaut donc :

$$\mathbb{P}\left(X_1=\alpha_1,\dots,X_N=\alpha_N\right) = \mathbb{E}\left[\prod_{i=1}^{\overline{K_{XY}}} p^{\alpha_{a_i}}\left(1-p\right)^{1-\alpha_{a_i}}\mathbf{1}_{\alpha_i=0,\forall i\notin\{a_i\}_{i=1,\dots,\overline{K_{XY}}}}\right], \tag{7.13}$$

ce qui est l'expression de la loi de $p\left(\delta_{a_1}+\cdots+\delta_{a_{\overline{K_{XY}}}}\right)$. □

On remarque en particulier que le nombre d'infecté·e·s de X après contact est entre 0 et $\overline{K_{XY}}$. Étudions à présent l'image de la mesure ν (on sait déjà qu'elle est différente de $\psi(\mu)$, car le nombre d'infecté·e·s peut aller de 0 à N). On peut décrire la loi de ses marginales :

Proposition 9. *Soit $(X_1,\dots,X_N)$ suivant une loi $\psi(\nu)$. Alors, pour tout $1\leqslant i\leqslant N$, X_i suit une loi de Bernouilli de paramètre $1-\left(1-\frac{p}{N}\right)^{\overline{K_{XY}}}$.*

Démonstration.

$$\begin{aligned}\mathbb{P}\left(X_i=1\right) &= \mathbb{E}\left[\mathbf{1}_{X_i=1}\right]\\ &= \mathbb{E}\left[\mathbb{E}\left[\mathbf{1}_{X_i=1}|(Y_j)_j,G\right]\right]\\ &= \mathbb{E}\left[\mathbb{E}\left[1-\prod_{i\sim j}\left(1-pY_j\right)|(Y_j)_j,G\right]\right]\\ &= \mathbb{E}\left[\mathbb{E}\left[1-\prod_{i\sim j}\left(1-pY_j\right)|G\right]\right]\end{aligned} \tag{7.14}$$

Or, les Y_j sont indépendants entre eux et indépendants de G. On peut donc écrire :

$$\begin{aligned}\mathbb{P}\left(X_i=1\right) &= \mathbb{E}_G\left[1-\prod_{i\sim j}\left(1-p\mathbb{E}[Y_j]\right)\right]\\ &= \mathbb{E}_G\left[1-\prod_{i\sim j}\left(1-\frac{p}{N}\right)\right].\end{aligned} \tag{7.15}$$

□

Remarque 30. *On a en général, pour ν de la forme $\mu_1 \otimes \cdots \otimes \mu_N$, avec $\mu_j \sim Ber(\alpha_j)$,*

$$X_i \sim Ber\left(\mathbb{E}_G\left[1 - \prod_{i\sim j}(1-p\alpha_j)\right]\right). \tag{7.16}$$

Les cas précédents correspondent aux cas extrêmes $(\alpha_1, \ldots, \alpha_N) = (1, 0, \ldots, 0)$ et $(\alpha_1, \ldots, \alpha_N) = \left(\frac{1}{N}, \ldots, \frac{1}{N}\right)$. On déduit de la proposition 8 que pour $(X_1, \ldots, X_N)$ de loi $\psi(\mu)$, X_i suit une loi de Bernouilli de paramètre $\frac{p\overline{K_{XY}}}{N}$. On constate alors que les paramètres des lois de Bernouilli que suivent les marginales de $\psi(\mu)$ et $\psi(\nu)$ sont équivalents dans l'asymptotique $p \ll 1$.

Ces deux exemples permettent de voir que les images de deux lois qui prédisent en moyenne le même nombre d'infecté·e·s dans Y peuvent même avoir des distributions typiques de natures totalement différentes (ici supportées par au plus $\overline{K_{XY}}$ éléments contre N) : il y a donc une sensibilité à l'hypothèse faite sur la condition initiale à masse égale. De plus, on retrouve naturellement les difficultés précédemment soulevées : la formule (7.6) n'a pas de raison de pouvoir être établie si l'on part d'une loi quelconque pour décrire l'état de l'infection au sein des classes. Or, l'image d'une loi particulière (typiquement, de Bernouilli indépendantes) n'est génériquement pas une loi aussi structurée d'un ensemble de variables de Bernouilli indépendantes.
On remarque dans les cas traités ci-dessus que les lois de chaque marginale sont les mêmes. Cela provient d'une propriété d'invariance plus générale :

Proposition 10. *Soit $\sigma \in \mathbb{S}_N$, $\mu \in \mathbb{P}\left(\{0,1\}^Y\right)$ une loi quelconque et $(Y_1, \ldots, Y_N)$ de loi μ. Alors (on fait un abus de notation en confondant une variable aléatoire et sa loi) :*

- $\psi(Y_1, \ldots, Y_N) = \psi\left(Y_{\sigma(1)}, \ldots, Y_{\sigma(N)}\right)$,
- *soit $(X_1, \ldots, X_N)$ de loi $\psi(Y_1, \ldots, Y_N)$. Alors $(X_1, \ldots, X_N) = \left(X_{\sigma(1)}, \ldots, X_{\sigma(N)}\right)$.*

Démonstration. Cela découle de l'invariance de la loi uniforme sur $\mathbb{G}_{\overline{K_{XY}}}(X, Y)$ sous l'action de $\mathbb{S}_N$. □

Tentative 4 : modèle sans mémoire.

Pour finir, on introduit un modèle microscopique probabiliste différent de ceux évoqués plus haut : on suppose qu'au matin n, chaque individu de la particule X a une probabilité I_X^n d'être infecté indépendamment des autres. Il est donc possible dans ce modèle (théorique) d'être infecté·e au jour 1, puis sain·e au jour 2, puis infecté·e au jour 3. On couple ce modèle au tirage d'un graphe $\overline{K_{XY}}$-uniforme. Bien que ce modèle n'ait que peu de sens d'un point de vue modélisation, en notant $S_X^n = \mathbb{E}[S_x]$ pour tout $x \in X$ la probabilité moyenne

d'infection au soir dans la particule X, on vérifie que les équations (7.6) se propagent.

En conclusion, le modèle par communauté décrit par l'équation (7.6) ne peut être vu comme l'évolution d'une quantité probabiliste issue d'un modèle microscopique cohérent. Il faut donc garder en tête que les quantités dont l'évolution est étudiée ne sont pas les moyennes dans le sens de la loi des grands nombres d'un modèle aléatoire microscopique. Elles correspondent plutôt à l'établissement d'un modèle d'évolution de proportions (déterministes) de susceptibles et d'infecté·e·s, construit à partir de quantités épidémiologiques (p, γ, ...) pouvant être trouvées dans la littérature spécialisée, qui fait souvent apparaître des modèles probabilistes.

7.3 Modèle microscopique déterministe

Considérons dans cette section le cas extrême où la partition de la population est la partition triviale de la population en singletons $\mathcal{P} = \{\{x_1\}, \ldots, \{x_p\}\}$. On suppose donc qu'on connaît l'ensemble des interactions dans la population.

Définition 6. *On appelle modèle microscopique le modèle défini par* (7.6) *dans le cas où* $\mathcal{P} = \{\{x_1\}, \ldots, \{x_p\}\}$ *et en l'absence de contacts indirects :*

$$\begin{cases} S_x^{n+1} &= S_x^n \prod_{y \in V} \left(1 - I_y^n p\right)^{K_{xy}} \\ I_x^{n+1} &= (1-\gamma) I_x^n + S_x^n \left(1 - \prod_{y \in V} \left(1 - I_y^n p\right)^{K_{xy}}\right) \\ R_x^{n+1} &= R_x^n + \gamma I_x^n. \end{cases} \tag{7.17}$$

On analysera dans le chapitre 8 l'erreur commise par un modèle condensé du type (7.6) par rapport au modèle microscopique dont la description est plus précise. Si l'on a vu que l'approche déterministe n'était pas la limite dans un sens "loi des grands nombres" d'un modèle probabiliste, illustrons cependant leur proximité sur deux exemples jouets.

On a simulé les deux modèles dans le cas d'une population $V = \{A, B, C\}$ avec $K_{AB} = K_{AC} = 1$ et $K_{BC} = 2$ pour un temps total de 5 jours. Même si les prédictions ne sont pas les mêmes, les deux schémas prédisent la même tendance, voir la table 7.1. En particulier, en introduisant le **risque** d'un individu comme la probabilité moyenne d'être infecté pendant l'expérience si l'on introduit une infection aléatoire dans l'établissement, et sa **dangerosité** comme le nombre moyen de personnes infectées si cette particule est la première infectée dans la population, on remarque sur la table 7.2 que les risques et dangerosités relatifs sont les mêmes dans cette situation (très) simple. On note en particulier que le risque et la dangerosité coïncident dans le modèle probabiliste à cause de la

reversibilité du phénomène de propagation.

Modèle	S	I	R
Déterministe	0.813	0.116	0.069
Probabiliste	0.787	0.131	0.081

Table 7.1: Comparaison des modèles déterministe et probabiliste pour une population à 3 personnes. On infecte initialement l'individu A, et on présente les quantités épidémiologiques de la personne C au bout de 5 jours. On a moyenné 10000 essais pour le modèle probabiliste et choisi une probabilité de transmission de 0.05 par contact et un taux de rémission par jour $\gamma = 0.2$.

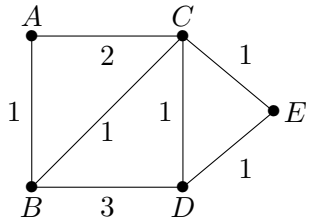


Figure 7.2: Le graphe des contacts journaliers entre 5 personnes étiquetées de A à E. Sur les arêtes figurent les nombres de contact journaliers entre les nœuds de l'arête.

On s'attend à ce que cette hiérarchie commune de la dangerosité de la population ait lieu dans des situations plus complexes, comme c'est encore le cas dans l'exemple représenté sur la figure 7.2. On a simulé les deux modèles avec les paramètres de contacts associés aux arêtes du graphe. Les risques et dangerosités des modèles sont représentés sur la figure 7.3. Une fois encore, les valeurs sont différentes en fonction de la nature de la modélisation, mais la hiérarchie est la même pour les deux modèles : C-B-D-A-E.

Modèle	Individu	Dangerosité	Risque
Déterministe	A	0.458	0.455
	B	0.495	0.496
	C	0.495	0.496
Probabiliste	A	0.47	0.47
	B	0.52	0.52
	C	0.52	0.52

Table 7.2: Comparaison des dangerosités et des risques des modèles déterministe et probabiliste.

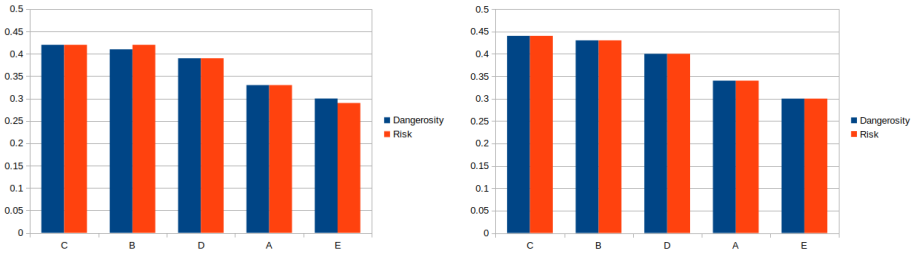


Figure 7.3: Les risques et dangerosités prédits par les deux modèles (gauche : déterministe, droite : probabiliste), pour les données de contacts du graphe de la figure 7.2. On a choisi $p = 0.05$, $\gamma = 0.2$ et répété le schéma pendant 5 jours.

7.4 Analyse de l'asymétrie risque/dangerosité sur un exemple

On a vu dans la section précédente que le risque et la dangerosité d'une particule différaient génériquement dans le modèle déterministe. Analysons ce phénomène sur un graphe linéique où l'on peut calculer explicitement les quantités épidémiologiques pour tout pas de temps. On considère comme sur la figure 7.4 trois personnes disposées sur un graphe linéique, avec un contact par jour pour chaque arête.

Comparons les situations initiales où 0 (ou 2) est initialement infecté au cas où l'infection part du nœud central 1. Dans le second cas, on peut écrire explicitement

$$\begin{cases} I_B^n &= (1-\gamma)^n \\ S_B^n &= 0 \\ S_A^n &= \prod_{k=0}^{n-1} \left(1 - p\,(1-\gamma)^k\right) \\ S_C^n &= \prod_{k=0}^{n-1} \left(1 - p\,(1-\gamma)^k\right). \end{cases} \tag{7.18}$$

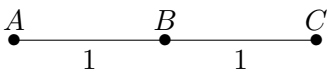
La dangerosité de la particule B vaut donc

$$d_B = 1 + 2\left(1 - \prod_{k=0}^{n-1} \left(1 - p\,(1-\gamma)^k\right)\right). \tag{7.19}$$

Si l'infection part de A, on a à présent

$$I_A^n = (1-\gamma)^n \tag{7.20}$$

Figure 7.4: Un graphe où la dangerosité et le risque diffèrent.



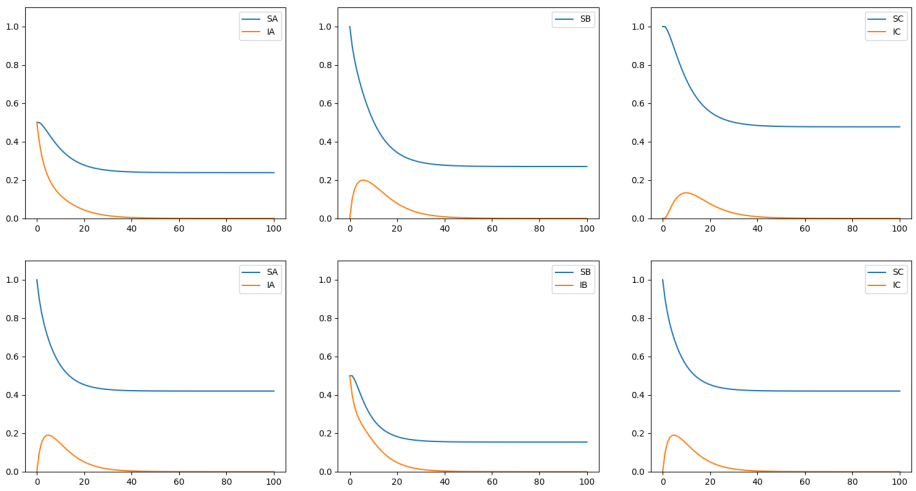


Figure 7.5: Comparaison de l'impact d'une infection de A (première ligne) ou de B (deuxième ligne) sur la population. On note que les impacts de A sur B et réciproquement sont similaires, mais pas identiques (S_B converge vers une valeur inférieure à 0.4 sur la première ligne alors que S_A converge vers une valeur supérieure sur la deuxième ligne.

d'où l'on déduit

$$S_B^n < \prod_{k=0}^{n-1} \left(1 - p\,(1-\gamma)^k\right). \tag{7.21}$$

Par symétrie entre A et C, on peut déduire $r_B > d_B$. On constate qu'une différence entre les deux situations réside dans la complexité des chaînes d'infection dans le cas où A est infectée initialement : une infection de B puis C qui réinfecte B est possible - même si elle n'a pas un grand sens d'un point de vue modélisation pour des nœuds singletons. L'impact de B sur A diffère donc de celui de A sur B. Sur la figure 7.5, on a simulé le modèle pour des conditions initiales moins dégénérées que dans l'exemple précédent, avec une infection initiale de 0.1. On voit que l'hétérogénéité demeure dans ce cas. Même si l'ensemble des chaînes menant de A à B est le même que l'ensemble des chaînes menant de B à A, la non linéarité dans le calcul de I dans (7.17) induit une irréversibilité dans la propagation de l'épidémie.

Chapitre 8

Condensation de modèles SIR

On s'intéresse dans ce chapitre à la condensation de modèles de propagation définis sur des graphes. Plus précisément, on cherche à identifier plusieurs nœuds du graphe que l'on considère comme identiques du point de vue de la modélisation de la propagation. On compare alors le modèle défini sur le graphe condensé au modèle dit "microscopique" défini sur le graphe original. On définit dans un premier temps la condensation de modèle dans un cadre abstrait qui permet de formuler les problématiques générales soulevées par ce processus. On étudie la condensation dans le cas de réseaux résistifs, puis dans le cas des modèles SIR introduits dans le chapitre 7.

8.1 Condensation de modèles sur graphe

Dans le cas où l'on a accès à une matrice $(K_{xy})_{x,y\in V}$ détaillant tous les contacts journaliers de la population, on peut définir conformément au chapitre 7 deux modèles SIR sur graphe : le modèle microscopique défini à l'échelle individuelle par (7.17), et le modèle par communautés défini par (7.6), où l'on a posé pour $X, Y \in \mathcal{P}$

$$K_{XY} = \sum_{x\in X, y\in Y} K_{xy}. \tag{8.1}$$

On peut dès lors étudier la différence entre ces deux modèles, et notamment l'évolution des écarts de prédiction en fonction du temps.
Définissons le processus de condensation dans un cadre plus général. On se donne un graphe microscopique $G = (V, E)$ et une loi d'évolution d'une quantité ϕ_t qui dépend des quantités K_{XY} définies sur les arêtes (X, Y) de G. On définit alors un nouveau graphe $\overline{G} = (\overline{V}, \overline{E})$ dit "condensé" en identifiant certains nœuds de G, et une nouvelle matrice $\overline{K}$ à partir K et d'une loi de sommation des blocs identifiés qui dépend du problème (moyenne, somme, ...). On définit alors une nouvelle loi d'évolution $\overline{\phi}_t$ sur $(\overline{G}, \overline{K})$: on dit que $\overline{\phi}$ est le modèle condensé de ϕ. Un exemple d'une telle condensation de graphe est représenté

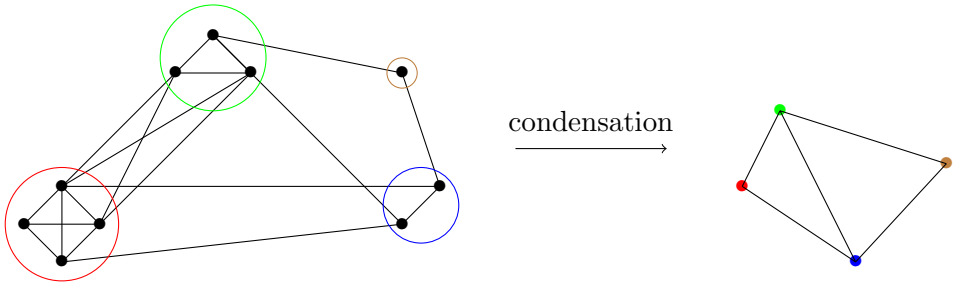


Figure 8.1: À gauche, un graphe dont les longueurs des arêtes sont inversement proportionnelles à leurs connectivités K_{XY}. On a ensuite condensé les nœuds dans les cercles, et choisi pour loi de sommation la somme des conductances : le graphe condensé est représenté à droite.

sur la figure 8.1. On étudie dans ce qui suit les deux principales questions soulevées par ce processus de condensation :

- Comment mesurer l'écart entre $\overline{\phi}$ et ϕ ?
- Étant donnés un graphe microscopique, une matrice et une loi de sommation, quels points identifier pour que $\overline{\phi}$ et ϕ soient proches ?

On traite ces questions dans le cas de la condensation de réseaux électriques, puis des modèles SIR développés dans le chapitre 7. La considération de réseaux électriques est motivée par l'analogie (formelle) suivante. Dans le cas d'un modèle SIR associé à un graphe G et une matrice K sans contacts indirects, lorsque $p \ll 1$, on peut réécrire l'équation sur I de (7.6)

$$I_X^{n+1} - I_X^n - S_X^n \mathcal{L} I_X^n = \beta_X S_X^n I_X^n - \gamma I_X^n \tag{8.2}$$

où $\beta_X = p \sum_{X \sim Y} K_{XY}$ et $\mathcal{L}$ est un opérateur de propagation sur G associé à K, qui s'écrit

$$\mathcal{L}\phi_X = \sum_{X \sim Y} K_{XY}(\phi_X - \phi_Y), \tag{8.3}$$

avec $X \sim Y$ si (X, Y) est une arête de G. On peut voir par cette formulation $\mathcal{L}$ comme un opérateur elliptique discret. Dans le cas d'un graphe linéique dont les conductances sont constantes, l'équation (8.3) est la discrétisation du laplacien en dimension 1. La résolution du problème de Laplace sur un réseau de résistance fait intervenir le même type d'opérateur lorsqu'on essaye d'écrire la dépendance entre tension et intensité dans un réseau via les lois d'Ohm et de Kirchoff.

8.2 Condensation de réseaux résistifs

Un réseau résistif est composé d'un graphe symétrique connexe $G = (V, E)$ et d'une matrice symétrique de conductances K entre nœuds de G. On pose $K_{xy} = 0$ si (x, y) n'est pas une arête de G, et $K_{xy} > 0$ si $(x, y) \in E$. On distingue un ensemble de nœuds externes $\Gamma \subset V$ de l'ensemble des nœuds internes $V \setminus \Gamma$.

8.2.1 Équation de Poisson discrète sur réseau résistif

On s'intéresse ici au cas d'un réseau où chaque nœud interne est connecté à une source qui injecte de l'intensité dans le circuit, et où les nœuds externes sont connectés à la terre (et donc de potentiel nul). On utilise pour décrire le comportement de ce système les deux lois suivantes régissant l'établissement de l'équilibre dans un système électrique :

$$i = cu \quad \text{(Loi d'Ohm)} \tag{8.4}$$

où i est l'intensité traversant une arête, c sa conductance et u la différence de potentiel entre ses extrémités ;

$$\sum_{k=1}^{n} i_k^x = f_x \quad \text{(Loi de Kirchoff)} \tag{8.5}$$

où les $(i_k^x)_{k=1,\dots,n}$ sont les intensités (signées) arrivant à un point x du circuit et f_x est l'intensité injectée en x. Sous ces hypothèses, on trouve le potentiel dans le circuit en résolvant le problème de Poisson suivant :

$$\left\{ \begin{array}{rcll} -\mathcal{L}_K u & = & f & \text{sur } V \setminus \Gamma \\ u & = & 0 & \text{sur } \Gamma \end{array} \right. \tag{8.6}$$

où $\mathcal{L}_K$ est l'opérateur laplacien discret défini par

$$\mathcal{L}_K u(x) = \sum_{y \in G} K_{xy}(u_y - u_x) \quad \forall x \in G. \tag{8.7}$$

La première ligne de (8.6) exprime la conservation de l'intensité aux nœuds, alors que la seconde est une condition limite nécessaire pour avoir unicité de la solution.

Proposition 19. *Si G est connexe,* $\mathrm{Ker}\mathcal{L}_K$ *est l'ensemble des fonctions constantes sur G et est donc de dimension* 1.

Démonstration. Soit $u \in G^{\mathbb{R}}$ tel que $\mathcal{L}_K u = 0$. D'après (8.7) on peut écrire pour tout $x \in G$

$$u_x = \frac{\sum_{y \in G} K_{xy} u_y}{\sum_{y \in G} K_{xy}}. \tag{8.8}$$

u_x est donc dans l'enveloppe convexe des valeurs de ses voisins. Soit $x \in G$, et $x_0 \in G$ tel que u atteint son maximum en x_0. Comme G est connexe, on dispose de $x_0, \ldots, x_k = x$ un chemin dans G de x_0 à x. On montre par récurrence que pour $j = 0, \ldots, k-1$, $u_{x_j} = u_{x_0}$. Si $u_{x_j} = u_{x_0}$, comme u_{x_j} est dans l'enveloppe convexe de ses voisins et que $(x_j, x_{j+1}) \in E$, on obtient $u_{x_j} = u_{x_{j+1}}$. Par conséquent $u_{x_0} = u_x$ et u est constante. □

On définit à présent le processus de condensation d'un réseau résistif étant donné une partition en sous-communautés. Soit $\mathcal{P}$ une partition de G, avec $\Gamma \in \mathcal{P}$. On identifie les nœuds de chaque classe en les contraignant à avoir le même potentiel (en les connectant par des conductances infinies). Comme dans les circuits électriques deux conductances en parallèle sont équivalentes à leur somme, on pose comme conductance équivalente entre les classes X et Y

$$\overline{K_{XY}} = \sum_{x \in X, y \in Y} K_{xy}. \tag{8.9}$$

En posant $f_X = \sum_{x \in X} f_x$, on cherche à présent une solution de

$$\left\{ \begin{array}{rcll} -\mathcal{L}_{\bar{K}} w & = & f & \text{sur } \mathcal{P} \setminus \{\Gamma\} \\ w_\Gamma & = & 0 & \end{array} \right. \tag{8.10}$$

où $w : \mathcal{P} \longrightarrow \mathbb{R}$. On définit ensuite $\bar{u} : V \longrightarrow \mathbb{R}$ l'interpolation de w sur G par $\bar{u}_x = w_X$ pour tout $X \in \mathcal{P}$ et $x \in X$.

Remarque 31. *La démarche de calculer la solution de la version discrétisée d'une EDP par calcul sur un maillage plus grossier peut-être reliée aux méthodes dites "multigrille". Généralement utilisées pour calculer des solutions approchées d'équations aux dérivées partielles linéaires, elles sont basées sur le calcul récursif de l'erreur sur différents maillages pour obtenir une convergence rapide de toutes les fréquences (voir [70] pour une introduction détaillée). Le point de départ de la méthode multigrille est l'observation du fait que lors de l'utilisation d'un algorithme itératif pour résoudre une EDP linéaire, l'erreur devient lisse mais pas nécessairement petite. On remarque donc que seules les hautes fréquences (vis-à-vis du maillage) sont efficacement calculées. On calcule alors les basses fréquences sur un maillage plus grossier - pour lequel certaines basses fréquences seront hautes. Dans le contexte de condensation, le graphe condensé peut être considéré comme le maillage grossier et le graphe exact comme le maillage raffiné.*

Définissons un cadre pour comparer les potentiels introduits. Soit

$$\mathcal{V} = \{v : V \longrightarrow \mathbb{R},\, v = 0 \text{ sur } \Gamma\}, \tag{8.11}$$

muni de la norme $\|v\|_K^2 = \frac{1}{2} \sum_{x,y} K_{xy}(v_x - v_y)^2$.

Remarque 32. *Par équivalence des normes en dimension finie, on obtient l'existence d'une constante $C > 0$ telle que pour tout $v \in \mathcal{V}$,*

$$\|v\|_2 \leqslant C\|v\|_K. \tag{8.12}$$

L'équation (8.12) *est une équation de type Poincaré, avec une constante C qui dégénère lorsque le graphe grandit, avec l'existence de longs chemins partant de la frontière. En effet, soit un chemin de longueur n $x_1, \ldots, x_n$ avec $x_1 \in \Gamma$, de variations locales $(v_{x_i} - v_{x_{i+1}})$ d'ordre 1. Alors v_{x_n} est potentiellement d'ordre n (par exemple dans le cas d'un graphe linéique de taille n et Γ un de ses nœuds extrémaux).*

Comme dans le contexte des EDP, écrivons la formulation variationnelle de l'équation (8.6). En multipliant des deux côtés par ϕ, on a pour tout $x \in V \setminus \Gamma$

$$\sum_{y \in V} K_{xy}\phi_x(u_x - u_y) = f_x\phi_x. \tag{8.13}$$

En sommant sur les x, on obtient la formulation variationnelle suivante du problème de Poisson :

$$\sum_{x,y \in V} K_{xy}(u_x - u_y)(\phi_x - \phi_y) = \sum_{x \in V} f_x\phi_x. \tag{8.14}$$

L'équation (8.14) est une condition d'optimalité pour la minimisation de la fonctionnelle strictement convexe

$$J : \begin{cases} \mathcal{V} & \longrightarrow & \mathbb{R} \\ v & \longmapsto & \|v\|_K^2 - \sum\limits_{x \in V} f_x v_x. \end{cases} \tag{8.15}$$

On peut alors définir u comme le minimiseur de J sur $\mathcal{V}$. De même, en définissant

$$\mathcal{V}_\mathcal{P} = \{v \in \mathcal{V},\ v \text{ constante sur } X \in \mathcal{P}\}, \tag{8.16}$$

on obtient

$$\bar{u} = \operatorname*{argmin}_{v \in \mathcal{V}_\mathcal{P}} J(v). \tag{8.17}$$

La proposition suivante donne une première estimation formelle de la distance entre les deux potentiels pour la norme $\|\cdot\|_K$:

Proposition 20. *On a l'estimation suivante :*

$$\|u - \bar{u}\|_K \leqslant \inf_{v \in \mathcal{V}_\mathcal{P}} \|u - v\|_K. \tag{8.18}$$

Démonstration. Par le lemme de Cea (voir [21]), on a $\bar{u} = P_{\mathcal{V}_\mathcal{P}}(u)$, où $P_{\mathcal{V}_\mathcal{P}}$ est la projection orthogonale sur $\mathcal{V}_\mathcal{P}$ pour la norme $\|\cdot\|_K$. □

On peut calculer $\bar{u}$ à partir de u : pour tout $X \in \mathcal{P}$ et $x \in X$,

$$\bar{u}_x = \frac{\sum_{y \in X} K_y u_y}{\sum_{y \in X} K_y}, \tag{8.19}$$

où $K_y = \sum_{x \sim y} K_{xy}$. Le potentiel condensé $\bar{u}$ est donc une moyenne pondérée de u sur chaque classe de la partition $\mathcal{P}$. En injectant la formulation (8.19) dans (8.18), on obtient la proposition suivante.

Proposition 21. *On peut estimer*

$$\|u - \bar{u}\|_K \leqslant 2 \left(\frac{1}{2} \sum_{X \in \mathcal{P}} \sum_{x \in X} K_x \left(u_x - \frac{\sum_{y \in X} K_y u_y}{\sum_{y \in X} K_y} \right)^2 \right)^{\frac{1}{2}}. \tag{8.20}$$

On obtient le comportement attendu, qui est que u et $\bar{u}$ doivent être proches lorsque u varie peu sur chaque classe de $\mathcal{P}$. La partie de droite de (8.20) peut être interprétée comme la somme des variances dans chaque classe par rapport à la norme naturelle, pondérée par les conductivités K_y.

Un cas limite avec une estimée explicite de la distance au potentiel condensé. Analysons en détail un cas particulier où l'on peut calculer explicitement une borne sur la distance entre le potentiel et sa condensation.
Pour toute matrice K et $X \subset V$ tels que $X \cap \Gamma = \emptyset$, considérons la partition $\mathcal{P}_X$ où la seule classe de $\mathcal{P}$ qui n'est pas un singleton est X. Supposons par ailleurs que que pour toute paire $(x, y) \in X \times X$, il existe un chemin de x à y dans X composé d'arêtes de conductances positives. Soit $\epsilon > 0$ et la matrice de conductances

$$K^\epsilon_{xy} = K_{xy} + \frac{\mathbf{1}_{x,y \in X}}{\epsilon} K_{xy}. \tag{8.21}$$

Quand ϵ tend vers 0, les conductances internes à X définies par la matrice K^ϵ explosent. Soit u_ϵ le problème de Poisson sans contrainte défini par (8.6) pour la matrice de conductances K^ϵ et $\bar{u}_\epsilon$ le problème condensé correspondant selon la partition $\mathcal{P}_X$.

Proposition 22. *Quand ϵ tend vers 0,*

$$\|\bar{u}_\epsilon - u_\epsilon\|_K = O(\epsilon). \tag{8.22}$$

Démonstration. Soit $\bar{u}$ la solution du problème de Poisson condensé (8.10) pour la partition $\mathcal{P}_X$: $\bar{u}$ minimise la fonctionnelle J sur $\mathcal{V}_\mathcal{X} := \{v \in \mathcal{V}, v \text{ constante sur } X\}$. En définissant la forme bilinéaire

$$b(u, v) = \sum_{x,y \in X} K_{xy}(u_x - u_y)(v_x - v_y), \tag{8.23}$$

on a $\mathcal{V}_\mathcal{X} = \mathrm{Ker}(b)$. Comme $b(u,v) = (Bu, Bv)_K$, avec

$$Bu : \begin{cases} V & \longrightarrow & \mathbb{R} \\ x & \longmapsto & \begin{cases} u(x) & \text{si } x \in X \\ 0 & \text{sinon} \end{cases} \end{cases} \tag{8.24}$$

et $(\ ,\ )_K$ est le produit scalaire associé à K, on peut utiliser le corollaire 2.4 dans [43] pour obtenir

$$\|\bar{u} - u_\epsilon\|_K = O(\epsilon). \tag{8.25}$$

$\bar{u}_\epsilon$ minimise

$$J_\epsilon : \begin{cases} \mathcal{V} & \longrightarrow & \mathbb{R} \\ v & \longmapsto & \frac{1}{2} \sum_{x,y \in V} K_{xy} (1 + \frac{\mathbf{1}_{x,y \in X}}{\epsilon})(v_x - v_y)^2 - \sum_{x \in V} f_x v_x. \end{cases} \tag{8.26}$$

sur $\mathcal{V}_\mathcal{X}$, mais sur l'ensemble des champs constants sur X on a $J_\epsilon = J$. On obtient donc $\bar{u}_\epsilon = \bar{u}$. □

Cette proposition illustre le fait qu'une condensation est plus précise pour la norme $\|\cdot\|_K$ si l'on a condensé des arêtes reliées par des conductances élevées.

8.2.2 Équation de Laplace sur un réseau résistif

On étudie à présent la propagation d'un courant dans un circuit où l'on a imposé le potentiel à deux nœuds : on contraint le potentiel à valoir 0 à une source s et 1 à un puits t, puis on calcule le courant résultant dans le circuit. De manière équivalente, cela revient à brancher une pile de 1 volt aux extrémités (s,t) du système. Ce problème a été étudié d'un point de vue probabiliste dans [27], et est connu sous le nom de "pressure drop" en mécanique des fluides, voir par exemple [22]. On peut calculer le potentiel V en résolvant le problème de Laplace discret suivant :

$$\begin{cases} -\mathcal{L}_K u &= 0 & \text{sur } G \setminus \{s,t\} \\ u_s &= 0 \\ u_t &= 1. \end{cases} \tag{8.27}$$

On peut résoudre le problème (8.27) en inversant une matrice : en notant les nœuds de G $\{1, \ldots, n-2, s, t\}$, considérons la matrice

$$\begin{aligned} A_{ij} &= \frac{K_{i,j}}{K_i} && \text{pour } i \leqslant n-2, j = 1, \ldots, n,\ j \neq i \\ A_{ii} &= -1 && \text{pour } i \leqslant n-2 \\ A_{ij} &= \delta_{ij} && \text{pour } i > n-2, j = 1, \ldots, n, \end{aligned} \tag{8.28}$$

où $K_i = \sum_{j \neq i} K_{ij}$. L'équation (8.27) peut être réécrite $Au = B$, où $B = (0, \ldots, 0, 1)$.
Comme on l'a fait précédemment pour l'équation de Poisson, on peut définir le modèle condensé associé à une partition $\mathcal{P}$, en imposant au potentiel à être constant sur chaque classe de $\mathcal{P}$.
On suppose par la suite que la classe de la source et du puits sont des singletons, et on pose encore $\overline{K_{XY}} = \sum_{x \in X, y \in Y} K_{xy}$. Le modèle condensé s'écrit alors :

$$\left\{ \begin{array}{rcll} -\mathcal{L}_{\bar{K}} w & = & 0 & \text{sur } \mathcal{P} \setminus \{s,t\} \\ w_s & = & 0 & \\ w_t & = & 1. & \end{array} \right. \tag{8.29}$$

On définit ensuite $\bar{u} : G \longrightarrow \mathbb{R}$ l'interpolation de w sur G : pour tout $X \in \mathcal{P}$ et $x \in X$, $\bar{u}_x = w_X$. Soit C la matrice définie par

$$\begin{array}{ll} C_{ij} = \dfrac{\bar{K}_{X_i,X_j}}{N_j \overline{K_{X_i}}} & \text{pour } i \leqslant N-2, j = 1, \ldots, N,\ X_i \neq X_j \\ C_{ij} = -\delta_{ij} & \text{pour } i \leqslant N-2, j \in X_i \\ C_{ij} = \delta_{ij} & \text{pour } i > N-2, j = 1, \ldots, N, \end{array} \tag{8.30}$$

où X_i est la classe de i, $\overline{K_{X_i}} = \sum_{X \neq X_i} \bar{K}_{X,X_i}$, N est le nombre total de classes et N_j le cardinal de X_j. L'équation (8.29) s'écrit $C\bar{u} = B$, avec $B = (0, \ldots, 0, 1)$.
On cherche à présent à trouver des conditions sur K et $\mathcal{P}$ pour que u et $\bar{u}$ soient proches (pour une norme à préciser). On peut estimer pour tout norme matricielle l'erreur de condensation par

$$\begin{array}{rl} \|u - \bar{u}\| = & \|A^{-1}B - C^{-1}B\| \\ \leqslant & \|A^{-1}\| \|C - A\| \|C^{-1}B\|. \end{array} \tag{8.31}$$

Pour la norme 2, on obtient

$$\|A - C\|_2^2 \leqslant \|A^{-1}\|_2^2 \sum_{i=1}^{n-2} \left(\sum_{\substack{j=1 \\ j \notin X_i}}^{n-2} \left(\frac{K_{ij}}{K_i} - \frac{\bar{K}_{X_i X_j}}{\bar{K}_{X_i} N_j} \right)^2 + \sum_{\substack{j \neq i \\ j \in X_i}} \left(\frac{K_{ij}}{K_i} \right)^2 \right). \tag{8.32}$$

Dans le premier terme $\dfrac{K_{ij}}{K_i} - \dfrac{\bar{K}_{X_i X_j}}{\bar{K}_{X_i} N_j}$, $\dfrac{K_{ij}}{K_i}$ est la part des conductances partant de i due à la connexion avec j, alors que

$$\frac{\bar{K}_{X_i X_j}}{\bar{K}_{X_i} N_j} = \frac{\bar{K}_{X_i X_j}}{N_j} \frac{1}{\bar{K}_{X_i}} \tag{8.33}$$

est le ratio entre la conductance moyenne entre les particules de X_i et X_j sur la conductance totale de X_i à la totalité du réseau. Ce ratio représente donc la part de la conductance reliant X_i à l'extérieur qui est due aux connexions avec X_j. Ce terme est donc minimal dans le cas où les nœuds de X_i sont similairement connectés au reste du réseau.
Le deuxième terme est une perte due à l'estimation grossière (8.31) : il diminue lorsqu'il n'y a pas de connexion interne dans un groupe. On voit néanmoins que pour une conductance interne totale $\sum_{j\in X_i} K_{ij}$ fixée, il est minimal dans la situation isotrope où K_{ij} est constante pour $j \in X_i$.

Remarque 33. *Il y a différentes manières de définir les blocs diagonaux de C, c'est-à-dire les C_{ij} pour $j \in X_i$. En effet, on peut ajouter une constante β_i à chaque C_{ij} pour $j \in X_i$ et $j \neq i$, puis retrancher $\beta_i(N_i - 1)$ au terme diagonal C_{ii}, sans modification de $C^{-1}B$ - l'estimation précédente correspondant au cas $\beta = 0$. La borne optimale dans* (8.32) *est atteinte pour*

$$\beta_i = \frac{\sum_{j\in X_i} \frac{K_{ij}}{K_i}}{(N_i-1)^2 + (N_i - 1)}. \tag{8.34}$$

Dans le cas idéal $\dfrac{K_{ij}}{K_i} = c_i$ pour tout $j \in X_i$, on obtient $\beta_i = \dfrac{c_i}{N_i}$, et le deuxième terme de (8.31) *devient $\dfrac{(N_i-1)^2 c_i}{N_i^3}$: il s'annule lorsqu'il n'y a aucune connexion interne dans les classes identifiées.*

Malgré les aspects négatifs de l'estimation (8.31), le comportement du premier terme illustre l'intuition suivante : des points connectés de la même manière au reste du réseau tendent à avoir le même potentiel dans l'équation de Laplace.

8.3 Condensation de modèles SIR

On étudie dans cette section la condensation d'un modèle SIR sur graphe. Étant donnée une matrice de contacts microscopique $(K_{xy})_{x,y\in V}$ et une partition $\mathcal{P}$, on considère $(S_x^n, I_x^n, R_x^n)_{x\in V, n\geqslant 0}$ solution du modèle SIR microscopique défini par (7.17), avec des conditions initiales constantes sur chaque classe $X \in \mathcal{P}$. En définissant pour tous $X, Y \in \mathcal{P}$

$$\overline{K_{XY}} = \frac{\sum_{x\in X, y\in Y} K_{xy}}{N_X} \tag{8.35}$$

où N_X est le cardinal de X, soit $(S_X^n, I_X^n, R_X^n)_{X\in\mathcal{P}, n\geqslant 0}$ solution du modèle SIR condensé par communautés suivant :

$$\left\{\begin{array}{lcl} S_X^{n+1} & = & S_X^n \prod\limits_{Y\in\mathcal{P}} (1-pI_Y^n)^{\overline{K_{XY}}} \\ I_X^{n+1} & = & S_X^n \left(1-\prod\limits_{Y\in\mathcal{P}} (1-pI_Y^n)^{\overline{K_{XY}}}\right) + (1-\gamma)\, I_X^n \\ R_X^{n+1} & = & R_X^n + \gamma I_X^n, \end{array}\right. \tag{8.36}$$

c'est-à-dire le modèle introduit au chapitre 7 sans contacts indirects, de même conditions initiales que le modèle microscopique. On cherche à estimer l'erreur commise par le modèle condensé par rapport au modèle microscopique lorsqu'on a remplacé la donnée exacte des contacts encodée dans la matrice K_{xy} par sa version moyennée $\overline{K_{XY}}$.

8.3.1 Estimation de la qualité de la condensation

Définissons l'erreur commise au temps n par

$$\epsilon_n = \max\left(\max_{X\in\mathcal{P},\, x\in X} \left|S_x^n - S_X^n\right|,\ \max_{X\in\mathcal{P},\, x\in X} \left|I_x^n - I_X^n\right|\right), \tag{8.37}$$

c'est-à-dire le plus grand écart de prédiction entre les deux modèles au temps n. On utilise la croissance de cette erreur pour quantifier la qualité d'une partition $\mathcal{P}$ de la population.

La proposition suivante donne une estimée de l'erreur au temps n en fonction de K.

Proposition 23. *En posant $K_{xY} = \sum_{y\in Y} K_{xy}$ pour $x \in V$ et $Y \subset V$, et $\overline{K_{XY}} = \dfrac{K_{XY}}{N_X}$ où N_X est le cardinal de X, on a pour tout $n \geqslant 1$*

$$\epsilon_n \leqslant \left(\left(1+2p\max_{x\in V} K_{xV}\right)^n - 1\right) \frac{\max\limits_{X\in\mathcal{P},\, x\in X} \sum_{Y\in\mathcal{P}} \left|K_{xY} - \overline{K_{XY}}\right|}{\max\limits_{x\in V} K_{xV}}. \tag{8.38}$$

Pour montrer la proposition 23, on utilise le lemme technique suivant :

Lemme 15. *Soit $(a_i)_{i=1,\dots,n}$, $(b_i)_{i=1,\dots,n}$ des réels de l'intervalle $[0,\, 1]$. On a*

$$\left|\prod_{i=1}^n a_i - \prod_{i=1}^n b_i\right| \leqslant \sum_{i=1}^n |a_i - b_i|\,. \tag{8.39}$$

Démonstration du lemme 15. Immédiat par récurrence. □

Démonstration de la proposition 23. On procède par récurrence : soit $n \in \mathbb{N}$.

$$\begin{aligned}
\left|S_x^{n+1} - S_X^{n+1}\right| &= \left| S_x^n \prod_{y\in V} \left(1 - pI_y^n\right)^{K_{xy}} - S_X^n \prod_{Y\in V} \left(1 - pI_Y^n\right)^{\overline{K_{XY}}} \right| \\
&\leqslant \left|S_x^n - S_X^n\right| \prod_{y\in V} \left(1 - pI_y^n\right)^{K_{xy}} + S_X^n \left| \prod_{y\in V} \left(1 - pI_y^n\right)^{K_{xy}} - \prod_{Y\in\mathcal{P}} \left(1 - pI_Y^n\right)^{\overline{K_{XY}}} \right| \\
&\leqslant \epsilon_n + \left| \prod_{Y\in\mathcal{P}} \prod_{y\in Y} \left(1 - pI_y^n\right)^{K_{xy}} - \prod_{Y\in\mathcal{P}} \left(1 - pI_Y^n\right)^{\overline{K_{XY}}} \right| \\
&\leqslant \epsilon_n + \left| \prod_{Y\in\mathcal{P}} \prod_{y\in Y} \left(1 - pI_y^n\right)^{K_{xy}} - \prod_{Y\in\mathcal{P}} \prod_{y\in Y} \left(1 - pI_Y^n\right)^{K_{xy}} \right| \\
&\qquad + \left| \prod_{Y\in\mathcal{P}} \left(1 - pI_Y^n\right)^{K_{xY}} - \prod_{Y\in\mathcal{P}} \left(1 - pI_Y^n\right)^{\overline{K_{XY}}} \right|
\end{aligned}$$

On utilise trois fois le lemme 15 pour obtenir successivement les estimations suivantes :

$$\left| \left(1 - pI_y^n\right)^{K_{xy}} - \left(1 - pI_Y^n\right)^{K_{xy}} \right| \leqslant p\epsilon_n K_{xy} \tag{8.40}$$

$$\left| \prod_{y\in Y} \left(1 - pI_y^n\right)^{K_{xy}} - \prod_{y\in Y} \left(1 - pI_Y^n\right)^{K_{xy}} \right| \leqslant p\epsilon_n K_{xY} \tag{8.41}$$

$$\left| \prod_{Y\in\mathcal{P}} \prod_{y\in Y} \left(1 - pI_y^n\right)^{K_{xy}} - \prod_{Y\in\mathcal{P}} \prod_{y\in Y} \left(1 - pI_Y^n\right)^{K_{xy}} \right| \leqslant p\epsilon_n K_{xV}. \tag{8.42}$$

D'autre part,

$$\left| \prod_{Y\in\mathcal{P}} \left(1 - pI_Y^n\right)^{K_{xY}} - \prod_{Y\in\mathcal{P}} \left(1 - pI_Y^n\right)^{\overline{K_{XY}}} \right| \leqslant p \sum_{Y\in\mathcal{P}} \left|K_{xY} - \overline{K_{XY}}\right|. \tag{8.43}$$

On obtient la borne

$$\left|S_x^{n+1} - S_X^{n+1}\right| \leqslant \epsilon_n \left(1 + pK_{xV}\right) + p \sum_{Y\in\mathcal{P}} \left|K_{xY} - \overline{K_{XY}}\right| \tag{8.44}$$

Un calcul similaire donne

$$\left|I_x^{n+1} - I_X^{n+1}\right| \leqslant \epsilon_n \left(1 - \gamma + 2pK_{xV}\right) + p \sum_{Y\in\mathcal{P}} \left|K_{xY} - \overline{K_{XY}}\right| \tag{8.45}$$

L'erreur vérifie donc

$$\begin{cases} \epsilon_{n+1} & \leqslant & \left(1 + 2p \max_{x \in V} K_{xV}\right) \epsilon_n + p \max_{X \in \mathcal{P},\, x \in X} \sum_{Y \in \mathcal{P}} \left| K_{xY} - \overline{K_{XY}} \right| \\ \epsilon_0 & = & 0 \end{cases} \tag{8.46}$$

On en déduit par comparaison avec une suite arithmético-géométrique :

$$\epsilon_n \leqslant \left(\left(1 + 2p \max_{x \in V} K_{xV}\right)^n - 1 \right) \frac{\max_{X \in \mathcal{P},\, x \in X} \sum_{Y \in \mathcal{P}} \left| K_{xY} - \overline{K_{XY}} \right|}{\max_{x \in V} K_{xV}}. \tag{8.47}$$

□

Dans la limite $p \max_{x \in V} K_{xV} \ll 1$, l'estimation (8.38) devient :

$$\epsilon_n \leqslant 2np \max_{X \in \mathcal{P},\, x \in X} \sum_{Y \in \mathcal{P}} \left| K_{xY} - \overline{K_{XY}} \right|. \tag{8.48}$$

Dans ce cas l'erreur croît linéairement en n, avec une pente donnée par la pire approximation faite lorsqu'on a remplacé K_{xY}, le nombre total de contacts de $x \in V$ avec un groupe $Y \in \mathcal{P}$ par sa version moyennée $\overline{K_{XY}}$.

8.3.2 Condensation exacte : un exemple

La borne donnée par la proposition 23 permet de dériver un premier cas où la condensation est exacte vis-à-vis du modèle microscopique :

Corollaire 3. *Supposons que pour tous groupes $X, Y \in \mathcal{P}$, K_{xY} ne dépend pas de $x \in X$ choisi. Alors les deux modèles prédisent les mêmes statistiques d'infection, c'est-à-dire que l'erreur ϵ_n définie par* (8.37) *est nulle pour tout temps.*

Démonstration. On a pour $X, Y \in \mathcal{P}$ et pour tout $x \in X$ $K_{xY} = \overline{K_{XY}}$, donc (8.38) donne $\epsilon_n = 0$ pour tout n. □

Remarque 34. *On peut aussi montrer par récurrence que la formule* (7.17) *ne dépend pas du représentant d'une communauté fixée.*

Étudions à présent le cas où l'on autorise les K_{xy} contacts entre x et y à avoir des probabilités de transmission variables $p_{xy}^1, \ldots, p_{xy}^{K_{xy}}$. Le modèle microscopique s'écrit alors

$$\begin{cases} S_x^{n+1} & = & S_x^n \prod_{y \in V} \prod_{s=1}^{K_{xy}} \left(1 - I_y^n p_{xy}^s\right) \\ I_x^{n+1} & = & (1-\gamma) I_x^n + S_x^n \left(1 - \prod_{y \in V} \prod_{s=1}^{K_{xy}} \left(1 - I_y^n p_{xy}^s\right)\right) \\ R_x^{n+1} & = & R_x^n + \gamma I_x^n. \end{cases} \tag{8.49}$$

Remarque 35. *Considérons le cas d'une population réelle où l'on a un suivi précis de l'ensemble des interactions comme dans l'étude présentée dans [32]. Dans cette étude, les employé·e·s d'un immeuble ont porté des capteurs permettant d'obtenir des données complètes de contact entre tou·te·s les participant·e·s. Les contacts ont alors une durée ; étant donné λ un taux de transmission par unité de temps, on associe à un contact de temps t une probabilité $p = 1 - e^{-t\lambda}$.*

Proposition 24. *Sous les hypothèses précédentes, on suppose que pour toute paire $(x, Y) \in V \times \mathcal{P}$, K_{xY} et l'ensemble des probabilités $(p^s_{xy})_{y\in Y, s=1,\dots,K_{xy}}$ ne dépendent pas du représentant $x \in X$ choisi. Alors le modèle microscopique généralisé* (8.49) *prédit des quantités épidémiologiques qui ne dépendent pas du représentant d'une communauté donnée. En définissant le modèle condensé généralisé*

$$\begin{cases} S_X^{n+1} &= S_X^n \prod\limits_{Y\in\mathcal{P}} \prod\limits_{s=1}^{K_{XY}} (1 - I_Y^n p^s_{XY}) \\ I_X^{n+1} &= (1-\gamma) I_X^n + S_X^n \left(1 - \prod\limits_{Y\in\mathcal{P}} \prod\limits_{s=1}^{K_{XY}} (1 - I_Y^n p^s_{XY}) \right) \\ R_X^{n+1} &= R_X^n + \gamma I_X^n, \end{cases} \tag{8.50}$$

où $(p^s_{XY})_{s=1,\dots,K_{XY}} = (p^i_{xy})_{y\in Y, i=1,\dots,K_{xy}}$ pour n'importe quel $x \in X$, il y a condensation exacte entre les deux modèles généralisés.

Démonstration. Il suffit de le montrer par récurrence en regroupant les termes dans les produits dans (8.49) et en utilisant le fait que les probabilités $(p^i_{xy})_{y\in Y, i=1,\dots,K_{xy}}$ ne dépendent pas de $x \in X$. □

Remarque 36. *L'hypothèse de la proposition 24 peut sembler restrictive. Considérons néanmoins une situation où des groupes de personnes bougent ensemble dans l'espace ($\mathbb{R}^2$, ou un graphe ...). On modélise un croisement entre deux groupes en associant à chaque personne un nombre fixe de contacts de durée donnée avec l'autre groupe (par exemple chaque personne du groupe X a 3 contacts d'une minute avec des gens du groupe Y). Cette situation satisfait les hypothèses précédentes.*

Finissons cette section avec une analyse de la situation précédente lorsque les probabilités sont faibles et de la forme $1 - e^{-\lambda t}$, avec t le temps de contact et λ le taux d'infection lors d'un contact. Dans ce cas, on peut écrire

$$1 - \prod_{Y\in\mathcal{P}} \prod_{s=1}^{\overline{K_{XY}}} (1 - I_Y^n p^s_{XY}) \approx \sum_{Y\in\mathcal{P}} I_Y^n \sum_{s=1}^{\overline{K_{XY}}} p^s_{XY}. \tag{8.51}$$

Conformément à la remarque 35, on pose $p^s_{XY} = 1 - e^{-\lambda t^s_{XY}} \approx \lambda t^s_{XY}$. L'infection dans X au temps n est alors

$$I^n_X \approx \lambda S^n_X \sum_{Y \in \mathcal{P}} I^n_Y \sum_{s=1}^{\overline{K_{XY}}} t^s_{XY}. \tag{8.52}$$

Il apparaît que la condition pour avoir condensation exacte dans ce cas est que pour toute paire $(X, Y) \in \mathcal{P}^2$, le temps total passé par un individu $x \in \mathcal{P}$ et des personnes de $\mathcal{P}$, $\sum_{y \in Y} \sum_{s=1}^{K_{xy}} t^s_{xy}$, ne dépende pas du représentant x choisi.

Chapitre 9

CrowdCovid : implémentation pour des établissements scolaires

On présente dans ce chapitre un algorithme de calcul du risque épidémiologique d'un emploi du temps dans un établissement scolaire. Destinée à des chef·fe·s d'établissements scolaires, l'implémentation prend la forme d'une interface web sur laquelle on introduit les plans de l'école et l'emploi du temps. L'algorithme calcule les déplacements de chaque classe, traduit les croisements en terme de contacts et compare les différents scénarios d'infection à l'aide du modèle SIR par communautés introduit dans le chapitre 7.

9.1 MODCOV9

Lors des premiers mois de la crise sanitaire en France, la plateforme MODCOV19[1] a été créée sous l'impulsion du CNRS[2] et de l'INSMI[3] afin de coordonner les différents projets de modélisation scientifique en épidémiologie. Les vocations de cette plateforme sont multiples :

- Réaliser un travail de veille scientifique sur les projets de modélisation liés à l'épidémie,
- Organiser des groupes de travail pour encourager les discussions entre projets au sein de MODCOV19,
- Identifier les besoins de données des projets de recherche et coordonner ces projets avec les organismes publics susceptibles de pourvoir à ces besoins.

Au printemps 2020, nous avons été contactés par MODCOV19 avec Bertrand Maury et Sylvain Faure (Laboratoire de Mathématiques d'Orsay) pour travailler sur la modélisation de

[1] https://modcov19.math.cnrs.fr/
[2] https://www.cnrs.fr
[3] https://www.insmi.cnrs.fr/

la propagation du Covid-19 dans les écoles. Le développement des modèles SIR du chapitre 7 et l'implémentation présentée dans celui-ci font suite à cette sollicitation. MODCOV19 a notamment débloqué un fond permettant d'embaucher un développeur pour la mise en place d'une application web nommée CroCo (CrowdCovid) hébergeant notre implémentation pour une utilisation en ligne.
Après avoir pris connaissance de la littérature en épidémiologie (modèles SIR déterministes ou probabilistes, microscopiques ou macroscopiques), nous avons identifié trois particularités de la dynamique des interactions sociales dans les établissements scolaires. D'une part, la population est structurée en classes (ou du moins en partie de classes en fonction des options) qui restent groupées lors de la semaine. D'autre part, les classes se déplacent entre un nombre fini de salles et de lieux identifiés - cantine, CDI, etc. - : on peut donc modéliser l'espace de déplacement par un graphe. Enfin, étant donné les plans de l'établissement, on peut reconstruire le déplacement de chaque individu sur le graphe à l'aide de son emploi du temps. Nous avons donc envisagé un algorithme en trois phases :

- Phase 1 : reconstruction des déplacements dans l'école à l'aide des emplois du temps et identification des situations de croisement,
- Phase 2 : traduction des croisements en termes de contacts directs, ou indirects dans le cas où deux classes se succèdent dans la même salle pour prendre en compte la transmission par aérosol,
- Phase 3 : utilisation d'un modèle SEIR pour calculer l'ampleur d'une infection théorique dans l'établissement.

Nous avons choisi une implémentation qui prenait en entrée un emploi du temps (sous forme de tableur pour être compatible avec l'export des logiciels d'aide à la conception d'emploi du temps de l'Éducation Nationale) et les plans de l'établissement sous forme de graphe. L'algorithme retourne alors un score associé à l'emploi du temps, qui représente l'ampleur moyenne de l'infection si une épidémie se déclenche dans l'école, ainsi que des visuels permettant une analyse plus fine de la dynamique infectieuse. Le fonctionnement général de l'algorithme est représenté sur la figure 9.1. La mise en place d'une application web rend cet outil utilisable par des chef·fe·s d'établissement désireux de comparer différents scénarios d'emploi du temps dans son école. Le modèle d'épidémiologie choisi est un modèle SEIR (sain·e·s - infecté·e·s mais pas encore contagieux·ses - infectieux·ses - remis·e·s) par communautés du type de celui développé dans le chapitre 7. Une communauté est ici un groupe d'élèves partageant les mêmes options dans l'emploi du temps, par exemple : "6emeA, LV1 anglais, demi-pensionnaire".

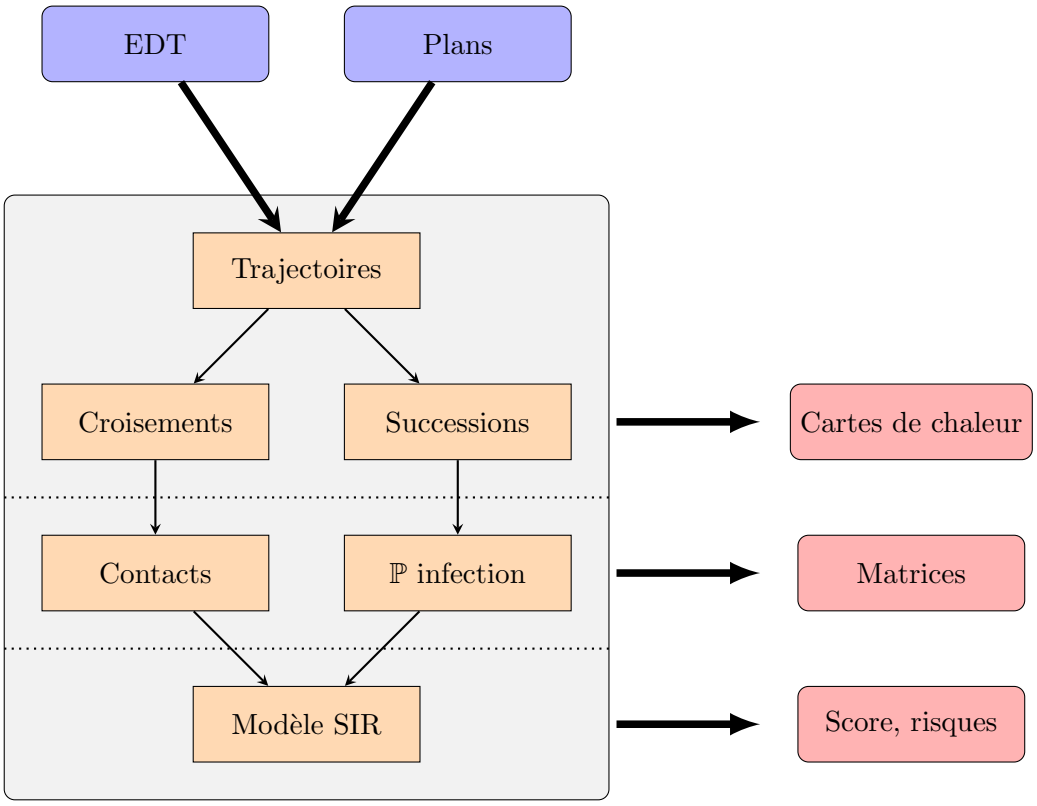


Figure 9.1: Fonctionnement général de l'algorithme CroCo (dans l'encadré gris). On entre des emplois du temps et des plans, on récupère à l'issue de chaque phase (délimitées par pointillés) des cartes de chaleur, des matrices, puis un score associé à des risques. Lors de la première phase, on reconstitue les déplacements puis on détecte les croisements et successions. La deuxième phase traduit ces occurrences en contacts et probabilité d'infection indirecte. Le modèle épidémiologique est itéré lors de la troisième phase.

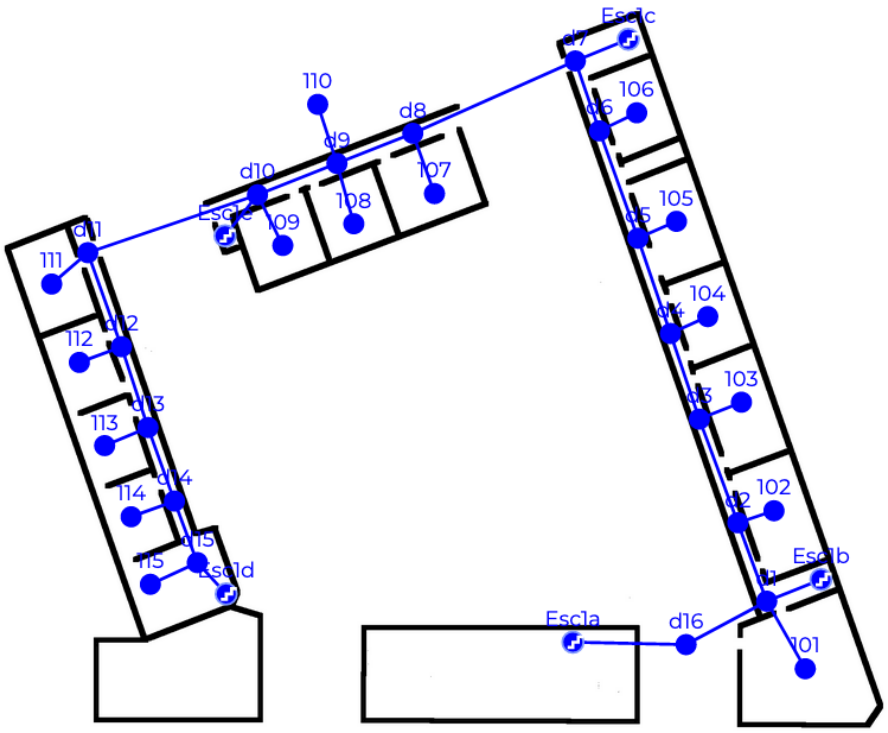


Figure 9.2: Le graphe symbolique d'un établissement (en bleu) superposé à une image d'un plan (en noir). Les nœuds représentent des salles, des couloirs ou des escaliers. Les arêtes sont les passages possibles entre ces lieux.

9.2 Entrées

On présente ici les entrées de l'application web. Deux données doivent être renseignées : les plans de l'école et l'emploi du temps complet. On peut également entrer des paramètres d'épidémiologie mais des paramètres par défaut sont proposés.

9.2.1 Graphe de l'école

La première donnée requise pour le programme est un graphe symbolique de l'établissement, dont les nœuds représentent les salles, les escaliers et les couloirs, et les arêtes un passage possible entre deux nœuds. Pour ce faire, on commence par télécharger un plan 2D de l'école (on peut charger plusieurs plans s'il y a plusieurs étages), comme représenté sur la figure 9.2. Une interface est alors présente pour superposer au plan un graphe symbolique que l'on construit à la souris. On peut choisir pour chaque nœud son type (classe, couloir, etc.) : des paramètres d'épidémiologie par défaut lui sont alors attribués. On peut modifier au besoin ces paramètres par défaut si l'on a une information supplémentaire sur le nœud considéré.

Name	Class	Regime	Option 1	Option 2	Groups
STUDENT2	6E	Half-boarder	ENGLISH		GR1
STUDENT10	6E	Half-boarder	ENGLISH		GR1
STUDENT33	6E	Half-boarder	ENGLISH		GR1
STUDENT41	6E	Extern	ENGLISH		GR1
STUDENT49	6E	Half-boarder	ENGLISH		GR1
STUDENT50	6E	Half-boarder	ENGLISH		GR1
STUDENT78	6E	Half-boarder	ENGLISH		GR1
STUDENT94	6E	Half-boarder	ENGLISH		GR1
STUDENT98	6E	Extern	ENGLISH		GR1
STUDENT107	6E	Half-boarder	ENGLISH		GR1
STUDENT164	6E	Half-boarder	GERMAN	ENGLISH	GR2
STUDENT168	6E	Half-boarder	GERMAN	ENGLISH	GR2
STUDENT195	6E	Half-boarder	GERMAN	ENGLISH	GR2
STUDENT196	6E	Half-boarder	GERMAN	ENGLISH	GR2
STUDENT199	6E	Half-boarder	GERMAN	ENGLISH	GR2
STUDENT205	6E	Half-boarder	GERMAN	ENGLISH	GR2
STUDENT217	6E	Half-boarder	GERMAN	ENGLISH	GR2
STUDENT222	6E	Extern	GERMAN	ENGLISH	GR2
STUDENT225	6E	Half-boarder	GERMAN	ENGLISH	GR2
STUDENT226	6E	Half-boarder	GERMAN	ENGLISH	GR2
STUDENT231	6E	Half-boarder	GERMAN	ENGLISH	GR2
STUDENT235	6E	Extern	GERMAN	ENGLISH	GR2
STUDENT238	6E	Half-boarder	GERMAN	ENGLISH	GR2
STUDENT239	6E	Half-boarder	GERMAN	ENGLISH	GR2
STUDENT240	6E	Half-boarder	GERMAN	ENGLISH	GR2
STUDENT242	6E	Half-boarder	GERMAN	ENGLISH	GR2
STUDENT244	6E	Half-boarder	GERMAN	ENGLISH	GR2
STUDENT453	6E	Half-boarder	ENGLISH		GR2

Duration	Day and time	Teacher	Field	Class	Room
1h00	Monday 09:15	Bob	Biology	3C	room1
1h00	Monday 10:15	Bob	Biology	3A	room1
1h00	Monday 14:00	Bob	Biology	3D	room1
1h00	Tuesday 08:15	Bob	Biology	4D	room1
1h00	Tuesday 10:15	Bob	Biology	3C	room2
1h00	Tuesday 11:15	Bob	Biology	5B	room3
1h00	Wednesday 08:15	Bob	Biology	3D	room6
1h00	Wednesday 08:15	Damien	Biology	4C	room4
1h00	Wednesday 09:15	Bob	Biology	4B	room5
1h00	Wednesday 10:15	Elena	Biology	5A	room2
1h00	Wednesday 11:15	Bob	Biology	4E	room3
1h00	Thursday 08:15	Bob	Biology	3D	room3
1h00	Thursday 08:15	Carole	Physics	3A	room2
1h00	Thursday 09:15	Bob	Biology	3C	room1
1h00	Thursday 10:15	Bob	Biology	4D	room6
1h00	Thursday 14:00	Bob	Biology	5B	library
1h00	Thursday 15:00	Bob	Biology	3A	library
1h00	Monday 15:00	Carole	Physics	4E GR1, 4E GR2	room2
1h00	Monday 15:00	Damien	Biology	4E GR1	room4
1h00	Monday 16:00	Damien	Biology	4E GR2	room3
1h00	Monday 10:15	Elena	English	4CD ANG2	room2
1h00	Monday 10:15	Francis	Spanish	4BD ESP2	room2

Figure 9.3: Un emploi du temps typique. À gauche, la feuille renseignant les données sur les élèves. À chaque élève correspond une classe et un ensemble d'options. À droite, l'emploi du temps, avec la durée de chaque cours, sa date, son enseignant·e, son lieu.

Des connexions entre niveaux différents sont prises en compte s'il y a des escaliers.

9.2.2 Emploi du temps

Le programme requiert également un emploi du temps, sous la forme d'un tableur au format *.xls*. Ce tableur doit comporter deux feuilles : une feuille contenant les métadonnées sur la population, c'est-à-dire la liste des étudiant·e·s et de leurs particularités (leurs options, leur régime et leur classe), afin de partitionner la population en communautés restant toujours ensemble ; une feuille contenant l'emploi du temps. Cet emploi du temps est typiquement l'export d'un des logiciels d'aide à la conception d'emploi du temps utilisés en France (UnDeuxTemps, Pronote, EDT, ...). Il doit renseigner pour chaque cours la date, le·la professeur·e, la classe et la salle. Différentes orthographes sont prises en compte pour le nom de la classe, la seule contrainte étant la correspondance entre les deux feuilles du tableur, et entre les noms de salle de l'emploi du temps et du graphe. Un exemple d'un tel emploi du temps est présenté sur la figure 9.3.

9.2.3 Paramètres

L'utilisateur.trice peut spécifier ses propres paramètres avant de lancer le calcul, même si des valeurs par défaut sont proposées. L'intégralité des paramètres du modèle est listée dans la table 9.1. Ces paramètres sont divisés en plusieurs catégories. Certains paramètres servent au calcul des trajectoires et à la détection des contacts ; d'autres paramètres sont des paramètres de modélisation servant à traduire les événements de croisement en termes de contacts. Les paramètres restants sont les paramètres d'épidémiologie. On détaillera le sens et le choix des paramètres dans la section 9.5.

Étape	Paramètre	Signification (unité)	Valeur par défaut
Calcul des trajectoires	dt_{mot}	Pas de temps des déplacements (s)	2
	$dt_{contacts}$	Pas de temps des contacts (s)	20
	$thresh_{contacts}$	Distance maximale de contact (m)	10
	v	Vitesse (m/s)	1
Contacts	Affinité classe1 - classe1	Affinité entre étudiant·e·s d'une même classe (entre 0 et 1)	0.7
	Affinité classe1 - classe2	Affinité entre étudiant·e·s de différentes classes (entre 0 et 1)	0.3
	Pondération	Pondération des contacts du lieu (entre 0 et 1)	1
	Nombre de contacts	Nombre de contacts par personne lorsqu'il y a un croisement	1 à 3
Épidémiologie	nb_{impact}	Durée avant qu'une infection soit détectée	5
	$rate_{air}$	Taux d'élimination d'un aérosol (/h)	6
	$rate_{surf}$	Taux d'élimination des dépôts surfaciques (/h)	0.5
	dep_{air}	Taux d'émission des aérosols	1
	dep_{surf}	Taux d'émission des dépôts surfaciques	1
	abs_{air}	Coefficient d'absorptionon des aérosols	1
	abs_{surf}	Coefficient d'absorptionon des dépôts surfaciques	1
	T_{inc}	Temps d'incubation (j)	5
	gamma	Taux de guérison + mort (/j)	0.2
	maxtime	Temps maximal de contagiosité des dépôts (h)	6
	epsilon	Nombre initial d'infecté·e·s dans la population	1
	$direct_{prob}$	Probabilité d'une infection par contact direct de 20 secondes	0.0003
	$proba_{air}$	Probabilité d'infection par aérosol si deux personnes se succèdent dans une pièce 10min chacune	0.001
	$proba_{surf}$	Probabilité d'infection par surface si deux personnes se succèdent dans une pièce 1h chacune	0

Table 9.1: Paramètres de CroCo.

9.3 Fonctionnement de l'algorithme

Détaillons le coeur de calcul, implémenté en python. Il est divisé en 13 étapes :

- Étape 1 : lecture du fichier *.xls* d'emploi du temps
- Étape 2 : analyse de la liste des élèves
- Étape 3 : analyse de la liste des cours
- Étape 4 : création des populations (groupes d'élèves, professeurs)
- Étape 5 : lecture des plans et création d'un graphe avec la librairie NetworkX
- Étape 6 : calcul des trajectoires sur le graphe
- Étape 7 : calcul des historiques de présence de chaque particule aux nœuds/arêtes
- Étape 8 : Calcul des précontacts (occurrence simultanée de deux particules au même nœud/arête)
- Étape 9 : calcul des cartes de chaleur des précontacts
- Étape 10 : calcul des précontacts indirects (succession des particules au même nœud/arête)
- Étape 11 : calcul des contacts directs
- Étape 12 : calcul des contacts indirects
- Étape 13 : calcul du modèle SEIR et du score.

9.3.1 Parsage de l'emploi du temps et des plans (étapes 1,2,3,4 et 5)

Ces étapes d'une grande technicité ont pour but de convertir les entrées en données exploitables en python. On construit un objet graphe à l'aide de la librairie python *NetworkX* qui porte les attributs de la géométrie, ainsi que les paramètres de la table 9.1 qui varient en fonction du lieu. La principale difficulté est le parsage des emplois du temps - à l'orthographe variable en fonction de l'emploi du temps - en tableaux de données de la librairie *pandas*. On reconstitue ensuite à l'aide des options les communautés (sous-ensemble des classes restant ensemble au cours de l'emploi du temps).
À l'issue de cette étape, on dispose du graphe, de la liste des cours et de la population sous format exploitable en python.

9.3.2 Calcul des chemins et des précontacts (étapes 6,7,8,9 et 10)

Les calculs commencent avec la reconstitution de chaque trajectoire. Pour chaque changement de salle de l'emploi du temps on calcule le déplacement dans le graphe de la particule considérée en cherchant le plus court chemin entre les nœuds considérés. On calcule alors le chemin précis (en espace et en temps) en faisant se déplacer la particule à vitesse constante le long du graphe en discrétisant avec un pas de temps $\mathrm{dt}_{\mathrm{mot}}$. On dispose donc de l'ensemble des positions successives de chaque particule au cours de l'emploi du temps. On attribue réciproquement à chaque nœud/arête l'information des particules qui y séjournent afin de tracer des cartes de chaleur de présence.
On calcule ensuite les précontacts directs : on détecte les occurrences de particules à distance inférieure à $\mathrm{thresh}_{\mathrm{contacts}}$. On exporte alors les cartes de chaleur (voir section 9.4). On détecte enfin les précontacts indirects, c'est-à-dire la succession de particules au même élément du graphe.

Remarque 37. *Ces précontacts directs et indirects posent des problèmes de mémoire, de grands DataFrame pouvant être générés à cette étape. Pour un établissement typique, le programme doit tourner sur une mémoire vive de 32 ou 64 Gbit.*

9.3.3 Calcul des contacts (étapes 11 et 12)

On convertit ensuite les précontacts en terme de contacts directs ou indirects. La première étape est de convertir les précontacts en événements de croisement : il faut identifier ces événements (notamment leur début et leur fin) à partir des données de cooccurrence à certains éléments du graphe. On prend en compte les contacts d'une particule avec elle même (il y a des contacts entre éléments d'un particule toute la journée). On associe ensuite à chaque événement de contact un nombre de contacts par personne et une pondération (paramètre "Number of contacts" et "Weighting" dans la table 9.1). Ce nombre de contact par personne est le coefficient $\overline{K_{XY}}$ dans les chapitres 7 et 8 : il représente le nombre constant de contacts qu'a un individu de la particule X avec des éléments de Y lors d'un

croisement. La pondération représente la dangerosité du lieu, qui peut varier suivant que l'on se trouve en extérieur, à la cantine ou encore en salle de classe. On traite différemment les contacts professeur/élève : on suppose que lors d'un cours un professeur est en contact avec chaque élève mais avec une faible pondération pour ne pas donner trop d'importance à ces contacts. On génère à cette étape une matrice de contacts (voir la section 9.4).
On ne convertit pas les précontacts indirects en de contacts réels, mais on veut calculer les probabilités de transmission pour chacun de ces précontacts. En effet, afin d'utiliser le modèle SIR (7.6), on doit calculer la probabililité d'infection d'un individu de la deuxième particule s'il y a un·e unique infecté·e dans la première. En notant λ_c le taux de contamination par unité de concentration du dépôt, on peut exprimer la probabilité d'infection pendant dt si la concentration du dépôt est $D(t)$:

$$\mathbb{P}(\text{Infection entre } t \text{ et } t+dt | \text{L'individu est sain à } t) = \lambda_c D(t). \tag{9.1}$$

En notant X la date de l'infection, on peut réécrire l'expression précédente

$$\frac{\mathbb{P}(X \in [t, t+dt])}{\mathbb{P}(X \geqslant t)}. \tag{9.2}$$

En notant $F_X(t)$ la fonction de répartition de X, on obtient

$$\frac{\mathrm{d}F_X}{\mathrm{d}t} = \lambda_c D(t)(1 - F_X(t)). \tag{9.3}$$

On peut calculer $D(t)$ avec les hypothèses du modèle : on dispose des taux d'emission et d'élimination des dépôts (voir table 9.1) et des dates de présence de la première particule. On peut alors calculer la probabilité d'infection lors de l'événement, c'est-à-dire $F_X(t_f)$, où t_f est la date de fin du contact indirect.

9.3.4 Épidémiologie (étape 13)

On connaît à cette étape l'ensemble des contacts directs et indirects. Étant donné un contact direct de temps t et de pondération w, la probabilité d'infection lors de celui-ci s'écrit

$$p = 1 - e^{-tw\lambda} \tag{9.4}$$

où λ est le taux instantané d'infection lors d'un contact, pouvant être calculé à partir de la probabilité d'infection lors d'un contact direct $\text{direct}_{\text{proba}}$ (voir table 9.1). Pour chaque particule, on fait tourner un modèle SEIR du type (7.6), avec pour condition initiale une population saine à l'exception d'une infection dans la particule considérée. On itère le modèle pendant $\text{nb}_{\text{impact}}$ jours (période à l'issue de laquelle une épidémie est détectée). On obtient l'impact de chaque particule sur l'établissement si l'infection initiale part de cette particule. On exporte à cette étape un score (la moyenne du nombre d'infecté·e·s en faisant varier la particule initiale infectée) et une matrice d'infection.

9.4 Visuels

Dans cette section on détaille les visuels générés par notre algorithme CroCo.

9.4.1 Score

On définit le score associé à un emploi du temps comme l'ampleur moyenne de l'infection si un·e infecté·e pénètre l'établissement. En notant D_Y le nombre d'infecté·e·s total à l'issue de la période de propagation lorsque la particule Y est initialement contaminée, le score s'écrit

$$\frac{\sum_{Y\in\mathcal{P}} N_Y D_Y}{\sum_{Y\in\mathcal{P}} N_Y}. \tag{9.5}$$

Bien que formulé comme un nombre d'infecté·e·s moyen dans la population, ce score n'a que peu de sens absolu à cause de la variabilité des paramètres et de la complexité du phénomène modélisé. Il est plus raisonnable d'interpréter la différence relative entre deux scores d'emplois du temps afin de prendre une décision plutôt que de chercher à lire ce score comme une prédiction du nombre effectif de malades dans l'école lors de l'épidémie.

9.4.2 Cartes de chaleur

On affiche ensuite deux types de cartes de chaleur pour visualiser les zones à risque. Sur la gauche de la figure 9.4.2, on représente la carte de chaleur globale de l'ensemble des contacts sur un exemple jouet. Comme attendu, la majorité des contacts ont lieu dans les classes, lieux où les élèves passent le plus de temps. Les discontinuités sont dues à la discrétisation en espace des trajectoires. Des cartes de présence personnalisées sont également disponibles, comme sur la droite de la figure 9.4.2.

9.4.3 Matrices

On génère trois types de matrices pour analyser la structure de l'épidémie au sein de la population : une matrice de contacts directs, deux matrices de contacts indirects, et une matrice d'infection.
La matrice de contacts directs a pour entrée i, j

$$M_{ij} = \sum_{\text{contacts } X_i\sim X_j} \frac{\overline{K_c}}{N_{X_j}}(1 - e^{-wt\lambda}), \tag{9.6}$$

où l'on a sommé sur chaque événement de croisement entre les particules, où $\overline{K_c}$ est le nombre total de contacts entre les particules lors du croisement, w la pondération, t le temps de croisement, λ le taux d'infection instantané et N_{X_j} le cardinal de X_j. Cette formule est construite sur des heuristiques pour établir un visuel monotone en le nombre de contacts. Elle dérive du constat suivant : s'il y a très peu d'infecté·e·s dans le système, le vecteur des

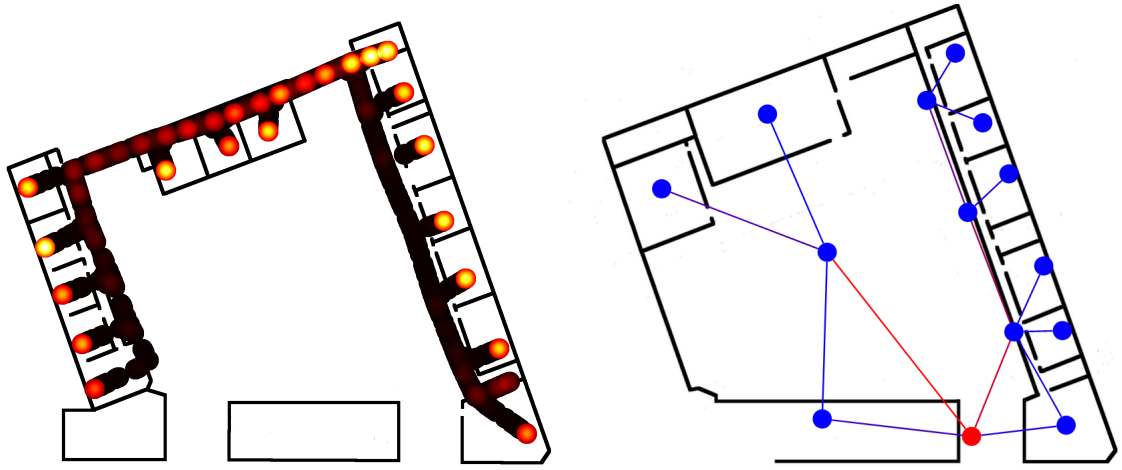


Figure 9.4: Cartes de chaleur générées par Croco. À gauche, la carte des contacts globaux. À droite, la carte de présence d'une particule sur le graphe. Les couleurs chaudes correspondent à un contact ou une présence fréquente.

infections vérifie $I^{n+1} = M^n I^n$, où M^n est définie par (9.6) en sommant sur les contacts du jour n. Dans la formule (9.6), on ne discrimine pas en fonction du jour de contact, pour ne pas favoriser un certain jour de l'emploi du temps et pour avoir un nombre unidimensionnel pour chaque paire de particules. La matrice est symétrique, car le nombre de contact est le nombre total de contacts entre X_i et X_j réparti uniformément entre chaque individu de X_i. On construit ensuite deux matrices de contacts indirects correspondant aux deux modes d'infection considérés (par aérosols, par dépôt surfacique). Leurs entrées en position (i, j), correspondant à l'impact de X_j sur X_i sont calculées à l'aide de la formule

$$1 - \prod_{\text{Croisement } X_j \to X_j} (1 - p_c) \tag{9.7}$$

où l'on prend le produit (la somme probabiliste) des événements où X_i succède à X_j, et où p_c est la probabilité d'une infection pour un individu de X_i s'il y a un·e infecté·e dans X_j lors de l'événement de contact indirect. On prend le produit sur l'ensemble des croisements dans tout l'emploi du temps : la situation revient à étudier la possibilité d'une infection indirecte de X_j à X_i en ne tenant compte d'aucune chaîne d'infection secondaire.
On génère enfin une matrice d'infection, dont la case en position (i, j) contient le nombre d'infecté·e·s dans X_j au bout de la période considérée si l'infection part de X_i. De telles matrices sont représentées sur la figure 9.5. En particulier, on note que l'impact d'une particule se lit en colonne. Le score est alors la moyenne des normes 1 des colonnes (pondérées par les effectifs des particules).

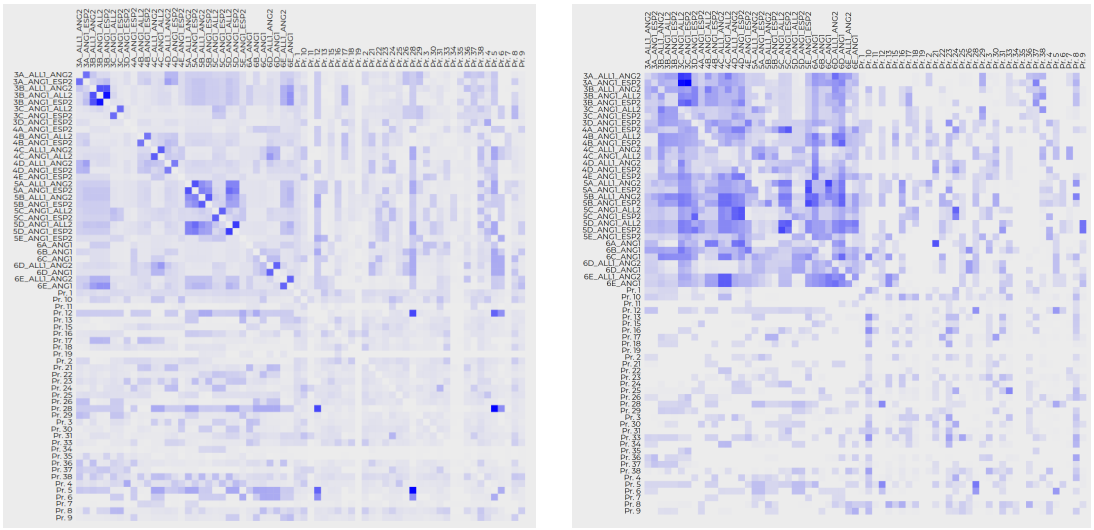


Figure 9.5: À gauche, la matrice des contacts directs, calculée avec (9.6). À droite, la matrice des contacts indirects.

9.4.4 Risque, dangerosité et score

On fournit enfin des visuels purement épidémiologiques dérivés des calculs de propagation d'infection de l'algorithme. Conformément au chapitre 7, on définit le risque et la dangerosité d'une particule à l'aide de la matrice d'infection introduite ci-dessus. En notant P cette matrice, on a les formules :

$$\begin{aligned} \textbf{risque}(X_i) &= \frac{\sum_{X_j} N_j M_{ij}}{\sum_{X_j} N_j} \\ \textbf{dangerosite}(X_j) &= \frac{\sum_{X_i} N_i M_{ij}}{\sum_{X_i} N_i}. \end{aligned} \tag{9.8}$$

En d'autres termes, le risque d'une particule est la fraction moyenne de la particule infectée si l'infection initiale part de quelqu'un tiré aléatoirement dans l'établissement ; la dangerosité d'une particule est l'ampleur de l'infection dans l'école si celle-ci part de la particule en question. On affiche également le score total de risque, moyenne pondérée des risques ou des dangerosités précédemment définis. Ces visuels sont représentés sur la figure 9.6 pour la comparaison de deux emplois du temps. L'un a un score deux fois plus élevé que l'autre : on choisira donc sur un plan purement épidémiologique cet emploi du temps de score minimal.

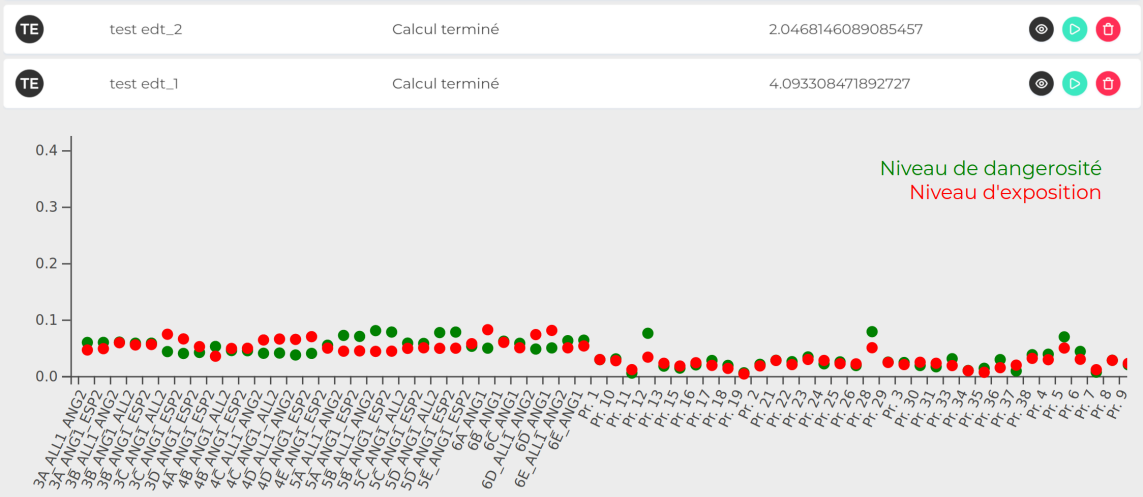


Figure 9.6: En haut, le score de deux emplois du temps sur un lycée pilote. En bas, le risque (ou exposition) et la dangerosité de chaque particule sur l'un de ces emplois du temps.

9.5 Étude paramétrique

Le calibrage des paramètres de calcul des trajectoires de la table 9.1 ne pose pas de problème particulier. On prend ainsi des pas de temps les plus petits possibles pour avoir une description la plus précise possible des déplacements - ce qui fait juste augmenter la mémoire et le temps de calcul. Une vitesse de $1ms^{-1}$ et un seuil de contact de $10m$ apparaissent réalistes (si deux particules sont à moins de $10m$, elles sont susceptibles d'entrer en contact). Par ailleurs, les paramètres de contact sont des paramètres de modélisation et peuvent donc être fixés arbitrairement. Côté épidémiologie, le nombre de jours avant détection de l'épidémie, le taux de rémission ou les taux d'élimination des dépôts peuvent être trouvés aisément dans la littérature. On a laissé des coefficients d'émission et des coefficients d'absorption neutres, c'est-à-dire par défaut égaux à 1, mais qui peuvent être augmentés ou diminués en fonction de la maladie et du lieu considéré. Comme pointé dans la section 9.5, les paramètres de taux d'infection lors de contacts directs ou indirects sont difficiles à trouver dans la littérature. On a donc décidé d'opter pour une démarche inverse : on a cherché à fixer ces valeurs afin d'obtenir des scores réalistes sur des établissements tests contactés par MODCOV19. On a fixé le paramètre d'infection par contact surfacique à 0, au vu du peu de contaminations par ce mode lors de l'épidémie de COVID-19. Pour un dialogue plus aisé avec la littérature spécialisée en épidémiologie, on a remplacé ces taux par des probabilités d'infection lors d'une situation fixée : p_c la probabilité d'infection lors d'un contact de 20 secondes pour les contacts directs, et p_a la probabilité d'infection lorsque deux individus se succèdent dans une pièce standard pendant 10 minutes chacun. Ces probabilités sont en effet rencontrées plus fréquemment en épidémiologie ; il y a en outre

une correspondance entre probabilité et taux et l'on peut aisément convertir l'un en l'une et réciproquement. En cherchant à obtenir un score de 2% en prenant en compte les contacts directs uniquement, et de 1% en prenant en compte seulement les contacts indirects, on obtient $p_c = 0.0003$, et $p_a = 0.003$, pour un risque joint de $4,21\%$. En comparaison, p_c est usuellement de 0.001 pour la grippe (voir [68]) et a été estimée à 0.0004 pour le COVID-19 par une méthode inverse, voir [65].
On rappelle que l'estimation des paramètres, si elle est nécessaire pour un affinage des résultats et un modèle plus réaliste, n'est pas absolument nécessaire ici. En effet, notre modèle n'a pas vocation à prédire un nombre réel d'infecté·e·s dans l'établissement, mais à afficher un score de risque subjectif permettant de comparer des emplois du temps. Le modèle et son implémentation ont été développés pour avoir un comportement monotone en fonction des des situations à risque. Nous pensons que dans des situations pratiques, un écart relatif important entre deux emplois du temps est révélateur d'une situation effectivement plus risquée dans l'un des cas, et ce malgré une large indécision sur les paramètres du modèle. Nous avons laissé de nombreux paramètres accessibles dans le modèle pour permettre un affinage dans le cas où ces paramètres seraient bien connus et mesurés efficacement.

Annexe

Preuve du lemme 9

La démonstration du lemme suivant a été suggérée par Filippo Santambrogio.

Lemme 9. *Soit Ω un compact et μ_1, μ_2 deux mesures absolument continues telles que $\mu_1 \leqslant_* \mu_2$. Alors $P_{K_1}(\mu_1) \leqslant_* P_{K_1}(\mu_2)$.*

Démonstration. Soient $\mu_1 \leqslant_* \mu_2$ deux mesures absolument continues par rapport à la mesure de Lebesgue. Soit $f_n : \mathbb{R} \longrightarrow \mathbb{R}_+$ une suite de fonctions C^1 strictement convexes, telle que $f_n'(0) = f_n(0) = 0$ et $f_n'(\infty) = n$. On choisit f_n qui converge simplement vers $f = \infty \mathbf{1}_{]1,\infty[}$. Il est possible de prendre $\|f_n - g_n\|_\infty \leqslant \frac{1}{n}$, avec $g_n(x) = n(x-1)\mathbf{1}_{x>1}$, comme représenté sur la figure 9.7. On considère les problèmes d'optimisation suivants

$$\rho_j^n \in \operatorname*{argmin}_{\rho \in W_2(\Omega)} W_2^2(\rho, \mu_j) + \int_\Omega f_n(\rho(x))\,\mathrm{d}x. \tag{9.9}$$

avec la convention $+\infty$ si ρ n'a pas de densité. Les conditions d'optimalité s'écrivent

$$\varphi_j + f_n'(\rho_j^n) = c_j \tag{9.10}$$

où φ_j est un potentiel de Kantorovich de ρ_j^n à μ_j et c_j est une constante. On veut montrer $\rho_1^n \leqslant_* \rho_2^n$. Comme f_n' est strictement croissante, il suffit de montrer que $f_n'(\rho_n^1) \leqslant f_n'(\rho_n^2)$ p.p. Notons

$$m = \inf_{y \in \Omega} f_n'(\rho_2^n(y)) - f_n'(\rho_1^n(y)). \tag{9.11}$$

On peut écrire

$$m = \inf_{y \in \Omega} c_2 - c_1 + \varphi_1 - \varphi_2. \tag{9.12}$$

m est atteint en un $y_0 \in \Omega$ tel que :

$$\left\{ \begin{array}{rcl} \nabla\varphi_1(y_0) & = & \nabla\varphi_2(y_0) \\ \mathrm{Hess}(\varphi_1 - \varphi_2)_{|y_0} & \geqslant & 0 \end{array} \right. \tag{9.13}$$

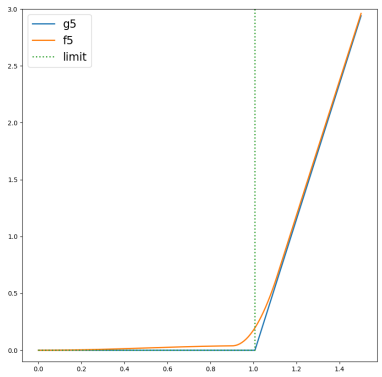


Figure 9.7: En bleu, la fonction g_5. En orange, son approximation par la fonction convexe lisse f_5. La limite commune à (f_n) et (g_n) est représentée en vert.

De plus pour $y \in \Omega$ on a

$$\rho_j^n(y) = |\det(I - D^2\varphi_j)_{|y}| \times \mu_j\left(y - \nabla\varphi_j(y)\right), \tag{9.14}$$

car le plan de transport de ρ_j^n à μ_j est de la forme $T(x) = x - \nabla\phi_j(x)$. Supposons que $\operatorname{infess}(\rho_2^n - \rho_1^n) < 0$. Alors $\operatorname{infess}(f_n'(\rho_2^n) - f_n'(\rho_1^n)) < 0$ donc $m < 0$. Par conséquent,

$$|\det(I - D^2\varphi_2)_{|y_0} \times |\mu_2\left(y_0 - \nabla\varphi_2(y_0)\right) < |\det(I - D^2\varphi_1)_{|y_0}| \times \mu_1\left(y_0 - \nabla\varphi_1(y_0)\right). \tag{9.15}$$

D'une part, on a

$$\begin{aligned} \mu_2(y_0 - \nabla\varphi_2(y_0)) &= \mu_2(y_0 - \nabla\varphi_1(y_0)) \\ &\geqslant \mu_1(y_0 - \nabla\varphi_1(y_0)). \end{aligned} \tag{9.16}$$

Par ailleurs, notons $A = Id - D^2\varphi_2(y_0)$, $B = Id - D^2\varphi_1(y_0)$, $a = \det(A)$, $b = \det(B)$. A et B sont les matrices hessiennes de fonctions convexes donc sont dans $S_2^+(\mathbb{R})$. Soit

$$\gamma \left\{ \begin{array}{ccc} [0,1] & \longrightarrow & \mathbb{R} \\ t & \longmapsto & \det(B + t(A - B)). \end{array} \right. \tag{9.17}$$

La dérivée de γ peut être exprimée à l'aide de la comatrice de B :

$$\gamma'(t) = \operatorname{Tr}({}^T\operatorname{Com}(B)(A - B)). \tag{9.18}$$

Elle est positive d'après (9.13). On obtient $a \geqslant b$, ce qui est absurde avec (9.15) et (9.16). On a donc que $\rho_1^n \leqslant \rho_2^n$ p.p.
Montrons à présent la convergence des ρ_j^n. Par optimalité de ρ_j^n, on a

$$\begin{aligned} W_2^2(\rho_j^n, \mu_j) &\leqslant W_2^2(\rho_j^n, \mu_j) + \int_\Omega f_n(\rho_j^n) \\ &\leqslant W_2^2(\rho_j, \mu_j) + \int_\Omega f_n(\rho_j) \end{aligned} \tag{9.19}$$

où $\rho_j \in \underset{W_2(\Omega)}{\operatorname{argmin}} W_2^2(\rho_j, \mu_j) + \displaystyle\int_\Omega f(\rho_j)$.
En comparant f_n et g_n (qui tendent en croissant vers f), on obtient :

$$W_2^2(\rho_j^n, \mu_j) \leqslant W_2^2(\rho_j, \mu_j) + \int_\Omega f(\rho_j) + \frac{|\Omega|}{n}. \tag{9.20}$$

Soit $\tilde{\rho}_j$ une valeur d'adhérence de ρ_j^n. $\tilde{\rho}_j$ satisfait :

$$W_2^2(\tilde{\rho}_j, \mu_j) \leqslant W_2^2(\rho_j, \mu_j). \tag{9.21}$$

Montrons que $\tilde{\rho}_j \in K_1$. On aura alors par optimalité $\tilde{\rho}_j = \rho_j$ et donc

$$\rho_1 = \lim_{n\to\infty} \rho_1^n \leqslant_* \lim_{n\to\infty} \rho_2^n = \rho_2. \tag{9.22}$$

Soit $\epsilon > 0$ et $A = \{\tilde{\rho}_j > 1 + \epsilon\}$. Supposons $\lambda(A) > 0$. On a

$$\limsup_{n\to\infty} \rho_j^n(A) \geqslant \tilde{\rho}_j(A) \geqslant (1+\epsilon)\lambda(A). \tag{9.23}$$

On peut supposer, quitte à extraire

$$\int_\Omega \rho_j^n(x)\,\mathrm{d}x \geqslant \left(1 + \frac{\epsilon}{2}\right)\lambda(A). \tag{9.24}$$

Soit $A_n = A \cap \left\{\rho_j^n > 1\right\}$, on a

$$\left(1 + \frac{\epsilon}{2}\right)\lambda(A) \leqslant \int_{A_n} \rho_j^n(x)\,\mathrm{d}x + \int_{A\setminus A_n} 1\,\mathrm{d}x. \tag{9.25}$$

Pour tout $y > 1$, $y \leqslant \dfrac{f_n(y)}{n} + 1$ donc

$$\left(1 + \frac{\epsilon}{2}\right)\lambda(A) \leqslant \int_{A_n} \frac{f_n(\rho_j^n(x))}{n}\,\mathrm{d}x + \lambda(A_n) + \lambda\left(A\setminus A_n\right). \tag{9.26}$$

On obtient

$$\begin{aligned} \frac{n\epsilon}{2}\lambda(A) &\leqslant \int_{A_n} f_n(\rho_j^n(x))\,\mathrm{d}x \\ &\leqslant \int_\Omega f_n(\rho_j^n(x))\,\mathrm{d}x + W_2^2(\rho_j^n, \mu_j). \end{aligned} \tag{9.27}$$

Soit χ une mesure de probabilité dont la densité est majorée par 1 presque-partout. Par optimalité, on a

$$\begin{aligned} \frac{n\epsilon}{2}\lambda(A) &\leqslant \int_\Omega f_n(\chi(x))\,\mathrm{d}x + W_2^2(\chi, \mu_j) \\ &\leqslant \frac{\lambda(\Omega)}{n} + W_2^2(\chi, \mu_j) \end{aligned} \tag{9.28}$$

ce qui est absurde à la limite $n \to \infty$. On a donc $\lambda(A) = 0$: $\tilde{\rho}_j$ est admissible. □

Démonstration des lemmes 10 et 11.

Lemme 10. *Soit $u \in \mathbb{H}^1(\Omega)$, $1 \leqslant i \leqslant N-2$. Alors il existe c une constante ne dépendant que de u telle que*

$$\int_{L_i}\int_{L_i} |u(s)-u(t)|^2 \,\mathrm{d}s\,\mathrm{d}t \leqslant \frac{c}{N^3} \tag{4.20}$$

Si de plus $\lambda\left(L_i \cap \{u<0\}\right) > 0$, alors

$$\int_{L_i} u_+^2(s)\,\mathrm{d}s \leqslant \frac{c}{N^2}. \tag{4.21}$$

Démonstration. On se contente de montrer le lemme pour $i = 1, \dots, N-2$. On se donne $u \in C_c^\infty(\Omega)$. Montrons d'abord le deuxième point : u s'annule en un $x_0 \in J_i = \left[\frac{i-\frac{1}{2}}{N}, \frac{i+\frac{1}{2}}{N}\right]$. On a

$$u(x) = \int_0^1 \nabla u(x_0 + t(x-x_0)) \cdot (x-x_0)\,\mathrm{d}t \tag{9.29}$$

donc

$$\begin{aligned}
\int_{J_i} u^2(x)\,\mathrm{d}x &\leqslant \int_0^1 \int_{J_i} |x-x_0|^2 |\nabla u(x_0+t(x-x_0))|^2 \,\mathrm{d}x\,\mathrm{d}t \\
&= \int_{J_i} |x-x_0|^2 \int_{x_0}^x \frac{|\nabla u(s)|^2}{|x-x_0|}\,\mathrm{d}s\,\mathrm{d}x \\
&\leqslant \|\nabla u\|_2^2 \int_{J_i} |x-x_0|\,\mathrm{d}x \\
&\leqslant \frac{c}{N^2}.
\end{aligned} \tag{9.30}$$

Pour le premier point, on écrit en utilisant le même procédé :

$$\begin{aligned}
\int_{J_i}\int_{J_i} |u(x)-u(y)|^2 \,\mathrm{d}x\,\mathrm{d}y &\leqslant \int_{J_i}\int_{J_i}\int_0^1 |\nabla u(x+t(y-x))|^2 |x-y|^2 \,\mathrm{d}t\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_{J_i}\int_{J_i} |x-y|^2 \int_x^y \frac{|\nabla u(s)|^2}{|x-y|}\,\mathrm{d}s\,\mathrm{d}x\,\mathrm{d}y \\
&\leqslant \frac{c}{N^3}.
\end{aligned} \tag{9.31}$$

On démontre le cas général dans $H^1(\Omega)$ en approximant par des fonctions C_c^∞. □

Lemme 11. *Soit $\phi \in C^1([0,1]^2)$. Il existe une constante $c>0$ telle que pour tous N, i et j*

$$\int_{M_{i,j}} \left|\phi(x,y) - N\int_{\Gamma_l} \phi(t)\,\mathrm{d}t\right|^2 \mathrm{d}x\,\mathrm{d}t \leqslant \frac{c}{N^2}\int_{M_{i,j}} |\nabla\phi|^2. \tag{4.44}$$

Démonstration. Montrons le lemme pour $i = j = 0$ pour alléger les notations. On a

$$\begin{aligned}\int_{M_{i,j}} \left|\phi(x,y) - N\int_0^{\frac{1}{N}} \phi(0,u)\,\mathrm{d}u\right|^2 \mathrm{d}x\,\mathrm{d}y &= \int_{M_{i,j}} \left|N\int_{\Gamma_l} \phi(x,y) - \phi(t)\,\mathrm{d}t\right|^2 \mathrm{d}x\,\mathrm{d}y \\ &\leqslant 2\int_{M_{i,j}} \left|N\int_0^{\frac{1}{N}} \phi(x,y) - \phi(x,u)\,\mathrm{d}u\right|^2 \mathrm{d}x\,\mathrm{d}y \\ &\quad + 2\int_{M_{i,j}} \left|N\int_0^{\frac{1}{N}} \phi(x,u) - \phi(0,u)\,\mathrm{d}u\right|^2 \mathrm{d}x\,\mathrm{d}y.\end{aligned} \tag{9.32}$$

On peut alors écrire

$$\begin{aligned}\phi(x,y) - \phi(x,u) &= \int_0^1 (y-u)\partial_y\phi(x, sy + (1-s)u)\,\mathrm{d}s \\ \phi(x,u) - \phi(0,u) &= \int_0^1 x\partial_x\phi(sx,u)\,\mathrm{d}s\end{aligned} \tag{9.33}$$

et utiliser l'inégalité de Jensen pour dominer l'estimée par

$$\begin{aligned}2N\int_{M_{i,j}}\int_0^{\frac{1}{N}}\int_0^1 (y-u)^2\partial_y\phi(x, sy+(1-s)u)^2\,\mathrm{d}s\,\mathrm{d}u\,\mathrm{d}x\,\mathrm{d}y \\ + 2N\int_{M_{i,j}}\int_0^{\frac{1}{N}}\int_0^1 x^2\partial_x\phi(sx,u)\,\mathrm{d}s\,\mathrm{d}u\,\mathrm{d}x\,\mathrm{d}y.\end{aligned} \tag{9.34}$$

En posant

$$\begin{aligned}&\psi(s,u,x,y) = (x, sy + (1-s)u)) \\ &\zeta(s,u,x,y) = (sx,u) \\ &d\mu = (y-u)^2\,\mathrm{d}s\,\mathrm{d}u\,\mathrm{d}x\,\mathrm{d}y \\ &d\nu = x^2\,\mathrm{d}s\,\mathrm{d}u\,\mathrm{d}x\,\mathrm{d}y,\end{aligned} \tag{9.35}$$

on peut réécrire (9.34)

$$2N\left(\int_{M_{i,j}}\int_0^{\frac{1}{N}}\int_0^1 \partial_y\phi(v,w)^2\psi_*\mu(\mathrm{d}v,\mathrm{d}w) + \int_{M_{i,j}}\int_0^{\frac{1}{N}}\int_0^1 \partial_x\phi(v,w)\zeta_*\nu(\mathrm{d}v,\mathrm{d}w)\right). \tag{9.36}$$

Estimons à présent $\psi_*\mu([a,b]\times[c,d])$ pour tout rectangle $[a,b]\times[c,d]\subset M_{i,j}$:

$$\begin{aligned}
\psi_*\mu([a,b]\times[c,d]) &= (b-a)\int_0^c\int_c^d\int_0^{\frac{u-c}{u-y}}(u-y)^2\,\mathrm{d}s\,\mathrm{d}u\,\mathrm{d}y\\
&+(b-a)\int_0^c\int_d^1\int_{\frac{d-u}{y-u}}^{\frac{c-u}{u-y}}(u-y)^2\,\mathrm{d}s\,\mathrm{d}u\,\mathrm{d}y\\
&+(b-a)\int_c^d\int_0^c\int_{\frac{c-u}{y-u}}^1(u-y)^2\,\mathrm{d}s\,\mathrm{d}u\,\mathrm{d}y\\
&+(b-a)\int_c^d\int_c^d\int_0^1(u-y)^2\,\mathrm{d}s\,\mathrm{d}u\,\mathrm{d}y\\
&+(b-a)\int_c^d\int_d^1\int_{\frac{d-y}{u-y}}^1(u-y)^2\,\mathrm{d}s\,\mathrm{d}u\,\mathrm{d}y\\
&+(b-a)\int_d^1\int_c^d\int_0^{\frac{d-u}{y-u}}(u-y)^2\,\mathrm{d}s\,\mathrm{d}u\,\mathrm{d}y\\
&+(b-a)\int_d^1\int_0^a\int_{\frac{c-u}{y-u}}^{\frac{d-u}{y-u}}(u-y)^2\,\mathrm{d}s\,\mathrm{d}u\,\mathrm{d}y.
\end{aligned}\tag{9.37}$$

À l'exception du deuxième et du dernier terme, chaque terme comporte une intégrale de c à d. On peut donc directement les dominer par $\dfrac{c_1(b-a)(c-d)}{N^3}$ où $c_1>0$ est une constante. On peut cependant calculer

$$\int_0^c\int_d^1\int_{\frac{d-u}{y-u}}^{\frac{u-c}{u-y}}(u-y)^2\,\mathrm{d}s\,\mathrm{d}u\,\mathrm{d}y=\int_0^c\int_d^1(u-y)(d-c)\tag{9.38}$$

$$\int_d^1\int_0^a\int_{\frac{c-u}{y-u}}^{\frac{d-u}{y-u}}(u-y)^2\,\mathrm{d}s\,\mathrm{d}u\,\mathrm{d}y=\int_d^1\int_0^a(u-y)(d-c),\tag{9.39}$$

ce qui est suffisant pour montrer

$$\psi_*\mu([a,b]\times[c,d])\leqslant c_1\frac{\lambda([a,b]\times[c,d])}{N^3},\tag{9.40}$$

où λ est la mesure de Lebesgue. Un calcul similaire donne

$$\zeta_*\nu([a,b]\times[c,d])\leqslant c_2\frac{\lambda([a,b]\times[c,d])}{N^3},\tag{9.41}$$

avec $c_2>0$. On peut alors dominer (9.36) par

$$\frac{2c_1}{N^2}\int_{M_{i,j}}\partial_y\phi(v,w)\,\mathrm{d}v\,\mathrm{d}w+\frac{2c_2}{N^2}\int_{M_{i,j}}\partial_x\phi(v,w)\,\mathrm{d}v\,\mathrm{d}w\tag{9.42}$$

ce qui conclut la démonstration avec $c = \max(c_1, c_2)$.

□

Démonstration du théorème 3

Théorème 3. *Soit $\Omega = [0,1]$. Si $U \in \mathbb{H}^1(\Omega)$, la projection en norme $\|\cdot\|_2$ de U sur $C_\rho = \left\{u \in \mathbb{L}^2(\Omega),\, \nabla \cdot u \geqslant 0\right\}$ est dans $\mathbb{H}^1(\Omega)$.*

Démonstration. Pour rentrer dans les notations de [10], on pose $V = \mathbb{L}^2(\Omega)$, $W = \mathbb{H}^1(\Omega)$, et $\phi(x) = x$. La fonction de dualité associée à (W, W', ϕ) est :

$$J : \left\{ \begin{array}{ccc} W & \longrightarrow & W' \\ w & \longmapsto & \left\{ \begin{array}{ccc} W & \longrightarrow & \mathbb{R} \\ v & \longmapsto & \langle v,\, w\rangle_{\mathbb{H}^1(\Omega)} \end{array} \right. \end{array} \right. \tag{9.43}$$

Pour appliquer le théorème 1.1 de [10], on doit montrer que (u, C_ρ, Id_V) est J-compatible, avec $u = P_{C_\rho}(U)$. Pour $\epsilon > 0$, il faut trouver une solution $u_\epsilon \in C_\rho \cap W$ à $u_\epsilon + \epsilon J(u_\epsilon) = u$. L'équation est dans $\left(\mathbb{H}^1(\Omega)\right)'$:

$$\forall \phi \in \mathbb{H}^1(\Omega),\, \int_\Omega \phi(u - u_\epsilon) = \epsilon \int_\Omega \phi u_\epsilon + \epsilon \int_\Omega \nabla\phi \cdot \nabla u_\epsilon. \tag{9.44}$$

Par densité de $C^\infty(\overline{\Omega})$ dans $\mathbb{H}^1(\Omega)$, si $u_\epsilon \in \mathbb{H}^2(\Omega)$, il suffit de montrer

$$\left\{ \begin{array}{l} \forall \phi \in \mathbb{H}^1(\Omega),\, \displaystyle\int_\Omega \phi\left((u - u_\epsilon) - \epsilon u_\epsilon + \epsilon \Delta u_\epsilon\right) = 0 \\ u'_\epsilon(0) = u'_\epsilon(1). \end{array} \right. \tag{9.45}$$

Posons $\alpha = \dfrac{1+\epsilon}{\epsilon}$, soit $a, b \in \mathbb{R}$.

$$u_\epsilon^{a,b} : x \longmapsto e^{\sqrt{\alpha}x}\left(a + \int_0^x \frac{u(t)e^{-\sqrt{\alpha}t}}{2\epsilon\sqrt{\alpha}}\,\mathrm{d}t\right) + e^{-\sqrt{\alpha}x}\left(b + \int_0^x \frac{-u(t)e^{-\sqrt{\alpha}t}}{2\epsilon\sqrt{\alpha}}\,\mathrm{d}t\right) \tag{9.46}$$

fournit une famille de solutions $\mathbb{H}^2(\Omega)$ à $\Delta u_\epsilon = \alpha u_\epsilon - \frac{u}{\epsilon}$. On a $\mathbb{H}^2(\Omega) \subset W$, il reste à fixer a, b pour que $u_\epsilon^{a,b} \in C_\rho$. On va utiliser le lemme suivant :

Lemme 16. *Soit $u \in \mathbb{L}^2(\Omega)$, telle que $\nabla \cdot u \geqslant 0$ dans $\mathbb{H}^{-1}(\Omega)$. Alors $\forall x \in \Omega$, $\exists c \in \mathbb{R}$ t.q.*

- *$\forall t \geqslant x$, $u(t) \geqslant c$ presque sûrement,*
- *$\forall t \leqslant x$, $u(t) \leqslant c$ presque sûrement.*

Démonstration. On commence par montrer que u' définit une mesure. Soit $\varphi, \phi \in C_c^\infty(\Omega)$ telles que $\phi = 1$ sur le support de φ. Comme $\langle u', \phi\|\varphi\|_\infty - \varphi\rangle \leqslant 0$ et $\langle u', \phi\|\varphi\|_\infty + \varphi\rangle \leqslant 0$, on a

$$|\langle u', \varphi\rangle| \leqslant \|\varphi\|_\infty \langle u', \phi\rangle. \tag{9.47}$$

u' est donc une distribution d'ordre 0 : elle s'identifie à une mesure de Radon μ. Soit $x \in \Omega$, et

$$g : \left\{ \begin{array}{ccc} \mathbb{R} & \longrightarrow & \mathbb{R} \\ t & \longmapsto & \mu([x,t[). \end{array} \right. \tag{9.48}$$

g appartient à $\mathbb{L}^1_{loc}(\Omega)$ car elle est bornée sur tout compact. La dérivée au sens des distributions de g est u' pour $\phi \in C_c^\infty(\Omega)$, $\langle g, \phi'\rangle = -\langle u', \phi\rangle$.
Montrons à présent que $\chi = g - u$ est constante : χ est une distribution sur Ω de dérivée nulle. Soit $\theta, \phi \in C_c^\infty(\Omega)$, t.q. $\displaystyle\int_\Omega \theta(x)\,\mathrm{d}x = 1$. Alors

$$\phi - \theta \int_\Omega \theta(x)\,\mathrm{d}x \tag{9.49}$$

est d'intégrale nulle, donc est la dérivée d'une certaine fonction $\psi \in C_c^\infty(\Omega)$. On obtient

$$0 = \langle \chi, \psi'\rangle = \langle \chi, \psi\rangle - \langle \chi, \theta\rangle \int_\Omega \theta(x)\,\mathrm{d}x. \tag{9.50}$$

Donc χ s'identifie à la constante $c = \langle \chi, \theta\rangle$. On a donc montré que $u(t) = c + \mu([x,t[)$ p.p., ce qui conclut. □

On revient à la démonstration du théorème 3. On a

$$\nabla \cdot u_\epsilon^{a,b}(x) = \sqrt{\alpha}\left(a e^{\sqrt{\alpha}x} - b e^{-\sqrt{\alpha}x} + \int_0^x \frac{-u(t)}{\epsilon\sqrt{\alpha}} \left(e^{\sqrt{\alpha}(x-t)} + e^{\sqrt{\alpha}(t-x)} \right) \mathrm{d}t \right). \tag{9.51}$$

Pour que les conditions aux limites sur $u_\epsilon^{a,b}$ soient respectées, on doit poser

$$a = b = \frac{1}{2\mathrm{sh}\sqrt{\alpha}} \int_0^1 \frac{u(t)}{\epsilon\sqrt{\alpha}} \mathrm{ch}(\sqrt{\alpha}(1-t))\,\mathrm{d}t. \tag{9.52}$$

On a alors

$$\nabla \cdot u_\epsilon^{a,b}(x) = \frac{1}{\epsilon}\left(\frac{\mathrm{sh}(\sqrt{\alpha}x)}{\mathrm{sh}(\sqrt{\alpha})} \int_0^1 u(t)\mathrm{ch}(\sqrt{\alpha}(1-t))\,\mathrm{d}t - \int_0^x u(t)\mathrm{ch}(\sqrt{\alpha}(x-t))\,\mathrm{d}t \right). \tag{9.53}$$

On applique le lemme précédent en x : soit $c > 0$ telle que $\forall t \leqslant x$, $u(t) \leqslant c$, $\forall t > x$, $u(t) \geqslant c$ presque-sûrement.

$$\begin{aligned}
\epsilon \mathrm{sh}(\sqrt{\alpha}) \nabla \cdot u_\epsilon^{a,b}(x) &= \int_0^x u(t) \left(\mathrm{ch}(\sqrt{\alpha}(1-t)) \mathrm{sh}(\sqrt{\alpha} x) - \mathrm{ch}(\sqrt{\alpha}(x-t)) \mathrm{sh}(\sqrt{\alpha}) \right) \mathrm{d}t \\
&\quad + \int_x^1 u(t) \mathrm{sh}(\sqrt{\alpha} x) \mathrm{ch}(\sqrt{\alpha}(1-t)) \, \mathrm{d}t \\
&\geqslant c \int_0^x \left(\mathrm{ch}(\sqrt{\alpha}(1-t)) \mathrm{sh}(\sqrt{\alpha} x) - \mathrm{ch}(\sqrt{\alpha}(x-t)) \mathrm{sh}(\sqrt{\alpha}) \right) \mathrm{d}t \\
&\quad + \int_x^1 \mathrm{sh}(\sqrt{\alpha} x) \mathrm{ch}(\sqrt{\alpha}(1-t)) \, \mathrm{d}t \\
&= 0
\end{aligned}$$

On a donc trouvé une solution dans $\mathbb{H}^2(\Omega) \cap C_\rho$ à l'équation (9.44). On peut alors appliquer le théorème 1.1 de [10] : la projection u est dans $W = \mathbb{H}^1(\Omega)$.

□

Bibliographie

[1] Acemoglu, D., Chernozhukov, V., Werning, I., & Whinston, M. D. (2021). Optimal targeted lockdowns in a multigroup SIR model. American Economic Review: Insights, 3(4), 487-502.

[2] Alder, B. J., Wainwright, T. E. (1959), Studies in Molecular Dynamics. I. General Method, J. Chem. Phys. **31** (2): 459.

[3] Allaire, G. (2007). Numerical analysis and optimization: an introduction to mathematical modelling and numerical simulation. OUP Oxford.

[4] Ames, W.F., Numerical methods for partial differential equations, Academic press, 2014.

[5] Ballard, P., The Dynamics of Discrete Mechanical Systems with Perfect Unilateral Constraints, Arch. Rational Mech. Anal. **154**, p.p. 199–274, 2000.

[6] Berthelin, F., Existence and weak stability for a pressureless model with unilateral constraint. Mathematical Models and Methods in Applied Sciences, **12** (02):249–272, 2002.

[7] Berthelin, F., Theoretical study of a multi-dimensional pressureless model with unilateral constraint, SIAM, Journal on Mathematical Analysis, 2017, vol. **49**, no. 3, pp. 2287-2320.

[8] Blackwell, Tim M., and P. Bentley. "Don't push me! Collision-avoiding swarms." Proceedings of the 2002 Congress on Evolutionary Computation. CEC'02 (Cat. No. 02TH8600). Vol. 2. IEEE, 2002.

[9] Brezis, H., Functional Analysis, Sobolev Spaces and Partial Differential Equations, ed. Springer.

[10] Brézis, H., & Stampacchia, G. (1968). Sur la régularité de la solution d'inéquations elliptiques. Bulletin de la Société Mathématique de France, 96, 153-180.

[11] Borgwardt, K. M., Gretton, A., Rasch, M. J., Kriegel, H. P., Schölkopf, B., & Smola, A. J. (2006). Integrating structured biological data by kernel maximum mean discrepancy. Bioinformatics, 22(14), e49-e57.

[12] Bouchut, F., Brenier, Y., Cortes, J., Ripoll, J.-F., A Hierarchy of Models for Two-Phase Flows, Journal of Nonlinear Science, **10**(6):639–660, December 2000.

[13] Brennen, C. E., & Brennen, C. E. (2005). Fundamentals of multiphase flow.

[14] Bourbaki, N., Espaces Vectoriels Topologiques, Masson, Paris, 1981.

[15] Bunne, C., Meng-Papaxanthos, L., Krause, A., & Cuturi, M. (2021). Jkonet: Proximal optimal transport modeling of population dynamics. arXiv preprint arXiv:2106.06345.

[16] Bourdin, F., & Maury, B. (2021). Multibody and macroscopic impact laws: a Convex Analysis Standpoint. In Trails in Kinetic Theory (pp. 97-139). Springer, Cham.

[17] Cancès, C., Gallouët, T. O., & Monsaingeon, L. (2017). Incompressible immiscible multiphase flows in porous media: a variational approach. Analysis & PDE, 10(8), 1845-1876.

[18] Cantó, B., Coll, C., & Sánchez, E. (2017). Estimation of parameters in a structured SIR model. Advances in Difference Equations, 2017(1), 1-13.

[19] Carlier, G., Duval, V., Peyré, G., & Schmitzer, B. (2017). Convergence of entropic schemes for optimal transport and gradient flows. SIAM Journal on Mathematical Analysis, 49(2), 1385-1418.

[20] Carrillo, J. A., Gualdani, M. P., & Toscani, G. (2004). Finite speed of propagation in porous media by mass transportation methods. Comptes Rendus Mathematique, 338(10), 815-818.

[21] Céa, J. (1964). Approximation variationnelle des problèmes aux limites. In Annales de l'institut Fourier (Vol. 14, No. 2, pp. 345-444).

[22] Chilton, T. H., & Colburn, A. P. (1931). II—pressure drop in packed tubes1. Industrial & Engineering Chemistry, 23(8), 913-919.

[23] Chisholm, R.L. & Firtel, R.A. Insights into morphogenesis from a simple developmental system. Nat. Rev. Mol. Cell Biol. 5, 531–541 (2004).

[24] Crowe, C. T. (2005). Multiphase flow handbook. CRC press.

[25] Degond, P., Hua, J., Navoret, L., Numerical simulations of the Euler system with congestion constraint, Journal of Computational Physics Volume **230**, Issue 22, 10 September 2011, pp. 8057–8088.

[26] Degond, P., Minakowski, P., Navoret, L., Zatorska, E., Finite volume approximations of the Euler system with variable congestion, Computers and Fluids (2017).

[27] Doyle, P. G., & Snell, J. L. (1984). Random walks and electric networks (Vol. 22). American Mathematical Soc..

[28] Ekeland, I., Temam, R., Analyse convexe et problèmes variationnels, ed. Dunod.

[29] Falk, R. S. (1974). Error estimates for the approximation of a class of variational inequalities. Mathematics of Computation, 28(128), 963-971.

[30] Feydy, J., Séjourné, T., Vialard, F. X., Amari, S. I., Trouvé, A., & Peyré, G. (2019, April). Interpolating between optimal transport and mmd using sinkhorn divergences. In The 22nd International Conference on Artificial Intelligence and Statistics (pp. 2681-2690). PMLR.

[31] Frémond,M., Non-smooth thermomechanics. Springer-Verlag,Berlin, 2002.

[32] M. Génois, C. Vestergaard, J. Fournet, A. Panisson, I. Bonmarin, & A. Barrat, Data on face-to-face contacts in an office building suggest a low-cost vaccination strategy based on community linkers, Network Science, 3(3), 326-347 (2015).

[33] Golse, F., Saint-Raymond, L., The Navier–Stokes limit of the Boltzmann equation for bounded collision kernels, Invent. math. (2004) 155, pp. 81–161.

[34] Helbing, D. and Molnar, P. "Social force model for pedestrian dynamics." Physical review E 51.5 (1995): 4282.

[35] Joly, P., Some Trace Theorems in Anisotripic Sobolev Spaces, Siam J. Math Anal, Vol. **23**, No. 3, pp 799-819, 1994.

[36] Kermack, W. O., & McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character, 115(772), 700-721.

[37] Kim, I., & Mészáros, A. R. (2018). On nonlinear cross-diffusion systems: an optimal transport approach. Calculus of Variations and Partial Differential Equations, 57(3), 1-40.

[38] Liu, C., Zhao, Z., Brogliato, B.: Frictionless multiple impacts in multibody systems: Part I. Theoretical framework. Proceedings of the Royal Society A, Mathematical, Physical and Engineering Sciences, 464(2100), 3193–3211 (2008).

[39] Loomis, W. (Ed.). (2012). The development of Dictyostelium discoideum. Elsevier.

[40] Mauras, S., Cohen-Addad, V., Duboc, G., Dupré la Tour, M., Frasca, P., Mathieu, C., ... & Viennot, L. (2021). Mitigating COVID-19 outbreaks in workplaces and schools by hybrid telecommuting. PLoS computational biology, 17(8), e1009264.

[41] Maury, B., A time-stepping scheme for inelastic collisions, Numerische Mathematik, Volume **102**, Number 4, pp. 649 - 679, 2006.

[42] Maury, B., Analyse fonctionnelle, exercices et problèmes corrigés, Ellipses, Paris, 2004.

[43] Maury, B. (2009). Numerical analysis of a finite element/volume penalty method. SIAM Journal on Numerical Analysis, 47(2), 1126-1148.

[44] Maury, B., Preux, A., Pressureless Euler equations with maximal density constraint : a time-splitting scheme, Topological Optimization and Optimal Transport: In the Applied Sciences **17**, 333 (2017).

[45] Maury, B., Roudneff-Chupin, A., & Santambrogio, F. (2010). A macroscopic crowd motion model of gradient flow type. Mathematical Models and Methods in Applied Sciences, 20(10), 1787-1821.

[46] Maury B., Roudneff-Chupin A., Santambrogio F., Venel J. Handling congestion in crowd motion modeling. Networks & Heterogeneous Media, 2011, 6 (3) : 485-519

[47] Maz'ya, V., Sobolev Spaces: with Applications to Elliptic Partial Differential Equations, Grundlehren der mathematischen Wissenschaften, Vol. **342**, Springer Science & Business Media, 2011.

[48] Mogilner, A., & Edelstein-Keshet, L. (1999). A non-local model for a swarm. Journal of mathematical biology, 38(6), 534-570.

[49] Morale, D., Capasso, V., & Oelschläger, K. (2005). An interacting particle system modelling aggregation behavior: from individuals to populations. Journal of mathematical biology, 50(1), 49-66.

[50] Moreau, J.-J., Décomposition orthogonale d'un espace hilbertien selon deux cônes mutuellement polaires, C. R. Acad. Sci. Paris, **255** (1962), 238–240.

[51] Moreau, J. J. (1977). Evolution problem associated with a moving convex set in a Hilbert space. Journal of differential equations, 26(3), 347-374.

[52] Moreau,J.J., Some numerical methods in multibody dynamics: application to granular materials, European Journal of Mechanics A/Solids, **13** (4), 93–114 (1994).

[53] Murray, J. D. (2002). Mathematical Biology: an introduction. Heidelberg: Springer.

[54] Nguyen, N.S., Brogliato, B. (2018), Comparisons of Multiple-Impact Laws For Multi-body Systems: Moreau's Law, Binary Impacts, and the LZB Approach. In: Leine R., Acary V., Brüls O. (eds) Advanced Topics in Nonsmooth Dynamics. Springer, Cham.

[55] Otto, F. (2001). The geometry of dissipative evolution equations: the porous medium equation.

[56] Perrin, C., Westdickenberg, M., One-dimensional granular system with memory effects, SIAM Journal of Mathematical Analysis, vol. **50** (6), p.5921-5946 (2018).

[57] Peyré G. Entropic Approximation of Wasserstein Gradient Flows. SIAM Journal on Imaging Sciences, 8(4), pp. 2323–2351, 2015.

[58] Radjaï, F., Dubois, F., (Editors), Discrete-element Modeling of Granular Materials, Wiley 2011.

[59] Radjai, F., Roux, S., Moreau, J.-J., Contact forces in a granular packing. Chaos: An Interdisciplinary Journal of Nonlinear Science, AIP, 1999, 9, pp.544–544.

[60] Ristow G.. Simulating granular flow with molecular dynamics. Journal de Physique I, EDP Sciences, 1992, 2 (5), pp.649-662.

[61] Roudneff-Chupin, A. (2011). Modélisation macroscopique de mouvements de foule. Phdthesis, PhD thesis, Université Paris-Sud XI.

[62] Santambrogio, F. (2015). Optimal transport for applied mathematicians. Birkäuser, NY, 55(58-63), 94.

[63] Schatzmann, M., A class of Nonlinear Differential Equations of Second Order in Time, Nonlinear Analysis, Theory, Methods & Applications, **2**, pp. 355–373, 1978.

[64] Silsbee, R. H. "Focusing in collision problems in solids." Journal of Applied Physics 28.11 (1957): 1246-1250.

[65] Temime, L., Gustin, M. P., Duval, A., Buetti, N., Crépey, P., Guillemot, D., ... & Opatowski, L. (2020). Estimating R0 OF SARS-COV-2 in healthcare settings. medRxiv.

[66] Torquato, S., Stillinger, F. H., Jammed Hard-Particle Packings: From Kepler to Bernal and Beyond, Reviews of Modern Physics, **82**, July-September 2010.

[67] Torquato, S., Truskett, T. M., Debenedetti, P. G., Is random close packing of spheres well defined?, Phys. Rev. Lett., **84** (2000), 2064–2067.

[68] Toth, D. J., Leecaster, M., Pettey, W. B., Gundlapalli, A. V., Gao, H., Rainey, J. J., ... & Samore, M. H. (2015). The role of heterogeneity in contact timing and duration in network models of influenza spread in schools. Journal of The Royal Society Interface, 12(108), 20150279.

[69] Troianiello, G. M., Elliptic Differential Equations and Obstacle Problems, University Series in Mathematics, Springer 1987.

[70] Trottenberg, U., Oosterlee, C. W., & Schuller, A. (2000). Multigrid. Elsevier.

[71] Villani, C.: Limites hydrodynamiques de l'équation de Boltzmann [d'après C. Bardos, F. Golse, D. Levermore, P.-L. Lions, N. Masmoudi, N., L. Saint-Raymond]. Séminaire Bourbaki, vol. 2000–2001, Exp. 893.

[72] Zhu, K., & Ying, L. (2014). Information source detection in the SIR model: A sample-path-based approach. IEEE/ACM Transactions on Networking, 24(1), 408-421.

université PARIS-SACLAY

ÉCOLE DOCTORALE
de mathématiques
Hadamard (EDMH)

Titre: Modélisation macroscopique de mouvements de foule à deux types, modèles SIR condensés.

Mots clés: Mouvements de foules, Transport optimal, Modèles SIR

Résumé: On étudie dans cette thèse la modélisation macroscopique des mouvements de foule dans le cas d'une population divisée en plusieurs types ayant des comportements différents, ainsi que le développement de modèles d'épidémiologie dits "SIR" permettant l'analyse de la propagation d'une maladie infectieuse dans une école. Ces deux problématiques ont été étudiées séparément : au sujet initial (les mouvements de foule) s'est superposé le problème de modélisation d'épidémie en milieu scolaire suite à une sollicitation de MODCOV19[a], une plateforme créée par le CNRS et l'INSMI pour centraliser et coordonner les projets de modélisation autour de l'épidémie de COVID-19. Cette thèse est donc composée de deux parties indépendantes.
On analyse d'une part la convergence de différents schémas numériques découlant d'approches différentes de l'équation de mouvement de foule à deux types - flot gradient, catching-up, volumes finis. On étudie également l'homogénéisation de modèles microscopiques de particules vers le problème macroscopique. On s'intéresse enfin au problème inverse d'identification des paramètres des modèles étant donné l'observation d'un mouvement de foule.
D'autre part, on développe un type de modèle SIR dit "condensé", où les quantités épidémiologiques sont définies à l'échelle de groupes d'individus. On analyse formellement la qualité du processus de condensation lorsque l'on a accès à l'ensemble des interactions dans la population, et on présente l'implémentation effective réalisée en collaboration avec MODCOV19.

[a]https://modcov19.math.cnrs.fr/

Title: Macroscopic Modeling of the Motion of a Crowd with Two Types, Condensed SIR Models

Keywords: Crowd Motion, Optimal Transport, SIR Models

Abstract: We study in this thesis the macroscopic modelling of crowd motion in the case of a population divided in several types that may have different behaviours, as well as the development of SIR models in order to analyse the spread of an infectious disease in a school. These two issues were studied separetely. As the original topic of this thesis was crowd motion, we answered to a proposition of MODCOV19[a] - a platform created by CNRS and INSMI to centralize and coordinate modeling projects on the COVID-19 outbreak - to design epidemiological models adapted to school media. This work is thus composed of two independent parts.
On the one hand we analyse the convergence of several numerical schemes that stem from different standpoints on the macroscopic crowd motion equation - gradient flow, catching-up, finite volumes. We study as well the homogenization of microscopic models of particles towards the macroscopic model. We eventually investigate the inverse problem of identifying of the parameters of a model, being observed the motion of a crowd.
On the other hand, we develop a class of "condensed" SIR models , where the epidemiological quantities are defined at the scale of groups of individuals. We formally analyse the quality of the condensation process when a full description of the interactions within the population is available. We then detail the implementation carried out in collaboration with MODCOV19.

[a]https://modcov19.math.cnrs.fr/

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France