Université Toulouse III - Paul Sabatier

Tapuscrit présenté pour l'obtention de l'

Habilitation à Diriger des Recherches

Spécialité : Mahématiques Appliquées

par

Frédéric de GOURNAY

de l'Institut de Mathématiques de Toulouse

Quelques applications de l'optimisation

Que nul n'entre ici s'il n'est optimiseur

Soutenance le 30 juin 2023, devant le jury composé de :

Rapportrice :	Mme. Julie Delon	Professeure des Universités
Rapporteur :	Mr. Simon Masnou	Professeur des Universités
Rapporteur :	Mr. Édouard Oudet	Professeur des Universités
Examinatrice :	Mme. Gersende Fort	Directrice de Recherche
Examinatrice :	Mme. Aude Rondepierre	Professeure des Universités
Parrain :	M. Pascal Noble	Professeur des Universités

Remerciements

La première chose qui est lue dans une thèse et dans une HDR est la section des remerciements. Ainsi, afin de forcer le lecteur à parcourir ce tapuscrit, j'ai donc décidé de disséminer mes remerciements dans mon Habilitation et je me contenterai donc de remercier dans cette section ceux que je n'ai pu citer plus tard. De plus, je ne vais mettre aucun nom, le lecteur devra se reconnaître lui-même[1].

Tout d'abord je dois remercier celles et ceux qui m'ont accompagné dans la préparation de cette HDR scientifiquement ou moralement : elles sont rapportrices, examinatrices, voisine de bureau, liée par les liens indissociables du mariage avec ma personne, simplement collègues ou même ma progéniture ; ils sont rapporteurs, parrains, collègues et sont aussi ceux qui viennent vous voir un jour à la pause café et qui -pour vous motiver à l'écriture de l'HDR- exigent de vous une date de soutenance dans moins de 6 mois.

Ensuite, je me dois de remercier les collaborateurs divers, qui sont d'une importance capitale. Il y a dans cette catégorie les mentors scientifiques, tous ceux qui m'ont appris quelque chose. Il y a aussi toutes les autres co-autrices et les autres co-auteurs. Mais aussi les équipes diverses, qu'elles soient scientifiques ou pédagogiques. Merci d'avoir travaillé avec moi.

Les "ou" de cette liste de remerciements n'étant pas exclusifs, je vais maintenant remercier mes collègues passés et présents des laboratoires ou départements. Pour l'immense majorité d'entre elles et d'entre eux, ce sont de gentilles gens aimables et aimés (par moi). Si quelque râleur[2] vient s'immiscer dans l'agréable ambiance d'un laboratoire ou d'un département pour la ternir, il ne fait que ressortir les qualités des autres et le plaisir que j'ai à les côtoyer. Notamment, si vous êtes directrice ou directeur de quelque-chose ou alors responsable d'autre chose et si dans nos interactions nous avons toujours travaillé dans une ambiance constructive, merci beaucoup.

Fi du travail, place à la famille. Je vis entouré de femmes jeunes[3] qui m'ont transformé en féministe acharné. Elles sont surtout le sel de mon existence et souvent ma plus grande joie. Il y a aussi dans la famille celles et ceux qui vous ont vu grandir, fait grandir et aidé à grandir. Merci à vous, je ne me sens pas tout le temps à votre hauteur, c'est là qu'il faut chercher la véritable origine des compliments dithyrambiques dont je peux quelquefois vous abreuver.

Ceux qui ont suivi le déroulement de ces remerciements s'attendent peut être à ce que je remercie mes compatriotes. Je crois que je vais plutôt remercier toute cette partie de l'humanité qui ne passe pas son temps à envoyer l'autre partie se tuer à la tâche, à la guerre ou sous les roues de son véhicule. Merci ainsi à ceux qui ont une conscience et notamment écologique. Nous sommes les plus nombreux, c'est juste que les autres braillent plus fort.

C'est fini, vous pouvez refermer ce tapuscrit :).

1. tandis que la lectrice devra se reconnaître elle-même, c'est comme cela...
2. j'ai des noms !!!
3. plus ou moins

Présentation générale du manuscrit

Présentation de cette HDR

J'ai décidé d'organiser le manuscrit de mon HDR de manière plus thématique que chronologique. De plus j'ai essayé de synthétiser les résultats principaux des différents articles de recherche tout en les mettant en perspective. Quand j'énonce les résultats dans ce tapuscrit, j'essaie de ne pas alourdir le message en omettant les hypothèses techniques, de même, j'ai essayé, quand je l'ai pu, de mettre les idées des preuves en retirant au maximum les détails pour ne garder que le goût général de la démonstration.

Il y a six grands chapitres thématiques dans cette HDR.

- Le premier chapitre concerne tout mon travail sur les problématiques d'**optimisation de forme** qui furent mon sujet de thèse avec Grégoire Allaire et François Jouve et que j'ai continué à explorer tout au long de ma carrière.
- Le deuxième chapitre traite des **problèmes inverses**, il recoupe surtout des travaux faits entre 2006 et 2012, en collaboration avec Otared Kavian et Yves Capdeboscq.
- Le troisième chapitre est consacré au **problème de Graëtz**, qui est un problème de convection-diffusion. Je commence à m'intéresser à ce problème avec Franck Plouraboué et Jérôme Fehrenbach vers 2010.
- Dans le quatrième chapitre, on s'intéresse aux problématiques sur l'**optimisation avec parcimonie** (avec des termes L^1), c'est avec Pierre Weiss puis Axel Flinth que je m'intéresse à ces problèmes vers 2018.
- Dans le cinquième chapitre, on trouvera des contributions sur le problème du **transport optimal**, ce sont des problématiques que j'ai regardé avec Léo Lebrat et Jonas Kahn et Pierre Weiss dès 2016.
- Finalement, le sixième chapitre est un recueil de travaux qui traitent de **réseau de neurones**, ce sont des problématiques que j'ai surtout regardé avec Pierre Weiss et Alban Gossard à partir de 2019.

Quelques remarques personnelles

Comme le reste de ce manuscrit est dédié à mon travail de recherche, je préfère dans ce paragraphe présenter ma personne et quelques remarques personnelles que j'ai sur mon propre travail.

Il est difficile et vain de trouver une unité thématique à la recherche de quelqu'un si cette unité est prise au sens du singulier et vise à réduire la personne à une simple description, qui, si elle n'a pas l'avantage de la compréhension de la complexité de l'individu, a celle de la simplicité de réduire quelqu'un à une identité. Une approche de recherche doit se faire dans la joie du butinage, dans l'excitation de la curiosité, même si cela implique d'être volage.

Ainsi, s'il n'y a pas d'unité dans ma recherche, il y a quand même de la cohérence. Cette cohérence est à chercher dans mes goûts scientifiques car l'on n'aime ce que l'on fait que si on ne fait que ce que l'on aime. Ainsi j'aime l'optimisation et les EDP. J'aime créer un code afin de bien comprendre le fonctionnement des théories mathématiques. Pour finir j'aime bien les à côtés de la recherche, qui font le reste du métier d'enseignant-chercheur. L'enseignement tout d'abord, m'a permis quelquefois, mais bien trop rarement à mon goût, de trouver des étudiants passionnés et surtout passionnants. L'administratif ensuite [4] me permet de m'investir dans la vie collective. Surtout ces deux activités sont balisées, récurrentes, rassurantes mêmes et servent de ligne de basse d'une partition professionnelle dont la mélodie [5] serait la recherche.

Présentation de l'impétrant

Il est d'usage semble-t-il d'ajouter à une HDR un petit *curriculum vitae*. Je vais me plier à cette règle tout en l'embellissant un peu sans tomber non plus dans les excès de l'autobiographie.

23 juillet 1980 : Naissance à Paris XX^e. Je n'ai pas grand chose à dire de ce jour qui marqua l'histoire car je n'ai pas beaucoup de souvenirs de ce jour là. Si je dois commencer mes remerciements, je pense qu'il faut que j'ai une pensée pour mes parents, Chantal et Jean-Marc, qui sont heureux de m'avoir [6].

23 Juillet 1996 : J'obtiens mon baccalauréat avec mention Bien. Cela fut une surprise car ma scolarité fut un peu perturbée, j'ai du sauter des classes pour éviter l'ennui. De surcroît, mes notes furent durement impactées par quelques fonctionnaires de l'Education Nationale avec lesquels je ne partageais rien d'autre que du mépris. Heureusement, il existe ces enseignants lumineux, je nomme ici Marie-Paul Cruveiller [7] et Mr Janvier, qui fut mon professeur de physique de 3ème et qui est à l'origine de mon goût pour la science.

23 Juillet 2001 : Après une classe préparatoire à Stanislas qui me ravit [8] et après un passage à l'école Polytechnique où j'ai compris assez vite que ma diplomation serait indépendante de mon comportement [9], je suis ingénieur. Ne comprenant pas l'appétit de mes camarades pour les carrières du secteur privé, je continue mon apprentissage de scientifique. Comme je n'ai rien fait dans mon école d'ingénieur, je n'ai pas de bourses, et le même niveau scientifique qu'en y entrant. Je me lance donc dans la lecture assidue des livres de Rudin pour préparer mon D.E.A à Paris V.I. (Analyse numérique).

23 Juillet 2002 : Je vais commencer ma thèse avec Grégoire Allaire que j'ai rencontré en stage de D.E.A. J'ai été impressionné dès le début par Grégoire et par ce sujet qui m'intriguait et qui me semblait presque magique. En fait, j'avais déjà rencontré Grégoire avant, dans la mesure où il est professeur à Polytechnique ; non pas en cours, je n'étais pas assez assidu pour cela, mais en rattrapage. Le D.E.A s'est bien passé, je m'en suis sorti avec une mention *Cum Laudae*. Ce n'était pas si mal pour quelqu'un qui ignorait tout des espaces vectoriels de dimension infinie en entrant [10]

23 Juillet 2005 : Ma thèse au CMAP est terminée. Elle s'appelle *Optimisation de formes par la méthode des courbes de niveaux*. Je n'ai jamais su si c'était par désespoir sur mon cas, mais Grégoire s'est fait épauler dans la direction de thèse par François Jouve. Je les en remercie tous les deux, ils m'ont façonné même si je n'aurai plus qu'une crainte toute ma vie, ne pas être à leur hauteur.

23 juillet 2006 : Je rentre en France après mon post-doc à Rutgers aux Etats-Unis sous la direction de Michael Vogelius. Je me suis remis finalement à l'homogénéisation et au gradient topologique et je suis recruté à Versailles. Je sais que je vais y rencontrer Yves Capdeboscq,

4. Et je pense bien que je suis le seul MCF à revendiquer du goût pour cela dans une HDR
5. Qui semble quelquefois être écrite le John Cage de 4'33"
6. Ce jour là.
7. Mon enseignante de CP ! !
8. Essentiellement car on recommence les mathématiques depuis le début et qu'on n'y fait pas de géométrie euclidienne du triangle.
9. Une variable aléatoire constante est indépendante de toutes les autres v.a., y compris de elle-même
10. Pour l'anecdote, je n'ai vraiment raté qu'une seule seule matière en D.E.A, l'homogénéisation, enseignée par François Jouve et Eric Bonnetier, d'après le livre d'un certain G. Allaire.

je connais et apprécie ses travaux mais je suis loin de connaître son humour [11]. Je vais aussi rencontrer deux autres mathématiciens qui m'ont terriblement marqué : Otared Kavian et Thierry Horsin.

23 juillet 2009 : Je collabore avec Jérôme Fehrenbach. Je ne le sais pas encore mais cela entraînera une cascade d'évènements qui changera ma vie pour toujours. Pour l'instant, c'est l'esprit joyeux et inconscient des évènements à venir que je prépare un voyage à Toulouse pour travailler avec lui.

23 juillet 2010 : Je reviens d'un séjour de recherche à l'université de Washington aux Etats-Unis. J'ai rencontré la jeune Stéphanie Bescos lors d'un séjour à Toulouse au mois d'octobre 2009. Ma décision est prise, je vais remettre mon HDR à un peu plus tard et je vais essayer de la rejoindre.

23 juillet 2011 : Ca y est, j'ai un poste à Toulouse depuis septembre 2010. J'ai rencontré plein de nouveaux collègues. Il y a ceux que je connaissais déjà (Sylvain Ervedoza, Mohammed Masmoudi, Jérôme Fehrenbach, Jean-Pierre Raymond) et ceux que j'apprécie déjà, parmi les INSAiens Aude Rondepierre, David Sanchez, Florent Chazel, Violaine Roussier-Michon, Sandrine Scott, Adeline Rouchon, Olivier Mazet, Alain Huard, Simona Grusea et évidemment celle que je n'ai pas besoin de nommer pour la remercier Cathy Maugis-Rabusseau. Il y a aussi les nouveaux collaborateurs qui me marqueront profondément que sont Franck Plouraboué et Pierre Weiss . Je ne sais pas si je les apprécie plus pour leur intelligence que pour leur bonne humeur.

23 juillet 2014 : Ma première fille Anna est née il y a plus d'un an et ma deuxième fille Alice viendra au monde dans deux jours. Je me sens un peu accaparé par la gestion des enfants. Une période se tourne avec la naissance des enfants et je commence à me rendre compte que je recentre mon activité sur ma famille.

23 juillet 2019 : C'est la fin de la thèse de mon premier étudiant, Léo Lebrat, je suis un peu inquiet car il va partir en post-doc en Australie *avant* de soutenir à Toulouse. Cela ne semble pas inquiéter mon co-directeur de thèse, Jonas Kahn [12].

23 juillet 2022 : J'ai 42 ans et depuis ces 8 dernières années beaucoup de collègues sont arrivés, il me faut ici penser à Robin Bouclier, Romain Duboscq, Charles Dossal, Pascal Noble. Sinon, je me suis marié avec Stéphanie en octobre 2021 et la COVID-19 a bouleversé toutes nos certitudes sur l'enseignement. Bientôt Alban Gossard va soutenir sa thèse, mais il a du travail pour ses derniers résultats numériques. J'en connais un qui va pas mal bosser cet été.

23 juillet 2023 : J'ai une petite idée de ce qui m'a marqué dans le début de l'année 2023, mais c'est une histoire qui reste à écrire...

11. Par gentillesse pour Yves qui est très recommandable, je ne m'étendrais pas sur ce sujet
12. Jonas je t'aime beaucoup, mais quelquefois,....

Table des matières

Chapitre 1

Optimisation de formes par la méthode des courbes de niveaux

1.1 Position du problème

J'ai commencé à m'intéresser pendant ma thèse à des problèmes d'optimisation d'équations aux dérivées partielles. Les solutions de ces équations dépendent des termes sources, des différents paramètres de l'équation mais aussi du domaine sur lequel elles sont posées. Les problèmes d'optimisation de formes prennent comme variable d'optimisation ce domaine. Nous écrirons de manière générique les problèmes d'optimisation de formes de la manière suivante :

Définition 1.1 Soit une équation aux dérivées partielles posée sur un domaine Ω dont u est la solution. L'objectif est de minimiser une fonctionnelle $\mathcal{J}$ dépendante de u et de Ω en utilisant comme variable d'optimisation le domaine Ω :

$$\arg\min_{\Omega\in\mathcal{U}_{ad}} \mathcal{J}(\Omega, u(\Omega)), \tag{1.1}$$

où $\mathcal{U}_{ad}$ est l'ensemble des domaines admissibles.

L'archétype des problèmes d'optimisation de forme consiste à prendre l'équation de l'élasticité linéaire (petite déformation/ grands déplacements) pour définir u qui est un champ de vecteur de $\mathbb{R}^d$ et qui résout l'équation suivante :

$$-\, div Ae(u) = f \text{ dans } \Omega \quad \text{avec } e(u) = \frac{1}{2}(\nabla u + \nabla u^T), \tag{1.2}$$

où A est un tenseur d'ordre 2, typiquement le tenseur d'élasticité de Hooke qui est défini pour toute matrice carrée ξ de taille d comme

$$A\xi = 2\mu\xi + \lambda(tr\xi)\mathrm{Id}, \quad \mu, \lambda \in \mathbb{R}, \mu > 0, 2\mu + d\lambda > 0.$$

En langage de la mécanique des solides, $\sigma = Ae(u)$ sont les contraintes, $e(u)$ sont les déformations et u le déplacement. Le bord de Ω est partitionné en sa partie de Dirichlet et de Neumann où sont appliquées les conditions aux bord éponymes

$$u = 0 \text{ sur } \Gamma_D \quad \sigma\cdot n = g \text{ sur } \Gamma_N.$$

Quitte à prendre un relèvement, nous pouvons toujours nous ramener à des conditions de Dirichlet homogènes sur le bord. Le problème d'optimisation de forme le plus simple consiste à minimiser la compliance, définie comme étant l'énergie du système par

$$\mathcal{J}(u) = \frac{1}{2}\int_\Omega Ae(u) : e(u) = \int_\Omega f \cdot u + \int_{\Gamma_N} g \cdot u.$$

Pour finir de donner un sens à (1.1), il suffit de préciser $\mathcal{U}_{ad}$. Nous allons nous limiter au cas simple où l'ensemble des domaines admissibles est l'ensemble des domaines Ω qui contiennent un ensemble ω donné et dont le volume total $|\Omega|$ est borné

$$\mathcal{U}_{ad} = \{\Omega \text{ tels que } \omega \subset \Omega \text{ et } |\Omega| \leq C\}.$$

On pourra par exemple considérer le cas où le support de f est contenu dans ω pour éviter de choisir comme Ω optimal l'ensemble vide.

1.1.1 Structure du gradient de forme

On s'intéresse à différentier la solution de l'équation aux dérivées partielles et ensuite la fonction coût par rapport à la forme du domaine. La structure nécessaire pour différentier est celle d'une variété, c'est-à-dire qu'à chaque domaine Ω on associe un espace vectoriel tangent et une application définie sur un sous-ensemble de cet espace tangent dans $\mathcal{U}_{ad}$.

Dans la version classique du gradient de forme [MS75 ; SZ92 ; Pir12], on suppose que la topologie du domaine ne varie pas. On se choisit un domaine de référence Ω_0 et on considère $\mathcal{U}_{ad}$ l'ensemble des domaines images de Ω_0 par un difféomorphisme $W^{1,\infty}$:

$$\mathcal{U}_{ad} = \{T(\Omega_0) \text{ tel que } T, T^{-1} \in W^{1,\infty}(\mathbb{R}^d, \mathbb{R}^d)\}. \tag{1.3}$$

On peut adjoindre à $\mathcal{U}_{ad}$ des contraintes supplémentaire de volume ou de périmètre sur les ouverts. Deux avantages s'offrent à cette approche : comme les difféomorphismes considérés forment un groupe pour la loi de composition, $\mathcal{U}_{ad}$ ne dépend pas du choix de Ω_0 dans $\mathcal{U}_{ad}$. Autrement dit $\mathcal{U}_{ad}$ est une classe d'équivalence des domaines définis par la relation d'équivalence "Il existe un difféomorphisme $W^{1,\infty}$ entre les deux domaines". De plus l'ensemble des difféomorphismes $W^{1,\infty}$ est une variété dont l'espace tangent est l'ensemble des champs de vecteur $W^{1,\infty}$. Ainsi, en écrivant

$$T(x) = x + \theta(x) \text{ pour } T \text{ proche de } Id,$$

l'espace tangent est défini comme l'ensemble des champs de vecteur $\theta \in W^{1,\infty}(\mathbb{R}^d, \mathbb{R}^d)$. Le calcul de la différentielle est l'objet de la Section 1.1.2, nous énonçons ici le théorème de structure d'Hadamard qui stipule qu'elle doit s'écrire (quand elle existe) comme une intégrale sur le bord

Théorème 1.2 — Théorème de structure de Hadamard. Soit une application $\mathcal{J}$ à valeurs réelles, on note $d\mathcal{J}.\dot{\theta}$ la différentielle au point $\theta = 0$ dans la direction $\dot{\theta}$ de l'application

$$\theta \mapsto \mathcal{J}(\Omega_\theta, u(\Omega_\theta)) \quad \Omega_\theta = \{x + \theta(x), x \in \Omega_0\}$$

Alors si cette différentielle existe, il existe un champ de vecteur p appelé "pression du gradient de forme" tel que

$$d\mathcal{J}.\dot{\theta} = \int_\Omega \operatorname{div}(\dot{\theta} p) \tag{1.4}$$

$$= \int_{\partial\Omega} (\dot{\theta} \cdot n) p \tag{1.5}$$

En un point x du bord de Ω, le signe de p indique si le domaine doit augmenter (p négative) ou diminuer (p positive) au point x pour minimiser $\mathcal{J}$.

Démonstration

La démonstration du théorème de structure repose sur le fait qu'au premier ordre Ω_θ ne dépend que de la composante normale de θ sur le bord. Effectivement, au premier ordre, la partie tangentielle de θ sur le bord ne fait que transporter les points de Ω_0 sur la frontière. A l'intérieur de Ω_0, le champ θ ne fait que transporter des points de l'intérieur vers l'intérieur. En utilisant un théorème de représentation, la différentielle étant une forme linéaire ne dépendant que de $\theta \cdot n$ sur le bord doit pouvoir s'ecrire sous la forme (1.5).

La formulation la plus générale du théorème 1.2 est certainement (1.4) qui n'a pas besoin d'une formule de Gauss et qui peut se voir comme un crochet de dualité, l'avantage de l'écriture de (1.5) est que la dépendance de la différentielle par rapport à $\dot{\theta}$ est explicite. Ainsi il apparaît clairement que seule la valeur de la pression sur le bord du domaine est importante et qu'elle peut être relevée dans le domaine et à l'intérieur comme souhaité.

1.1.2 Calcul du gradient de forme

L'obtention du gradient de forme et de la "pression" p intervenant dans l'équation (1.5) est une procédure relativement fastidieuse [MS75 ; Sim80 ; SZ92 ; AS07 ; Pir12] dont je vais résumer les différentes étapes clefs, nous l'appliquerons à l'équation de Laplace homogène avec une fonction à dériver relativement simple.

Proposition 1.3 Pour tout θ, considérons u_θ solution de :

$$-\mathrm{div}\nabla u_\theta = f \text{ dans } \Omega_\theta \quad u_\theta = 0 \text{ sur } \partial\Omega_\theta$$

Soit j une fonction à valeurs réelles et une fonction coût de la forme

$$\theta \mapsto \mathcal{J}(\theta) = \int_{\Omega_\theta} j(x, u_\theta(x))dx. \tag{1.6}$$

Si f, Ω_0 et j sont assez réguliers et si $\hat{u}$ est l'adjoint, défini dans (1.11), alors le Théorème 1.2 s'applique et

$$d\mathcal{J}.\dot{\theta} = \int_{\partial\Omega_0} (\dot{\theta} \cdot n)p \quad \text{avec} \quad p(x) = j(x, u_0(x)) + f(x)\hat{u}(x) - \nabla u_0(x) \cdot \nabla \hat{u}(x).$$

La démonstration de cette proposition se fait en plusieurs étapes qui composent le reste de cette section.

Forme variationnelle On met tout d'abord l'équation sous forme variationnelle, ainsi on cherche $u_\theta \in H_0^1(\Omega_\theta)$ qui vérifie

$$\int_{\Omega_\theta} \nabla u_\theta \nabla v = \int_{\Omega_\theta} fv \quad \forall v \in H_0^1(\Omega_\theta)$$

Avec un changement de variable donné par $T = Id + \theta$ dans l'intégrale, on se ramène à une forme variationnelle telle que la dépendance en θ ne s'exprime plus sur les intégrales mais sur les intégrandes. Après quelques calculs, on se ramène à chercher $\tilde{u}_\theta = u_\theta \circ T \in H_0^1(\Omega_0)$ tel que :

$$\int_{\Omega_0} A(\theta)\nabla\tilde{u}_\theta\nabla\tilde{v} = \int_{\Omega_0} F(\theta)\tilde{v} \quad \forall \tilde{v} \in H_0^1(\Omega_0)$$

Il est primordial à ce stade de remarquer que si v décrit $H_0^1(\Omega_\theta)$ alors $\tilde{v} = v \circ T$ décrit $H_0^1(\Omega_0)$ et que les espaces de Sobolev dans lesquels on cherche les variables ne dépendent plus de θ. Dans cet exemple, nous avons

$$A(\theta) = |det\nabla T|(\nabla T)^{-1}(\nabla T)^{-T} \text{ et } F(\theta) = |det\nabla T| f \circ T \tag{1.7}$$

Ensuite, en utilisant le même changement de variable, on obtient à partir de (1.6) la formule suivante pour la fonctionnelle $\mathcal{J}$:

$$\mathcal{J}(\theta) = \int_{\Omega_0} |det(\nabla T)| j(T(x), \tilde{u}_\theta(x)) dx \tag{1.8}$$

Différentiation Ensuite nous pouvons -via un théorème des fonctions implicites- différentier $\tilde{u}_\theta$ par rapport à θ. En fixant une direction $\dot{\theta}$ et en notant $\dot{u}$, $\dot{A}$, $\dot{F}$ les dérivées dans cette direction, nous avons :

$$\int_{\Omega_0} \nabla \dot{u} \nabla \tilde{v} = \int_{\Omega_0} \dot{F}\tilde{v} - \dot{A}\nabla u \nabla \tilde{v} \quad \forall \tilde{v} \in H_0^1(\Omega). \tag{1.9}$$

Avec ici, en dérivant (1.7), $\dot{F} = \mathrm{div}(\dot{\theta})f + \nabla f \cdot \dot{\theta}$ et $\dot{A} = div(\dot{\theta})Id - \nabla\dot{\theta} - (\nabla\dot{\theta})^T$. De même nous obtenons en dérivant (1.8) la formule suivante pour $\dot{\mathcal{J}} = d\mathcal{J}.\dot{\theta}$:

$$\dot{\mathcal{J}} = \int_{\Omega_0} div(\dot{\theta})j + \nabla_x j \cdot \dot{\theta} + \partial_u j.\dot{u} = \int_{\Omega_0} div\left(\dot{\theta}j\right) - \partial_u j.(\nabla u \cdot \dot{\theta}) + \partial_u j.\dot{u} \tag{1.10}$$

Introduction de l'adjoint Dans la formule (1.10), la dépendance de $\dot{\mathcal{J}}$ par rapport à $\dot{\theta}$ est linéaire, ce qui confirme que l'on a bien calculé la différentielle, mais le calcul effectif du gradient est hors de portée dans la mesure où celui-ci fait intervenir le calcul de $\dot{u}$ qui dépend implicitement de $\dot{\theta}$. Pour remédier à ce problème, on introduit le concept d'adjoint $\hat{u}$ qui résout l'equation

$$\int_{\Omega_0} \nabla v \nabla \hat{u} = \int_{\Omega_0} \partial_u j.v \quad \forall v \in H_0^1(\Omega_0). \tag{1.11}$$

L'adjoint est construit pour que l'on puisse prendre $v = \dot{u}$ dans (1.11) et $\tilde{v} = \hat{u}$ dans (1.9) et de tout substituer dans (1.10) pour trouver

$$\dot{\mathcal{J}} = \int_{\Omega_0} \mathrm{div}(j\dot{\theta}) - \partial_u j.(\nabla u \cdot \dot{\theta}) + \dot{F}\hat{u} - \dot{A}\nabla u \nabla \hat{u}$$

Mise sous forme de Hadamard La vérification du théorème de structure d'Hadamard (1.4) n'est pas immédiate, elle se fait en remarquant que $\dot{F} = \mathrm{div}(f\dot{\theta})$ et en utilisant la formule suivante, vraie pour tout $u, \hat{u}, \dot{\theta}$

$$\dot{A}\nabla u \nabla \hat{u} - \Delta u(\nabla \hat{u} \cdot \dot{\theta}) - \Delta \hat{u}(\nabla u \cdot \dot{\theta}) = \mathrm{div}(M\dot{\theta}) \text{ avec } M = (\nabla \hat{u} \cdot \nabla u)Id - \nabla \hat{u} \nabla u^T - \nabla u \nabla \hat{u}^T \tag{1.12}$$

Il suffit ensuite de remplacer $-\Delta u$ par f et $-\Delta \hat{u}$ par $\partial_u j$ pour obtenir :

$$\dot{\mathcal{J}} = \int_{\Omega_0} div(j\dot{\theta}) + div(f\dot{\theta})\hat{u} + f(\nabla \hat{u} \cdot \dot{\theta}) + div(M\dot{\theta}) = \int_{\partial\Omega_0} (j + f\hat{u})(\dot{\theta} \cdot n) + (M\dot{\theta}) \cdot n$$

On conclut ensuite en remarquant que les conditions de Dirichlet imposent

$$(M\dot{\theta}) \cdot n = -(\nabla u \cdot \nabla \hat{u})(\dot{\theta} \cdot n)$$

sur le bord pour compléter le calcul et on trouve ici une pression qui vaut

$$p = j + fw - \nabla u \cdot \nabla \hat{u}.$$

Discussion Toutes les étapes se justifient correctement, à l'exception notoire de la dernière. Effectivement, si la formule (1.12) est bien vraie au sens des distributions, elle est appliquée au sens L^1 et il n'y a aucune garantie que le terme $(\nabla u \cdot \nabla \hat{u})(\theta \cdot n)$ soit bien un terme L^1 de $\partial\Omega_0$ pour pouvoir appliquer la formule de calcul de la pression p.

1.1.3 Méthode de Level-set

La méthode des courbes de niveaux (level-set) permet de résoudre efficacement le problème d'optimisation de forme, elle peut être vue comme une re-paramétrisation des variables en supposant que $\mathcal{U}_{ad}$, l'ensemble des domaines admissibles s'écrit comme l'ensemble des courbes de niveaux de certaines fonctions. En se limitant à $\mathcal{F}$, un sous-ensemble des fonctions continues :

$$\mathcal{U}_{ad} = \{\Omega \text{ tel que } \exists \phi \in \mathcal{F} \text{ et } \Omega = \phi^{-1}(\mathbb{R}^-)\}$$

Il y a un avantage numérique certain à modifier la description de $\mathcal{U}_{ad}$. On prendra par exemple l'ensemble $\mathcal{F}$ comme l'ensemble des fonctions $P1$ sur un maillage triangulaire ou tétrahédral, ce qui nous permet d'avoir un ensemble de variables d'optimisation simples à manipuler et pour lesquelles, l'application qui à une fonction ϕ associe le domaine correspondant est aisément calculable. De plus, advecter un domaine Ω par rapport à un champ de vecteur θ revient, à advecter une courbe de niveau de la fonction ϕ correspondante, c'est-à-dire à suivre un pas de temps de l'équation

$$\frac{\partial \phi}{\partial t} - V|\nabla \phi| = 0 \quad \text{avec } V = \theta \cdot n \text{ sur } \partial\Omega \tag{1.13}$$

Grâce au théorème de structure de Hadamard, comme la dérivée de la fonction coût s'écrit sous la forme (1.5), il suffit de prendre comme vitesse $V = -p$ sur le bord de Ω et une extension quelconque à l'intérieur. L'apparente simplicité de la méthode de la courbe des niveaux cache une difficulté mathématique. Effectivement l'ensemble des domaines admissibles ainsi défini n'est plus l'ensemble introduit par (1.3) comme l'ensemble des domaines atteignables par un difféomorphisme et la différentielle calculée par rapport aux difféomorphismes peut n'avoir aucun rapport avec le gradient par rapport aux paramètres de ϕ (à supposer que ce dernier existe). C'est le cas notamment quand ϕ est modifié de telle sorte que sa courbe de niveau change de topologie.

1.1.4 Le gradient topologique

Le gradient topologique est un autre type de modification de domaine. Au lieu d'advecter la frontière par un champ de vecteur, comme dans la Section 1.1.2, on s'intéresse à la modification de la solution d'une edp lors de la création d'un trou à l'intérieur même du domaine. Les preuves de dérivation par rapport à la nucléation [EKS94 ; SZ99 ; Céa+00 ; GGM01 ; SZ01] sont assez fastidieuse, par contre, si on impose que le trou soit un "faux trou", la preuve du gradient topologique devient plus simple, nous nous baserons sur l'analyse de [CV03]. Si par exemple en électrostatique, on considère l'équation suivante :

$$-\operatorname{div}(\sigma_\varepsilon \nabla u_\varepsilon) = f \text{ dans } \Omega \text{ et } u_\varepsilon = 0 \text{ sur } \partial\Omega, \tag{1.14}$$

où on définit la conductivité par

$$\sigma_\varepsilon = \begin{cases} \sigma_0 \text{ sur } \Omega \setminus \varepsilon B \\ \sigma^\star \text{ sur } \varepsilon B \end{cases} \quad \text{et } \varepsilon B = \{\varepsilon x, x \in B\} \quad \sigma^\star \in \mathbb{R}^{+*}, \tag{1.15}$$

où $B \subset \mathbb{R}^d$ est un ouvert contenant 0. Le "trou" est εB et est rempli d'un matériau di-électrique. Un "vrai trou" consiste à prendre $\sigma^\star = +\infty$ (trou de Dirichlet, conducteur parfait) ou $\sigma^\star = 0$ (trou de Neumann, isolant parfait). Dans le cas d'un faux trou, il est relativement facile de calculer la différentielle de u_ε dans différentes normes [CV03] avec des techniques issues de la théorie de l'homogénisation [All12].

Le type de résultat du gradient topologique est le suivant :

Proposition 1.4 Si u_ε résout (1.14) et si u_0 est suffisament régulier autour de 0, il existe une matrice symétrique $\mathcal{M}$, indépendante de u_ε, telle que pour toute fonction test ϕ régulière autour de 0, on a :

$$\int_\Omega \sigma_0 \nabla(u_\varepsilon - u_0)\nabla\phi = \varepsilon^d \mathcal{M}\nabla u_0(0) \colon \nabla\phi(0) + o(\varepsilon^d). \tag{1.16}$$

Avec des techniques d'adjoints, c'est-à-dire en introduisant des équations similaires à (1.11) on remontera à toute dérivée de fonction coût via l'équation (1.16) qui est similaire à (1.9), pourvu que l'adjoint soit bien C^0 autour de 0.

Introduction de $\mathcal{M}$ En utilisant des inégalités d'énergie, on établit tout d'abord des bornes uniformes dans $H^1(\Omega)$ de u_ε et on s'intéresse à $d_\varepsilon = u_\varepsilon - u_0 \in H_0^1(\Omega)$ qui résout l'équation

$$\int_\Omega \sigma_\varepsilon \nabla d_\varepsilon \nabla\phi = (\sigma^\star - \sigma_0)\int_\Omega \mathbb{1}_{\varepsilon B}\nabla u_0 \nabla\phi \quad \forall\phi \in H_0^1(\Omega) \tag{1.17}$$

$$\int_\Omega \sigma_0 \nabla d_\varepsilon \nabla\phi = (\sigma_0 - \sigma^\star)\int_\Omega \mathbb{1}_{\varepsilon B}\nabla u_\varepsilon \nabla\phi \quad \forall\phi \in H_0^1(\Omega) \tag{1.18}$$

De l'équation (1.18), on déduit que (1.16) est équivalent à montrer, au sens faible-* des mesures, un développement semblable à

$$\varepsilon^{-d}\mathbb{1}_{\varepsilon B}\nabla u_\varepsilon = \mathcal{M}\nabla u_0(0)\delta_{x=0} + o(1). \tag{1.19}$$

On commence à montrer la convergence au sens des mesures du terme de gauche, pour se faire il suffit de montrer que $\|\varepsilon^{-d}\mathbb{1}_{\varepsilon B}\nabla u_\varepsilon\|_{L^1} = \mathcal{O}(1)$. Par définition de d_ε et régularité de u_0, il suffit de montrer que $\|\varepsilon^{-d}\mathbb{1}_{\varepsilon B}\nabla d_\varepsilon\|_{L^1} = \mathcal{O}(1)$. Grâce à un Cauchy Schwartz, on borne $\|\varepsilon^{-d}\mathbb{1}_{\varepsilon B}\nabla d_\varepsilon\|_{L^1}$ par le produit de $\|\varepsilon^{-d/2}\mathbb{1}_{\varepsilon B}\|_{L^2}$ et de $\|\varepsilon^{-d/2}\nabla d_\varepsilon\|_{L^2}$. Il est clair que $\|\varepsilon^{-d/2}\mathbb{1}_{\varepsilon B}\|_{L^2} = \mathcal{O}(1)$, il suffit donc de montrer que $\|\varepsilon^{-d/2}\nabla d_\varepsilon\|_{L^2} = \mathcal{O}(1)$. Cette borne est dérivée de (1.17) , via une borne en énergie, c'est-à-dire en prenant $v = d_\varepsilon$ dans (1.17).

Borne L^2 de la différence On sait d'après ce qui précède que le terme de gauche de (1.19) converge au sens des mesures. Il suffit maintenant de montrer que le terme de droite a la forme voulue. Il faut d'abord tirer de la borne $\|d_\varepsilon\|_{H^1} = \mathcal{O}(\varepsilon^{d/2})$ une borne en $\|d_\varepsilon\|_{L^2} = o(\varepsilon^{d/2})$. Pour ce faire on introduit p, l'adjoint de la norme L^2 de $\|d_\varepsilon\|_{L^2}^2$, solution de

$$\int_\Omega \sigma_0 \nabla p \nabla\tilde{v} = \int_\Omega d_\varepsilon \tilde{v} \quad \forall\tilde{v} \in H_0^1(\Omega). \tag{1.20}$$

Un peu de régularité elliptique et des injections de Sobolev nous donne pour un certain $q > 2$ $\|\nabla p\|_{L^q} = \mathcal{O}(\|d_\varepsilon\|_{L^2})$.

On utilise $\phi = p$ dans (1.18) et $d_\varepsilon = \tilde{v}$ dans (1.20) pour obtenir :

$$\|d_\varepsilon\|_{L^2(\Omega)}^2 = (\sigma_0 - \sigma^\star)\int_\Omega \mathbb{1}_{\varepsilon B}\nabla u_\varepsilon \nabla p.$$

On utilise à droite la borne $\|\nabla p\|_{L^q} = \mathcal{O}(\|d_\varepsilon\|_{L^2})$, la borne du paragraphe précédent $\|\mathbb{1}_{\varepsilon B}\nabla u_\varepsilon\|_{L^2} = \mathcal{O}(\varepsilon^{d/2})$ et une borne $\|\mathbb{1}_{\varepsilon B}\|_{L^q} = \mathcal{O}(\varepsilon^{\frac{d}{q}})$ avec $\frac{1}{q} + \frac{1}{2} + \frac{1}{p} = 1$, ce qui nous donne au final $\|d_\varepsilon\|_{L^2} = o(\varepsilon^{d/2})$.

Technique de polarisation Ensuite, on utilise une technique de polarisation. Il suffit d'introduire un problème de même nature que (1.17), en notant v_ε, v_0 les inconnues (respectivement perturbées et non perturbées) du nouveau problème et $\tilde{d}_\varepsilon = v_\varepsilon - v_0$. Ensuite, on pose une fonction test régulière ϕ et on utilise (1.17) pour obtenir la série d'approximations suivantes

$$\begin{aligned}\int_\Omega \phi\sigma_0 \nabla d_\varepsilon \nabla \tilde{d}_\varepsilon &= \int_\Omega \sigma_0 \nabla(\phi d_\varepsilon)\nabla \tilde{d}_\varepsilon - \int_\Omega \sigma_0 (\nabla\phi) d_\varepsilon \nabla \tilde{d}_\varepsilon \\ &= -k\int_\Omega \mathbb{1}_{\varepsilon B}\nabla v_0 \nabla(d_\varepsilon \phi) + o(\varepsilon) \\ &= -k\int_\Omega \mathbb{1}_{\varepsilon B}\nabla v_0 \nabla d_\varepsilon \phi - k\int_\Omega \mathbb{1}_{\varepsilon B}\nabla v_0 \nabla \phi d_\varepsilon + o(\varepsilon) \\ &= -k\int_\Omega \mathbb{1}_{\varepsilon B}\nabla v_0 \nabla d_\varepsilon \phi + o(\varepsilon)\end{aligned}$$

Où on utilise à répétition des bornes $\|d_\varepsilon\|_{L^2} = o(\varepsilon^{d/2})$,$\|\nabla d_\varepsilon\|_{L^2} = \mathcal{O}(\varepsilon^{d/2})$ et $\|\mathbb{1}_{\varepsilon B}\|_{L^2} = \mathcal{O}(\varepsilon^{d/2})$. En procédant de même et en changeant le rôle de d_ε et $\tilde{d}_\varepsilon$, on obtient pour tout ϕ.

$$\int_\Omega \mathbb{1}_{\varepsilon B}\nabla v_0 \nabla u_\varepsilon \phi = \int_\Omega \mathbb{1}_{\varepsilon B}\nabla u_0 \nabla v_\varepsilon \phi + o(\varepsilon^d) \tag{1.21}$$

En choisissant plusieurs v_0 tels que ∇v_0 soit un vecteur de la base canonique on trouve ensuite la formule (1.19), où $\mathcal{M}$ est défini de manière implicite avec les différentes limites faibles de $\mathbb{1}_{\varepsilon B}\nabla v_\varepsilon$.

Le gradient topologique Une fois que l'équation (1.19) est établie, on peut s'apercevoir que la limite de $\mathcal{M}$ existe quand $\sigma^\star$ tend vers 0 ou $+\infty$, on peut ainsi en déduire quelle serait la formule pour un trou de Neumann ou un trou de Dirichlet.

Bien évidemment, la remarque qui consiste à intervertir des limites en ε et $\sigma^\star$ ne forme pas une preuve et les démonstrations de gradient topologique sont plus complexes, mais font intervenir à un niveau ou à un autre ces fonctions tests v_ε et une notion de superconvergence similaire à (1.21) qui est la convergence au sens des mesures de l'énergie dans le lemme div-curl.

Le gradient topologique s'exprime par un tenseur $\mathcal{M}$ qui ne dépend que de la forme du trou et de sa nature (le paramètre $\sigma^\star$). Il est calculable point par point et indique si on doit creuser un trou à l'endroit donné et de quel forme il doit être creusé, pour peu que l'on soit capable de retrouver pour chaque tenseur admissible la forme du trou donné. Le calcul de l'ensemble des tenseurs admissibles se fait par des bornes de type Hashin-Shtrikman et ne sont pas disponibles pour tout type d'équations. Plus prosaïquement, on fixera l'ensemble des trous admissibles à l'avance, par exemple certaines ellipses et on calculera les tenseurs de polarisation admissibles.

1.2 [All+05] : Courbes de niveaux et gradient topologique

Dans [All+05], nous nous sommes principalement intéressés à mélanger la méthode des courbes de niveaux et celle du gradient topologique. En effet les travaux précédents [AJT02 ; AJT04] ou [SW00 ; OS01 ; WWG03] en optimisation de formes incluaient des méthodes d'advection de frontière uniquement et ne permettaient pas de modifier la topologie du domaine considéré. Numériquement, l'algorithme est capable de combler les trous mais pas de nucléer la forme. L'objectif de [All+05] était donc de présenter un algorithme incorporant en même temps l'advection de frontière et la nucléation par la méthode du gradient topologique.

L'idée de l'algorithme est relativement simple, il suffit de faire une itération de gradient topologique toutes les n itérations d'advection de frontière. Un exemple est montré dans la Figure 1.1 qui concerne l'optimisation d'un pylône électrique. L'équation de base est celle de l'élasticité linéaire et les contraintes sont que le pylône doit supporter des chargements sur ses bras (représentés par des flèches en haut à gauche de la Figure 1.1) et qu'il peut s'accrocher sur les coins de sa base (les noeuds de Dirichlet sont représentés par des boules en haut à gauche de la Figure 1.1). Le critère d'optimisation est ici la compliance avec une contrainte de volume gérée par un multiplicateur de Lagrange. On représente la forme à l'itération 6,11,16 et 21, juste après avoir effectué une nucléation par gradient topolgique.

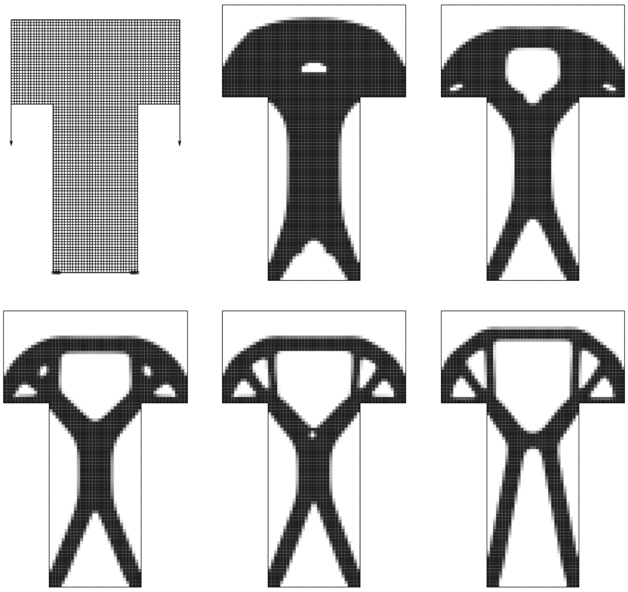


FIGURE 1.1 – Illustration de la Section 1.2 : Différentes étapes de l'évolution en $2d$ d'un problème d'optimisation de pylône, on représente le pylône juste après qu'un trou ait été nucléé dans la masse. La méthode level-set optimise la forme de ce trou et à convergence (après n itérations), on tente une nouvelle nucléation. Les itérations montrées sont celles juste après la nucléation.

L'intérêt de cette méthode est que nous montrons empiriquement que la forme optimale finale a une dépendance très faible à l'initialisation et ainsi on évite les nombreux minimas locaux. Nous faisons aussi des tests en $3d$ et montrons que le gradient topologique n'apporte pas grand chose dans nos tests dans ce cas par rapport à la méthode de level-set. La raison est qu'en $3d$ il est facile de créer un trou en pinçant deux côtés de la forme. En deux dimensions, même s'il est possible de créer un trou, il faut le créer depuis le bord et le faire évoluer vers l'intérieur de la forme, ce qui semble plus compliqué pour l'algorithme.

1.3 [Gou06],[GAJ05] : Régularisation de la vitesse

Dans [Gou06], nous répondons à deux problèmes. Le premier problème est de régulariser et le deuxième est d'étendre la vitesse d'advection V de la méthode des courbes de niveaux définie dans (1.13). Nous résolvons ces deux problèmes en changeant le produit scalaire qui donne le gradient de la vitesse. Nous montrons quel produit scalaire il convient de choisir et nous montrons qu'il accélère la vitesse de convergence de l'algorithme. Ces travaux sont dans la lignée des questions soulevées dans [Bur03 ; PBH04 ; MP10].

Pour tout problème d'optimisation de forme, le théorème de structure d'Hadamard énonce que la différentielle de la fonction coût dans la direction du champ de vecteur θ s'écrit ;

$$d\mathcal{J}.\dot{\theta} = \int_{\partial\Omega} (\dot{\theta} \cdot n) p,$$

Où p est la pression du gradient de forme. Parmi les termes de p, on retrouvera notamment, sur les bords où sont imposées des conditions de Neumann et pour l'élasticité linéaire un terme $Ae(u) \colon e(v)$ où v est l'adjoint de la fonction coût. La méthode des courbes de niveaux (1.13) nécessite de prendre $V = -p$ sur le bord de Ω ce qui donne bien au premier ordre une descente de gradient. Cependant pour l'équation d'advection de la courbe de niveau, on a besoin d'une vitesse V qui est définie en tout point de la grille cartésienne (ou sur toute cellule, cela dépend de la méthode numérique choisie pour résoudre (1.13)). En se concentrant sur le terme $Ae(u): e(v)$ de p, la méthode classique consistait à résoudre l'équation d'élasticité sur la même grille cartésienne qui définit ϕ et de mettre un matériau "mou" pour simuler les cellules vides. Le problème est que $Ae(u): e(v)$ n'a pas de réel sens physique dans le vide. Dans (1.13), nous choisissons plutôt d'imposer un espace de régularisation de la vitesse $\mathcal{H}$ avec un produit scalaire, ici

$$(V,W)_{\mathcal{H}} = \int_D \nabla V \cdot \nabla W + aVW \quad a > 0, \tag{1.22}$$

où D est la grille cartésienne et a un paramètre de régularisation et de résoudre

$$(V,W)_{\mathcal{H}} = -\int_{\partial\Omega} pW.$$

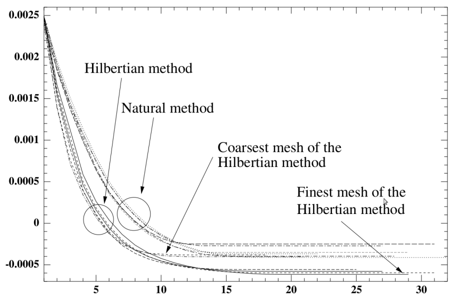


FIGURE 1.2 – Comparaison des courbes d'évolution de la compliance d'une console pour différents maillages avec une régularisation de la vitesse (Hilbertian method) et sans régularisation (Natural method).

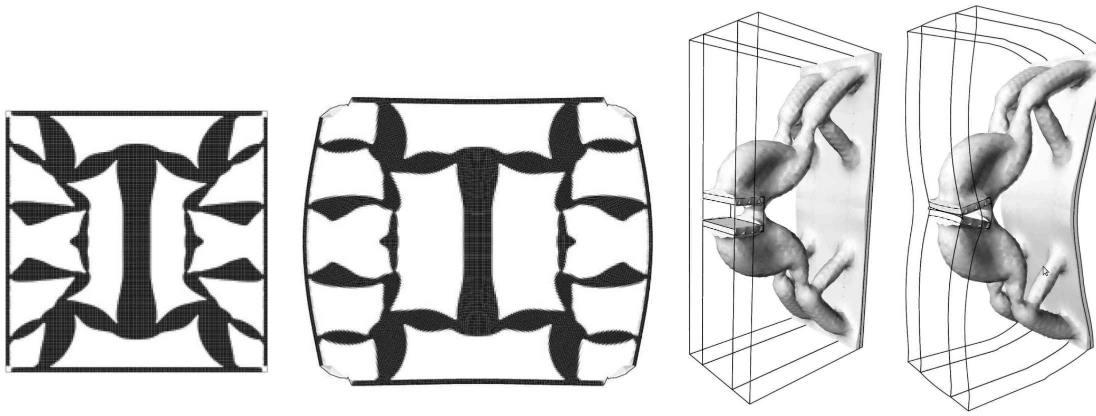


FIGURE 1.3 – Illustration de la Section 1.3 : Deux micro-mécanismes et leur configuration déformée. Le premier est un mécanisme à coefficient de Poisson négatif et le deuxième est une pince $3d$ activée par une pression sur son dos.

Cette méthode a l'avantage de diffuser la pression p en dehors de $\partial\Omega$ et de la régulariser. Le paramètre a, introduit dans (1.22), sert à régler la distance caractéristique de la diffusion et doit être choisi par l'utilisateur. La méthode peut aussi être interprétée comme une méthode de gradient dans l'espace $\mathcal{H}$. Nous montrons dans [Gou06] que la méthode de régularisation/extension de la vitesse améliore la convergence de la méthode des courbes de niveaux en temps de calcul et en valeur de l'optimum, voir Figure 1.2. L'amélioration en temps de calcul est due au paramètre a qui rend des vitesses plus régulières et pour lesquelles la CFL de la résolution numérique de (1.13) est plus lâche.

Dans [GAJ05] nous utilisons la méthode de régularisation de la vitesse pour l'advection de la level-set, avec la méthode de gradient topologique de [All+05] pour optimiser des micro-mécanismes compliants. Ce sont des mécanismes en dimension 2 qui peuvent être gravés à l'échelle micrométrique avec précision et qui doivent répondre à des cahiers des charges classiques de l'ingénierie [Sig97] (il faut des pinces en deux dimensions, des moteurs, des invertisseurs de force, des actuateurs linéaires, etc...). Nous utilisons nos deux méthodes pour produire des mécanismes compliants, j'ai choisi de représenter une pince en $3d$ qui est activée par une pression sur son dos et une mécanisme à coefficients de Poisson négatif qui est un mécanisme qui doit pousser sur les bords hauts etbas quand on tire sur les bords droits et gauche, voir Figure 1.3.

1.4 [Gou06] : Problèmes de valeurs propres

La question principale qui nous a interessée dans [Gou06] est d'optimiser par rapport au domaine la première valeur propre d'un opérateur quand elle est multiple, c'est-à-dire que l'espace propre associé est de dimension strictement plus grande que 1. L'archétype des problèmes est de trouver la forme Ω qui maximise la première valeur propre d'un système d'élasticité linéaire. En mettant des conditions de symétries bien choisies, on peut assurer que la première valeur propre ne sera pas multiple. Dans ce cas la première valeur propre cesse d'être différentiable et un algorithme idoine doit être trouvé.

Obtention du gradient de forme Dans [Gou06], nous obtenons la formule du gradient de forme de la première valeur propre de l'élasticité linéaire. Nous donnons dans l'équation suivante les termes principaux de la formule

Proposition 1.5 Soit un problème aux valeurs propres avec des opérateurs positifs $\mathcal{A}$ et $\mathcal{B}$ auto-adjoints dans un espace de Hilbert $\mathcal{H}$

$$\mathcal{A}u = \lambda \mathcal{B}u,$$

On suppose que $\mathcal{A}^{-1}$ est compact et $\mathcal{B}$ est continu, ainsi il existe une plus petite valeur propre λ de dimension finie. Soit $\mathcal{M}$ la sphère des vecteurs propres associés à λ renormalisés par $(\mathcal{B}u, u) = 1$. Si $\mathcal{A}$ et $\mathcal{B}$ sont différentiables par rapport à un paramètre θ, alors la dérivée directionnelle de λ par rapport à $\dot{\theta}$ est donnée par

$$\lambda'(\dot{\theta}) = \min_{u \in \mathcal{M}} (d\mathcal{A}.\dot{\theta}u, u) - \lambda(d\mathcal{B}.\dot{\theta}u, u) \tag{1.23}$$

Démonstration

La plus petite valeur propre minimize le quotient de Rayleigh et on a

$$\lambda = \min_u f(u) \text{ avec } f(u) = \frac{(\mathcal{A}u, u)}{(\mathcal{B}u, u)},$$

Le minimum étant réalisé sur E_λ, l'espace propre associé à λ. Si on perturbe un peu les opérateurs $\mathcal{A}$ et $\mathcal{B}$, on perturbe un peu f et seuls les vecteurs propres associés à λ, c'est-à-dire les minimiseurs actuels, sont candidats à être les nouveaux minimiseurs. Chacun de ces u contribue à modifier λ par $df(u).\dot{\theta}$ avec

$$df(u).\dot{\theta} = \frac{(d\mathcal{A}.\dot{\theta}u, u)(\mathcal{B}u, u) - (\mathcal{A}u, u)(d\mathcal{B}.\dot{\theta}u, u)}{(\mathcal{B}u, u)^2}.$$

Or λ est un minimum donc parmi les éléments de E_λ, celui qui gagne la compétition pour devenir le nouveau minimiseur est celui avec la plus petite dérivée. On obtient donc

$$\lambda'(\dot{\theta}) = \min_{u \in E_\lambda} df(u).\dot{\theta}$$

Il suffit de renormaliser u sur la sphère $\mathcal{M}$ pour conclure.

Il est à noter que le λ considéré n'est plus différentiable mais seulement directionnellement différentiable. Même si $d\mathcal{A}$ et $d\mathcal{B}$ sont linéaires par rapport à θ, λ' défini dans (1.23) ne l'est plus dès que la dimension de la sphère $\mathcal{M}$ (et donc de l'espace propre associé à λ) est strictement plus grande que 1.

Dans le cas de l'optimisation de forme, le théorème de structure de Hadamard s'applique et la formule de la dérivée directionnelle de forme de λ dans la direction $\dot{\theta}$ doit être du type

$$\lambda'(\theta) = \min_{u \in \mathcal{M}} \int_{\partial\dot{\Omega}} (\theta \cdot n) G(u, u), \tag{1.24}$$

avec $G(u, u)$ bilinéaire symétrique par rapport à u.

Choix de la direction de descente Nous nous intéressons ensuite au choix de la direction de descente et nous nous arrêtons sur une méthode de plus profonde montée pour maximiser λ. Moralement, nous allons résoudre un problème du type

$$\max_{\theta, \|\theta\|=1} \lambda'(\theta) + \int_{\partial\Omega} (\theta \cdot n) p \tag{1.25}$$

Le vecteur p représente des gradients par rapport à d'autres termes de la fonctions coût totale ou des contraintes (comme des contraintes de volume sur la forme) ramenées dans la fonction coût via leur multiplicateur de Lagrange.

Proposition 1.6 Soit s la dimension de l'espace propre E_λ et $t = \frac{s(s+1)}{2} + 1$. Il existe des coefficients r_{ijk}, c_k avec $1 \le i, j \le s$ et $1 \le k \le t$ tels que la résolution du problème (1.25) soit équivalente à la résolution du problème suivant

$$\max_{\|v\|_{\ell^2} \le 1} \min_{\|a\|_{\ell^2} = 1} \sum_{ijk} a_i a_j v_k r_{ijk} + c_k v_k. \tag{1.26}$$

Ce problème est un problème SDP (optimisation semi-définie positive) de petite dimension.

Démonstration

En notant $(e_i)_i$ une base orthonormale de vecteurs propres dans $\mathcal{M}$ et en notant $r_{ij} = G(e_i, e_j)$, la formule (1.24) se réécrit

$$\lambda'(\theta) = \min_{a \in \mathbb{R}^s \|a\|_{\ell^2} = 1} \sum_{i,j} a_i a_j \int_{\partial\Omega} (\theta \cdot n) r_{ij},$$

En suivant les idées de la Section 1.22 et en introduisant le produit scalaire (1.22), on introduit R_{ij} et P qui résolvent

$$(R_{ij}, W)_{\mathcal{H}} = \int_{\partial\Omega} r_{ij} W \quad (P, W)_{\mathcal{H}} = \int_{\partial\Omega} pW \quad \forall W.$$

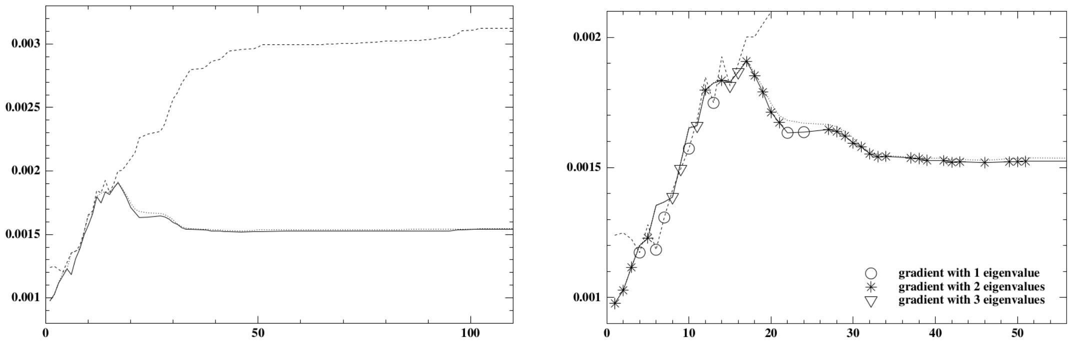


FIGURE 1.4 – Illustration de la Section 1.4 : Evolution des trois premières valeurs propres d'un problème d'élasticité linéaire $3d$ (gauche) et nombre de valeurs propres prises en compte dans le calcul de la direction de descente.

L'algorithme de plus profonde montée (1.25) se résume à trouver une vitesse d'advection de la méthode de levelset qui résout

$$\max_{V\in\mathcal{H},\|V\|_{\mathcal{H}}\leq 1}\ \min_{\|a\|_{\ell^2}=1}\sum_{ij} a_i a_j (R_{ij}, V)_H + (V, P)_H.$$

Ce problème est en fait posé en dimension finie, car seule la projection orthogonale de V sur $Vect((R_{ij})_{ij}, P)$ compte dans le calcul de l'optimum. On note $(E_k)_{k\leq t}$ une base orthonormée de $Vect((R_{ij})_{ij}, P)$ pour le produit scalaire de H, $(r_{ijk})_k$ les coordonnées de R_{ij} dans cette base et c_k les coordonnées de C. On cherche donc $V = \sum_k v_k E_k$ tel que le vecteur $v = (v_k)_k \in \mathbb{R}^t$ vérifie

$$\max_{\|v\|_{\ell^2}\leq 1}\ \min_{\|a\|_{\ell^2}=1}\sum_{ijk} a_i a_j v_k r_{ijk} + c_k v_k.$$

Ce problème est de petite dimension, le vecteur a est de taille s, la dimension de l'espace propre de λ. Le vecteur v est de taille $t = \frac{s(s+1)}{2} + 1$. Le calcul des R_{ij} ou de P ne pose pas de problème car le produit scalaire de $\mathcal{H}$ est le même au cours des itérations et le calcul des R_{ij} ou de P est le coût d'un algorithme de descente-remontée pour peu qu'une factorisation LU du produit scalaire ait été faite en amont.

Pour résoudre le problème (1.26), nous utilisons une boîte noire du type SDP (optimisation semi-définie positive), qui consiste à minimiser par rapport à Y une fonction coût linéaire (Y, Y_0) sous contrainte qu'une certaine matrice $E(Y)$ dont les coefficients dépendent linéairement de Y reste dans le cône des matrices positives

$$\min_{Y, E(Y)\geq 0}(Y, Y_0).$$

Ici, on introduit les variable $w, z \in \mathbb{R}$ et on prend $Y = [v, w, z]$, $Y_0 = [0, -1, 0]$ et

$$E(Y) = \begin{pmatrix} A(Y) - zId & 0 & 0 & 0 \\ 0 & C(Y) + z - w & 0 & 0 \\ 0 & 0 & Id & v \\ 0 & 0 & v^T & 1 \end{pmatrix},$$

où $A(Y)_{ij} = \sum_k r_{ijk} v_k$ et $C(Y) = \sum_k c_k v_k$. Il est facile de voir que la contrainte $E(Y) \geq 0$ est exactement les contraintes $\|v\|_{\ell^2} \leq 1$, $w \leq z + v_k c_k$ et $z \leq \min_{\|a\|_{\ell^2}=1} a_i a_j A(Y)_{ij}$. Comme on cherche à maximiser w, on résout bien le problème initial.

Dans la pratique, aucune valeur propre n'est jamais multiple, on considère donc $\mathcal{M}$ comme une boule dans l'espace engendré par les vecteurs propres dont les valeurs propres sont comprises entre λ et $(1 + \varepsilon) * \lambda$.

On donne aussi dans [Gou06] un certain nombre de tests numériques, ceux que je préfère sont illustrés dans la Figure 1.4, surtout entre les itérations 5 et 20. L'algorithme a trois valeurs propres à optimiser en même temps dont deux sont issues de la symétrie du problème. Entre les itérations 5 et 20, l'algorithme semble hésiter entre prendre une des valeurs propres ou alors prendre le couple symétrique ou encore prendre les 3 en même temps. Cela est du au paramètre de la tolérance qui considère deux valeurs propres comme égales. A l'itération 12, l'algorithme prend le couple symétrique, ce qui a un effet désastreux sur la dernière valeur propre qui oscille. Heureusement il se reprend et dans les itérations 16 à 18 optimise les 3 valeurs propres en même temps. Ensuite la valeur propre qui n'est pas issue de la symétrie s'éloigne et ne fait plus partie des préoccupations de l'algorithme qui vit sa vie avec le couple symétrique. Une brisure de la symétrie dues aux erreurs numériques advient vers l'itération 22 mais tout rentre dans l'ordre assez vite.

1.5 [GAJ08 ; AGJ09] : Compliance robuste

Dans [GAJ08], nous nous intéressons à la problématique d'optimiser la compliance d'un domaine dont le chargement peut être soumis à des perturbations, plus particulièrement à optimiser le pire des cas. Ce travail s'inscrit dans la lignée des travaux [CC99 ; CC04 ; All12 ; Che12].

Étant donné un opérateur $\mathcal{A}$, défini positif, à résolvante compacte un terme source f, une taille maximale des perturbations m, un opérateur de localisation des perturbations $\mathcal{B}$ continu et des perturbations $s = \mathcal{B}g$, le problème de compliance robuste revient à trouver la perturbation de f qui maximise la compliance de l'opérateur (ou "l'énergie")

$$\max_{s=\mathcal{B}g, \|g\|\leq m} \mathcal{C}(f+s) \quad \text{avec } \mathcal{C}(f) = \max_u -\frac{1}{2}(\mathcal{A}u,u)+(f,u),$$

Où la norme de $\|g\|$ est une norme issue d'un produit scalaire. Physiquement, étant donné un terme source f perturbé par s, le problème de compliance robuste revient à calculer la perturbation qui réalise le pire des cas de la compliance. C'est un problème de recherche de perturbation dite "pessimiste". L'objectif de [GAJ08] est de trouver des méthodes efficaces de calcul de cette compliance robuste et d'optimiser par rapport au domaine la compliance robuste. Le problème peut se re-écrire

$$\begin{aligned} \max_{\|g\|\leq m} \max_u -\frac{1}{2}(\mathcal{A}u,u)+(f+\mathcal{B}g,u) &= \max_u -\frac{1}{2}(\mathcal{A}u,u)+(f,u)+m\|\mathcal{B}^\star u\| \\ &= \max_{\|g\|\leq m} \frac{1}{2}(\mathcal{A}^{-1}(f+\mathcal{B}g), f+\mathcal{B}g) \end{aligned}$$

Par nature, c'est un problème quadratique en g avec contraintes quadratiques de dimension infinie. Nous étudions dans un premier temps la résolution du problème direct. La meilleure formulation de ce problème est d'introduire ρ, une variable duale de la contrainte $\|g\| \leq m$, de considérer le problème aux valeurs propres généralisé

$$\mathcal{A}e_i = \rho_i \mathcal{B}\mathcal{B}^\star e_i \tag{1.27}$$

dont la plus petite valeur propre ρ_0 est strictement positive et de noter E_{ρ_i} les espaces propres associés. Pour tout ρ, on calcule u_ρ la solution (si elle existe) de

$$(\mathcal{A} - \rho\mathcal{B}\mathcal{B}^\star)u = f.$$

Quand ρ est une valeur propre du problème généralisé, alors u_ρ existe si et seulement si f est orthogonal à E_ρ et on note u_ρ la solution orthogonale à E_ρ.

Proposition 1.7 L'algorithme de résolution du problème de compliance robuste est le suivant

1. Si f n'est pas orthogonal à E_{ρ_0} aller à l'étape (3).

2. On calcule u_{ρ_0}. Si $\|\mathcal{B}u_{\rho_0}\| \leq \dfrac{m}{\rho_0}$ alors l'ensemble des solutions est donné par
$$\{u_{\rho_0} + v, v \in E_0, \|\mathcal{B}v\| = \frac{m}{\rho_0} - \|\mathcal{B}u_{\rho_0}\|\}.$$
Finir l'algorithme
3. La fonction $g(\rho) = (f, u_\rho) + \frac{m^2}{\rho}$ est convexe et admet un unique point critique $\rho^\star$ sur $]0, \rho_0[$. La solution du problème est
$$\{u_{\rho^\star}\}.$$

Comment trouver $\rho^\star$. Dans le cas où l'algorithme nous indique qu'il faut minimiser la fonction convexe g, on utilise une méthode de Newton. Le calcul de u_ρ nécessite une résolution de $\mathcal{A} - \rho\mathcal{B}\mathcal{B}^\star$ et la formule de la dérivée seconde de g par rapport à ρ nécessite de calculer
$$v_\rho = \mathcal{A}^{-1}\mathcal{B}\mathcal{B}^\star(\mathcal{A} - \rho\mathcal{B}\mathcal{B}^\star)^{-1}\mathcal{A}u_\rho$$
L'obstruction usuelle au déploiement d'une méthode de Newton est le surcoût de calcul de la Hessienne. Ici, le surcoût n'est pas excessif, $\mathcal{A}$ a déjà été factorisé/préconditionné pour calculer le problème aux valeurs propres généralisé, de même que $\mathcal{A} - \rho\mathcal{B}\mathcal{B}^\star$ pour le calcul de u_ρ. Cependant chaque itération en ρ pour trouver $\rho^\star$, le minimum de g est relativement coûteuse car elle nécessite une factorisation de la matrice $\mathcal{A} - \rho\mathcal{B}\mathcal{B}^\star$. Afin d'accélérer la convergence de l'algorithme de Newton, on donne dans [GAJ08] de bonnes bornes sur $\rho^\star$ qui permettent de commencer les itérations plus proches de $\rho^\star$. On note $(e_i)_{i\in\mathbb{N}}$ les vecteurs propres du problème (1.27) associés à des valeurs propres non nulles. En renormalisant ces vecteurs propres par $(\mathcal{A}e_i, e_j) = \delta_{ij}$, en notant $f_i = (f, e_i)$ les coordonnées de f, et en notant πf la projection A-orthogonale de f sur le noyau de $\mathcal{B}\mathcal{B}^\star$, on obtient :
$$g(\rho) = \sum_{i\in\mathbb{N}} \frac{\rho_i (f_i)^2}{\rho_i - \rho} + \frac{m^2}{\rho} + \|\pi f\|^2$$
En supposant que les n premières valeurs propres du problème (1.27) sont calculées, on s'intéresse au problèmes de minimisations de g_m et g_M donnés par
$$g_M(\rho) = \sum_{i\leq n} \frac{\rho_i (f_i)^2}{\rho_i - \rho} + \frac{m^2}{\rho} \qquad g_m(\rho) = ((\mathcal{A}^{-1}f, f) - \sum_{i\leq n} f_i^2)\frac{\rho_n}{\rho_n - \rho} + g_M(\rho).$$
On montre dans [GAJ08] que ces deux problèmes d'optimisation convexe ont une solution respectivement notées ρ_m et ρ_M sur $]0, \rho_0[$ et que ces deux solutions fournissent un encadrement de $\rho^\star$ le minimum de g, au sens où $\rho_m \leq \rho^\star \leq \rho_M$. Les calculs d'optimisation sur g_M et g_m sont triviaux et la borne donnée permet d'initialiser correctement le problème de Newton de calcul de $\rho^\star$.

Optimiser la compliance robuste par rapport au domaine. Dans [GAJ08], nous nous intéressons aussi à la dérivation par rapport au domaine de la compliance robuste. Pour le calcul effectif de la dérivée, il vaut mieux regarder la formule de la compliance robuste donnée par
$$\mathcal{J} = \max_u -\frac{1}{2}(\mathcal{A}u, u) + (f, u) + m\|\mathcal{B}^\star u\| \tag{1.28}$$
on note $\mathcal{M}$ l'ensemble des maximiseurs de la compliance robuste et une dérivation formelle par rapport au domaine nous donne en notant $\mathcal{J}', \mathcal{A}', f', B'$ les dérivées directionnelle de forme par rapport à une direction θ
$$\mathcal{J}' = \max_{u\in\mathcal{M}} -\frac{1}{2}(\mathcal{A}'u, u) + (f', u) + \frac{m}{\|\mathcal{B}^\star u\|^2}(\mathcal{B}'\mathcal{B}^\star u, u)$$

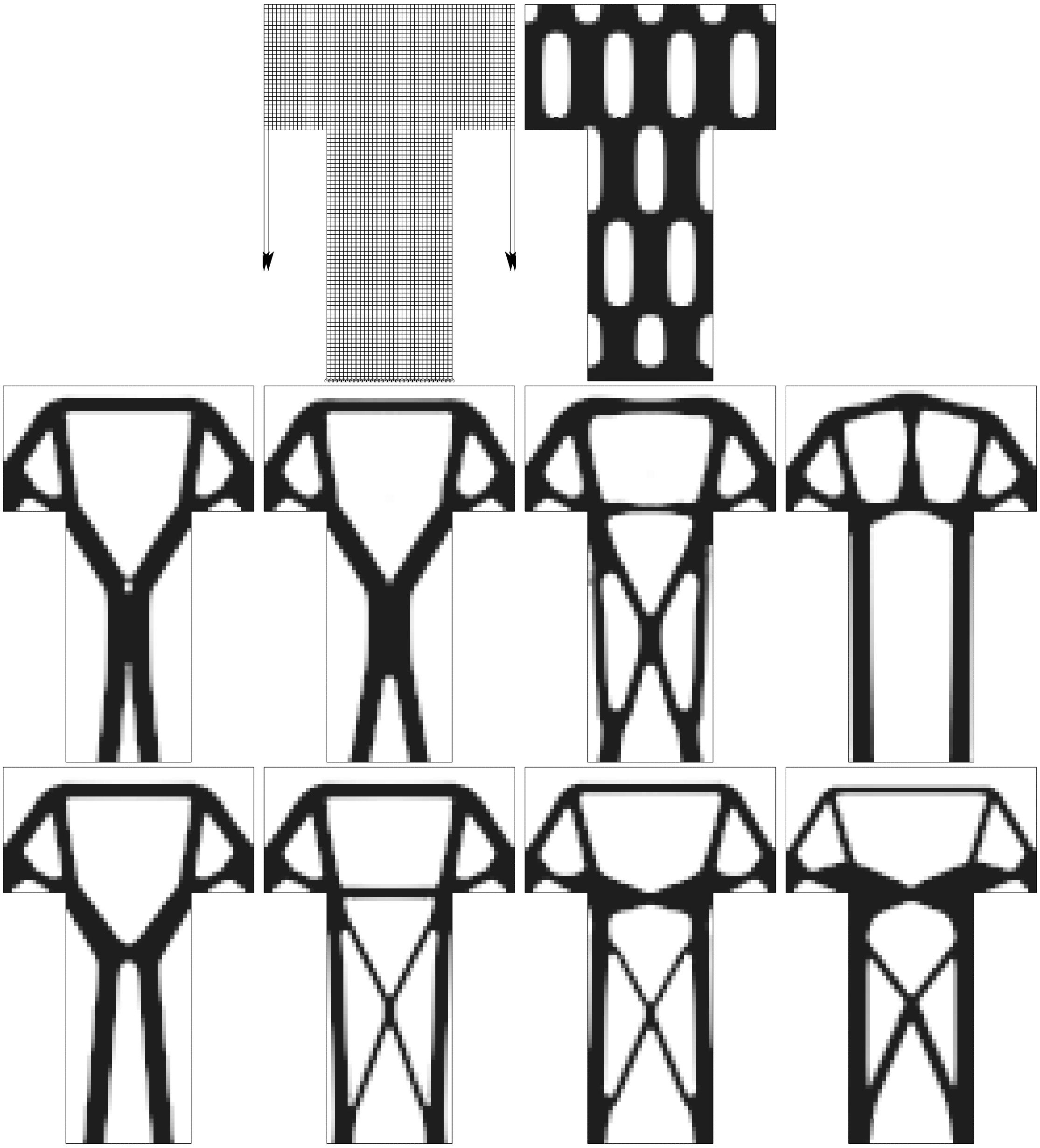


FIGURE 1.5 – Illustration de la Section 1.5 : L'exemple du pylone perturbé. En haut le maillage avec les forces appliquées (gauche) et l'initialisation (droite). Sur la ligne du milieu des perturbations verticales sont autorisées là où les forces sont appliquées. Sur la ligne du bas, des perturbations horizontales et verticales sont autorisées là où les forces sont appliquées. Les perturbations augmentent de la gauche vers la droite

On montre dans [GAJ08] que la formule (1.28) de dérivée formelle est juste et on l'exprime comme

$$\mathcal{J}' = \max_{u \in \mathcal{M}} \int_{\partial\Omega} (\theta \cdot n)(v(u,u) + l(u))$$

ici $\mathcal{M}$ est une sphère qui n'est pas centrée en 0, v est bilinéaire et l est linéaire. Le calcul du gradient se fait comme dans la Section 1.4 où l'on traitait le cas de valeurs propres multiples et où l'ensemble $\mathcal{M}$ était une sphère centrée en 0 et où $l = 0$. La compliance robuste n'est qu'une

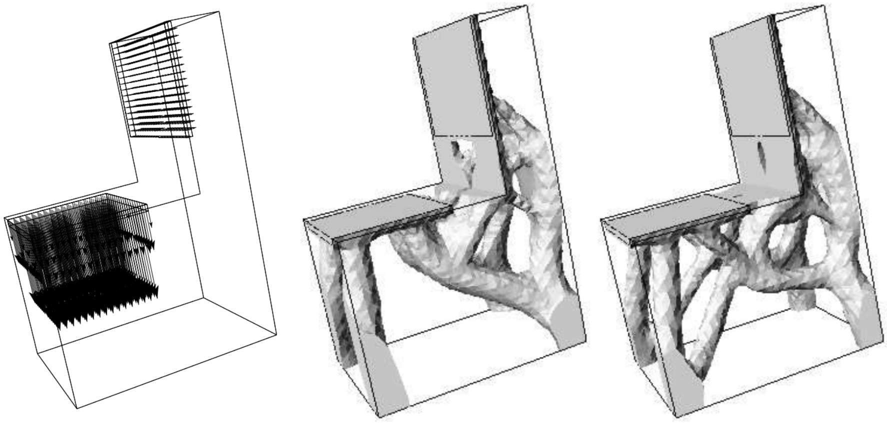


FIGURE 1.6 – Illustration de la Section 1.5 : Test de la chaise robuste, les chargements sont montrés à gauche, les perturbations sont localisées au même endroit que les chargements et sont supposées être dans la même direction. Au milieu, un facteur de perturbation $m = 0$ et à droite un facteur de perturbations $m = 1$.

extension des valeurs propres multiples et les arguments s'appliquent quasiment verbatim. Un exemple de résultat type est donné dans la Figure 1.5.

Dans [AGJ09], nous étendons la méthode à d'autres tests et nous regardons aussi l'optimisation de puissances L^p des contraintes, le test de la "chaise robuste" est présenté en Figure 1.6.

1.6 [FG18] Méthode de Gauss Newton

La question que nous nous sommes posés dans [FG18] était d'essayer d'implémenter une méthode de Gauss-Newton en optimisation de forme. La méthode de Gauss-Newton est une méthode d'ordre 2 qui ne nécessite pas de calcul de Hessienne. L'avantage de ne pas à avoir à calculer des dérivées secondes est que l'espace des formes atteignables par difféomorphisme est une variété non-Riemannienne et il n'existe pas de définition intrinsèque de la dérivée seconde dans un tel espace. Des notions d'optimisation d'ordre 2 peuvent quand même être développées et ont été tentées dans [NR00 ; Roc05 ; EH05 ; KK14 ; ACV12]. Dans [FG18] nous voulions implémenter une méthode de Gauss-Newton à l'optimisation de forme.

Définition 1.8 La méthode de Gauss-Newton est une méthode d'optimisation réservée aux problèmes de moindres carrés de la forme

$$\min_{\theta} \frac{1}{2}\|F(\theta)\|^2$$

et s'écrit

$$\theta_{k+1} = \theta_k + d_k \qquad \text{avec } (dF^\star \circ dF)d_k = -dF^\star F,$$

où dF est la différentielle de F au point θ_k et $dF^\star$ son adjoint. C'est une approximation de la méthode de Newton où l'application linéaire $(dF^\star \circ dF)$ remplace la Hessienne de la fonction objectif. Le terme $dF^\star F$ est bien le gradient de la fonction objectif.

Nous l'avons d'abord essayé sur un critère de moindre carré issu d'un problème inverse d'identification de trou. On suppose que pour un trou donné on a mesuré la solution d'une EDP sur un domaine d'observation ω et on cherche à retrouver la forme du trou. Pour cela on minimise par rapport à θ la fonction

$$\|F_1(\theta)\|^2 = \int_\omega |u(\theta) - u_t|^2,$$

où u_t sont les données et $u(\theta)$ est la solution d'une EDP posée sur le domaine $\Omega(\theta)$, avec $\omega \subset \Omega(\theta)$. Le produit scalaire utilisé est le produit scalaire L^2 sur ω ce qui donne $F_1(\theta) = u(\theta) - u_t$. Dans un deuxième temps, nous avons ensuite essayé la méthode de Gauss-Newton sur un problème de compliance avec contrainte de volume en écrivant

$$\|F_2(\theta)\|^2 = \frac{1}{2}\int_{\Omega_\theta} A\nabla u_\theta \colon \nabla u_\theta + \frac{\lambda}{2}\int_{\Omega_\theta} 1, \tag{1.29}$$

Où $u_\theta \in H^1_D(\Omega_\theta)$, le symbole D indique que l'on met des conditions des Dirichlet sur une partie de la frontière de Ω_θ et u_θ vérifie l'équation

$$\int_{\Omega_\theta} A\nabla u_\theta \nabla v = \int_{\Omega_\theta} fv \quad \forall v \in H^1_D(\Omega_\theta).$$

Le produit scalaire utilisé ici est

$$\langle (m, v), (\tilde{m}, \tilde{v})\rangle_{\mathcal{M}} = \frac{1}{2}\int_\Omega Am \colon \tilde{m} + \lambda \int_\Omega v \cdot \tilde{v}, \tag{1.30}$$

où $m, \tilde{m}$ sont des champs de vecteurs et $v, \tilde{v}$ des champs scalaires. Après avoir fait une intégration par partie pour ramener les intégrales dans (1.29) à des intégrales sur le domaine Ω, on obtient la formulation suivante pour $F_2(\theta) = (m_\theta, v_\theta)$

$$T = Id + \theta \quad m_\theta = \sqrt{|\det \nabla T|}\nabla u_\theta \nabla T^{-1} \quad v_\theta = \sqrt{|\det \nabla T|}.$$

L'introduction d'un terme de moindre carré dans le cas de la compliance est complètement artificielle et nous fûmes relativement surpris de l'excellent comportement de l'algorithme.

Calcul de dF et $dF^\star$ Nous montrerons le calcul dans le cas $F = F_2$. En utilisant le changement de variable usuel de l'optimisation de formes, on établit une formulation variationnelle pour $\tilde{u}_\theta = u_\theta \circ (Id + \theta)$ de la forme

$$\int_\Omega C(\theta)\nabla\tilde{u}_\theta \nabla v = \int_\Omega f(\theta) v$$

et on peut dériver par rapport à θ, en notant $du.\theta, df.\theta, dC.\theta$, les dérivées directionnelles, on obtient

$$\int_\Omega A\nabla(du.\theta)\nabla v = \int_\Omega -dC.\theta\nabla u \nabla v + \int_\Omega (df.\theta)v \tag{1.31}$$

En écrivant $F(\theta) = (F_1(\theta)\nabla\tilde{u}_\theta, F_2(\theta))$, le calcul de $dF.\theta$ est donné par

$$dF.\theta = (dF_1.\theta\nabla u + \nabla(du.\theta), dF_2.\theta)$$

Ainsi le calcul de $dF.\theta$ nécéssite de trouver $du.\theta$ donc de résoudre le problème (1.31).

Pour calculer $dF^\star$, il faut se donner un produit scalaire $\langle \bullet, \bullet\rangle_\Theta$ sur l'ensemble des paramètres (ici on choisit le produit scalaire H^1 sur les champs de vecteurs θ) et par définition, pour $z = (m, v)$:

$$\langle dF^\star z, \theta\rangle_\Theta = \langle z, dF.\theta\rangle_{\mathcal{M}} = \frac{1}{2}\int_\Omega Am \colon (dF_1.\theta\nabla u + \nabla(du.\theta)) + \lambda\int_\Omega + \lambda\frac{1}{2}\int_\Omega \operatorname{div}(\theta)v,$$

où le produit scalaire $\mathcal{M}$ est défini dans (1.30).

On se débarasse du terme $\nabla(du.\theta)$ en introduisant un adjoint p qui résout

$$\int_\Omega Am \colon \nabla\phi = \int_\Omega A\nabla\phi \cdot \nabla p$$

Puis on utilise l'adjoint pour transformer le terme $\int_\Omega Am \colon \nabla\phi$ en $\int_\Omega A\nabla(du.\theta)\nabla p$ et on utilise (1.31) pour obtenir une formule

$$\langle dF^\star z, \theta\rangle_\Theta = \int_\Omega l(\theta)$$

où l est linéaire et ne dépend que de z. On se sert ensuite de la représentation de θ par éléments finis pour calculer la forme linéaire $\theta \to \int_\Omega l(\theta)$ et on relève cette forme linéaire par le produit scalaire Θ pour obtenir $dF^\star$. Comme le produit scalaire Θ est indépendent de l'optimisation, il est factorisé une fois pour toute et inverser sa matrice est facile. Le calcul de la forme linéaire est juste un passage dans la méthode d'intégration éléments finis, au final le calcul de $dF^\star z$ ne nécessite que le calcul de l'adjoint p. Ainsi le calcul de $dF^\star z$ et de $dF.\theta$ nécessitent une résolution de système linéaire.

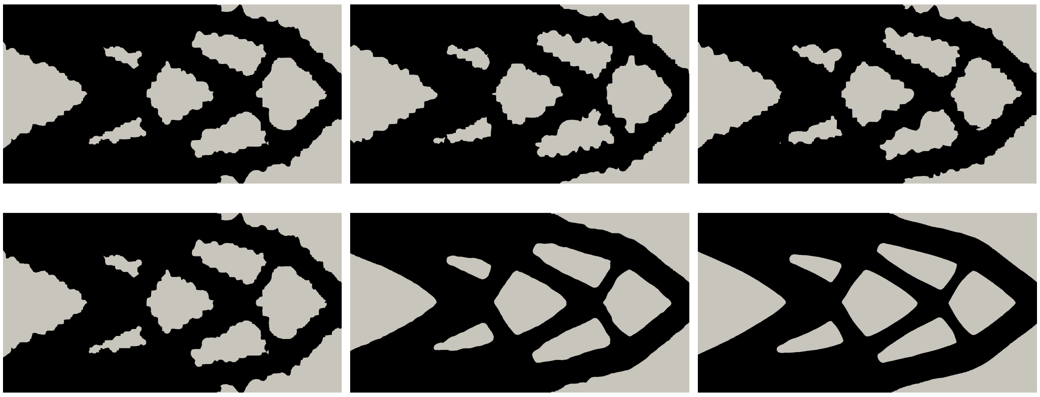


FIGURE 1.7 – Illustration de la Section 1.6 : Pour la compliance de la console, capacité de l'algorithme de Gauss-Newton à sortir des oscillations. En partant (à gauche) d'un problème oscillant on lance 2 itérations d'optimisation. En haut avec un algorithme de gradient et en bas avec un algorithme de Gauss-Newton. L'algorihtme de Gradient a bien plus du mal à compenser les oscillations pour trouver la direction de descente tandis l'algorithme de Gauss-Newton détruit les oscillations.

Calcul de la direction de descente Il est à noter que le calcul des opérateurs linéaires dF et $dF^\star$ n'est pas possible, nous n'avons accès qu'au produit matrice-vecteur $dF.\theta$ et $dF^\star z$ pour tout z, θ. De plus ces produits matrice-vecteur nécessitent une résolution du problème de l'EDP sous-jacente. Pour trouver la direction de descente, nous utilisons un algorithme itératif (GMRES en l'occurrence) qui ne requiert que la connaissance du produit matrice-vecteur avec un faible nombre d'itérations (10 en l'occurrence). Ainsi un calcul de direction de descente coûte 20 fois plus cher qu'un calcul d'évaluation de la fonctionnelle (et 10 fois plus cher qu'un calcul de gradient classique qui lui ne nécessite qu'un calcul d'adjoint). Ceci doit être pondéré par le fait que nous factorisons la matrice de l'EDP et que son inversion n'est plus que le coût d'un algorithme de descente/remontée. Cependant cette technique n'est pas disponible pour les EDP paraboliques par exemple.

Résultats Le comportement de l'algorithme est celui attendu, le nombre d'itération pour arriver à convergence diminue drastiquement, la convergence est beaucoup plus rapide mais le coût de chaque itération est augmenté par la résolution du système de Gauss-Newton donnant la direction d_k. Nous avons cependant montré aue les deux méthodes de gradient et de Gauss-Newton sont comparables en temps de calcul global et que la méthode de Gauss-Newton exhibe bien les structures relativement classiques des méthodes d'ordre 2. Cependant la dynamique d'évolution de la forme dans le cas de la méthode de Gauss-Newton est beaucoup plus satisfaisante, voir Figure 1.7. Dans le

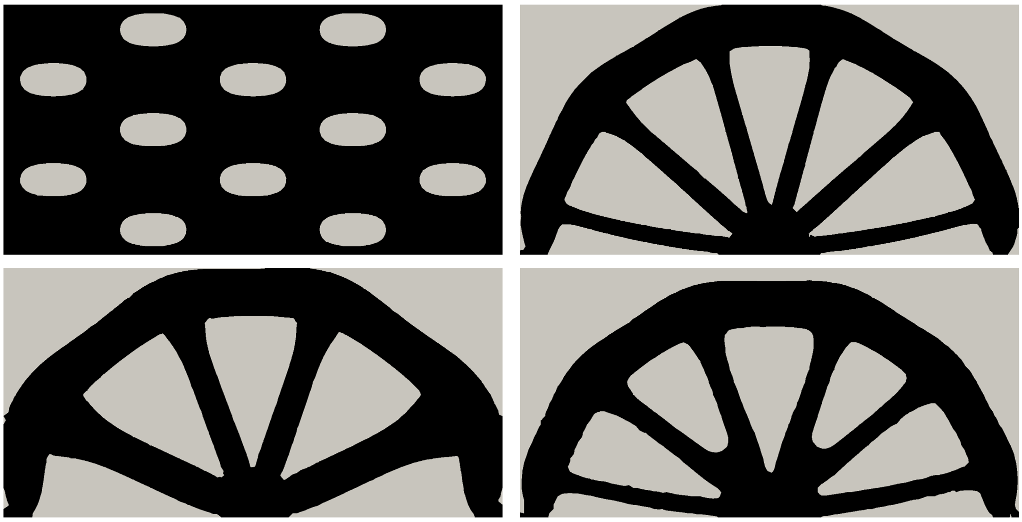


FIGURE 1.8 – Illustration de la Section 1.6 : Initialisation et formes finales pour l'optimisation d'une arche pour, respectivement, la méthode de Gauss-Newton,l'algorithme du gradient et un algorithme de régularisation de la vitesse.

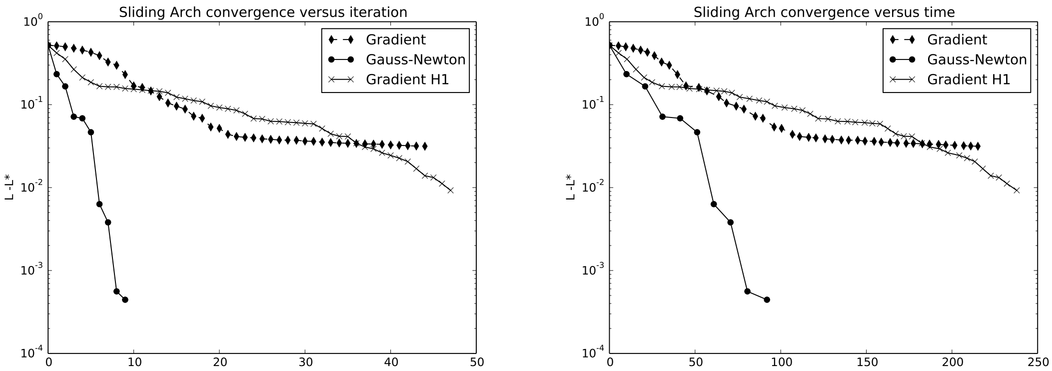


FIGURE 1.9 – Illustration de la Section 1.5 : Evolution de la convergence de la compliance pour le problème de l'arche. On trace en échelle logarithmique la différence entre la fonction objectif et sa valeur numérique optimale (obtenue pour Gauss-Newton). A gauche, les abscisses sont le nombre d'itérations et à droite les abscisses sont le temps de calcul. Les algorithmes considérés sont Gauss-Newton, l'algorithme du gradient et un algorithme de régularisation de la vitesse (H^1).

cas de la compliance de la console, une division par un facteur 5 du nombre d'itérations d'évolution est observée et la forme reste moins longtemps bloquée sur des minimas locaux voir Figure 1.7 et Figure 1.8 et Figure 1.9

1.7 Perspectives

J'ai deux travaux non-publiés sur le sujet de l'optimisation de formes, un sur l'optimisation du flambement et un sur l'optimisation de sensibilité. Le flambement est une approximation (linéaire) de la bifurcation d'un système mécanique non-linéaire. L'optimisation de sensibilité consiste à optimiser la forme d'un échantillon d'un certain matériau sur lequel on va faire une expérience physique (par exemple, on le soumet à un chargement et on observe la réponse du matériau) qui nous permet -via un problème inverse- de retrouver des paramètres du matériau. Ainsi une *expérience physique* se décompose ici en une prise de mesures physiques et un problème inverse d'assimilation de paramètres. La question de l'optimiseur de forme est de savoir comment optimiser le domaine pour que l'*expérience physique* soit la plus discrimante possible (retrouve le mieux les paramètres). C'est un problème d'optimisation bi-niveau car le problème inverse qui nous permet de retrouver le paramètre est souvent un problème d'optimisation en soi.

Une deuxième question est de transposer les techniques d'ordre 2 (les différentes flaveurs de Newton) au problème d'optimisation de formes. Nous nous y essayons un peu dans [FG18], mais il reste beaucoup à faire. Ce qui est intéressant est que l'espace naturel dans lequel le problème est posé est des domaines atteignables par difféomorphisme $W^{1,\infty}$ qui est une variété non-riemannienne de dimension infinie. On pourrait déjà se poser dans le cas riemannien (en utilisant des produits scalaires similaires à [Gou06]) pour essayer d'aborder le problème.

Une limitation des travaux de [Gou06] est l'introduction d'un produit scalaire pour travailler dans $W^{1,\infty}$. On devrait plutôt utiliser des techniques d'optimisation dans des espaces de Banach. Une contribution récente intéressante est [Her23].

Bibliographie

[All12] Grégoire Allaire. *Shape optimization by the homogenization method*. T. 146. Springer Science & Business Media, 2012.

[ACV12] Grégoire Allaire, Eric Cancès et Jean-Léopold Vié. "Second-order shape derivatives along normal trajectories, governed by Hamilton-Jacobi equations". In : *Calcolo* 49.3 (2012), p. 193-219.

[AGJ09] Grégoire Allaire, Frédéric de Gournay et François Jouve. "Stress minimization and robust compliance optimization of structures by the level set method". In : *8th World Congress on Structural and Multidisciplinary Optimization*. 2009.

[All+05] Grégoire Allaire, Frédéric de Gournay, François Jouve et A-M Toader. "Structural optimization using topological and shape sensitivity via a level set method". In : *Control and cybernetics* 34.1 (2005), p. 59.

[AJT02] Grégoire Allaire, François Jouve et Anca-Maria Toader. "A level-set method for shape optimization". In : *Comptes Rendus Mathematique* 334.12 (2002), p. 1125-1130.

[AJT04] Grégoire Allaire, François Jouve et Anca-Maria Toader. "Structural optimization using sensitivity analysis and a level-set method". In : *Journal of computational physics* 194.1 (2004), p. 363-393.

[AS07] Grégoire Allaire et Marc Schoenauer. *Conception optimale de structures*. T. 58. Springer, 2007.

[Bur03] Martin Burger. "A framework for the construction of level set methods for shape optimization and reconstruction". In : *Interfaces and Free boundaries* 5.3 (2003), p. 301-329.

[CV03] Yves Capdeboscq et Michael S Vogelius. "A general representation formula for boundary voltage perturbations caused by internal conductivity inhomogeneities of low volume fraction". In : *ESAIM : Mathematical Modelling and Numerical Analysis* 37.1 (2003), p. 159-173.

[Céa+00] Jean Céa, Stéphane Garreau, Philippe Guillaume et Mohamed Masmoudi. "The shape and topological optimizations connection". In : *Computer methods in applied mechanics and engineering* 188.4 (2000), p. 713-726.

[Che12] Andrej Cherkaev. *Variational methods for structural optimization*. T. 140. Springer Science & Business Media, 2012.

[CC99] Andrej CHERKAEV et Elena CHERKAEVA. “Optimal design for uncertain loading condition”. In : *Homogenization : In Memory of Serguei Kozlov*. World Scientific, 1999, p. 193-213.

[CC04] Elena CHERKAEV et Andrej CHERKAEV. “Principal compliance and robust optimal design”. In : *The Rational Spirit in Modern Continuum Mechanics*. Springer, 2004, p. 169-196.

[EH05] Karsten EPPLER et Helmut HARBRECHT. “A regularized Newton method in electrical impedance tomography using shape Hessian information”. In : *Control and Cybernetics* 34.1 (2005), p. 203.

[EKS94] Hans A ESCHENAUER, Vladimir V KOBELEV et Axel SCHUMACHER. “Bubble method for topology and shape optimization of structures”. In : *Structural optimization* 8.1 (1994), p. 42-51.

[FG18] Jérôme FEHRENBACH et Frédéric de GOURNAY. “Shape optimization via a levelset and a Gauss-Newton method”. In : *ESAIM : Control, Optimisation and Calculus of Variations* (2018).

[GGM01] Stéphane GARREAU, Philippe GUILLAUME et Mohamed MASMOUDI. “The topological asymptotic for PDE systems : the elasticity case”. In : *SIAM journal on control and optimization* 39.6 (2001), p. 1756-1778.

[Gou06] Frédéric de GOURNAY. “Velocity extension for the level-set method and multiple eigenvalues in shape optimization”. In : *SIAM journal on control and optimization* 45.1 (2006), p. 343-367.

[GAJ05] Frédéric de GOURNAY, Grégoire ALLAIRE et François JOUVE. “Optimisation de forme de micro-mécanismes compliants par la méthode des courbes de niveaux”. In : t. 2. Actes du 7ème colloque national en calcul des structures. 2005, p. 229-234.

[GAJ08] Frédéric de GOURNAY, Grégoire ALLAIRE et François JOUVE. “Shape and topology optimization of the robust compliance via the level set method”. In : *ESAIM : Control, Optimisation and Calculus of Variations* 14.1 (2008), p. 43-70.

[Her23] Philip J HERBERT. “Shape Optimisation with $W^{1,\infty}$: A connection between the steepest descent and Optimal Transport”. In : *arXiv preprint arXiv :2301.07994* (2023).

[KK14] Henry KASUMBA et Karl KUNISCH. “On computation of the shape Hessian of the cost functional without shape sensitivity of the state variable”. In : *Journal of Optimization Theory and Applications* 162.3 (2014), p. 779-804.

[MP10] Bijan MOHAMMADI et Olivier PIRONNEAU. *Applied shape optimization for fluids*. Oxford university press, 2010.

[MS75] François MURAT et Jacques SIMON. “Etude de problèmes d’optimal design”. In : *IFIP Technical Conference on Optimization Techniques*. Springer. 1975, p. 54-62.

[NR00] A NOVRUZI et JR ROCHE. “Newton’s method in shape optimisation : a three-dimensional case”. In : *BIT Numerical Mathematics* 40.1 (2000), p. 102-120.

[OS01] Stanley J OSHER et Fadil SANTOSA. “Level set methods for optimization problems involving geometry and constraints : I. frequencies of a two-density inhomogeneous drum”. In : *Journal of Computational Physics* 171.1 (2001), p. 272-288.

[Pir12] Olivier PIRONNEAU. *Optimal shape design for elliptic systems*. Springer Science & Business Media, 2012.

[PBH04] Bartosz PROTAS, Thomas R BEWLEY et Greg HAGEN. “A computational framework for the regularization of adjoint analysis in multiscale PDE systems”. In : *Journal of Computational Physics* 195.1 (2004), p. 49-89.

[Roc05] Jean Rodolphe ROCHE. “Adaptative Newton-like Method for Shape Optimization”. In : *Control and Cybernetics* 34 (2005), p. 363-377.

[SW00] James A Sethian et Andreas Wiegmann. “Structural boundary design via level set and immersed interface methods”. In : *Journal of computational physics* 163.2 (2000), p. 489-528.

[Sig97] Ole Sigmund. “On the design of compliant mechanisms using topology optimization”. In : *Journal of Structural Mechanics* 25.4 (1997), p. 493-524.

[Sim80] Jacques Simon. “Differentiation with respect to the domain in boundary value problems”. In : *Numerical Functional Analysis and Optimization* 2.7-8 (1980), p. 649-687.

[SZ99] Jan Sokolowski et Antoni Zochowski. “On the topological derivative in shape optimization”. In : *SIAM journal on control and optimization* 37.4 (1999), p. 1251-1272.

[SZ01] Jan Sokolowski et Antoni Zochowski. “Topological derivatives of shape functionals for elasticity systems”. In : *Optimal Control of Complex Structures*. Springer, 2001, p. 231-244.

[SZ92] Jan Sokolowski et Jean-Paul Zolesio. “Introduction to shape optimization”. In : *Introduction to Shape Optimization*. Springer, 1992, p. 5-12.

[WWG03] Michael Yu Wang, Xiaoming Wang et Dongming Guo. “A level set method for structural topology optimization”. In : *Computer methods in applied mechanics and engineering* 192.1-2 (2003), p. 227-246.

CHAPITRE 2

Différents problèmes inverses

Dans ce chapitre nous détaillons plusieurs problèmes que j'ai eu à résoudre à la sortie de ma thèse. Ils ont tous en commun d'appartenir à la classe des problèmes inverses d'identification. C'est-à-dire qu'on cherche à partir d'informations sur la solution d'une EDP à retrouver des informations sur l'EDP en question, par exemple ses paramètres. Ils sont formalisés comme des problèmes d'optimisation sur les paramètres de l'EDP en prenant comme fonction objectif un terme dit "d'attache aux données" qui mesure la différence entre les donnée observées expérimentalement et les données simulées avec le jeu de paramètre en considération.

2.1 [Cap+09] Le 0-Laplacien

Le premier problème qui nous intéresse est celui du 0-Laplacien. Considérons un appareil de tomographie par impédance électrique, c'est à dire un appareil capable d'infliger des courants électriques dans le corps humain via des électrodes et de mesurer le couple courant/potentiel sur ces électrodes. On suppose donc que nous avons à notre disposition plusieurs $(u_i)_{i=1..p}$ solutions de

$$div(\sigma\nabla u_i) = 0 \text{ dans } \Omega, \tag{2.1}$$

tels que $I_i = \int_{\partial\Omega} u_i(\sigma\nabla u_i \cdot n)$ est connu avec précision. On couple cet appareil avec un échographe, qui est capable de focaliser des ultra-sons sur une zone prédéterminée du corps humain, disons autour d'un point x_0. On postule que cette compression change la conductivité électrique localement de manière déterministe. On note σ^ε la conductivité perturbée et u_i^ε le potentiel électrique perturbé par les ultrasons. On peut supposer que le potentiel du dispositif n'a pas été modifié, on a donc $u_i^\varepsilon = u_i$ sur le bord. On suppose que l'énergie électrique perturbée I_i^ε est mesurée avec précision.

En utilisant la formule du tenseur de polarisation donnée dans la Proposition 1.4, on peut trouver un développement de l'énergie (produit courant/potentiel) dépensée dans les électrodes au premier ordre. Si ε est le rayon de la perturbation, la différence d'énergie électrique injectée dans le système, vaut

$$I_i^\varepsilon - I_i = \mathcal{M}\nabla u_i(x_0)\nabla u_i(x_0),$$

avec $\mathcal{M}$ un tenseur supposé connu et dépendant de $\sigma(x_0)$. Puis il suffit de faire varier la position de x_0, pour connaître $\sigma\nabla u \cdot \nabla u$.

Dans [Cap+09] nous nous sommes interessés, dans la lignée de [Amm+08] à la question suivante : Étant donné plusieurs $(u_i)_{i=1..p}$ solutions de (2.1), peut-on retrouver la carte σ depuis les données de $\sigma\nabla u_i \cdot \nabla u_i$? On remarque tout d'abord qu'en infligeant des courants électriques de la forme I,J,$I+J$, on est capable de calculer, par la formule du parallélogramme les données $\sigma\nabla u_i \cdot \nabla u_j$

pour tout i, j et x_0. Ici u_i (resp. u_j) est le potentiel électrique de réponse au courant I (resp J) sur le jeu d'électrodes.

Si $\sigma\nabla u_i \cdot \nabla u_j$ est connu sur $\omega \subset \Omega$, est-il possible retrouver σ sur ω ? La question peut être reformulée comme étant le 0-Laplacien [1] car si $d_{ij} = \sigma\nabla u_i \cdot \nabla u_j$, il s'agit de résoudre

$$div(d_{ii}\frac{\nabla u_i}{|\nabla u_i|^2}) = 0 \text{ dans } \omega,$$

ce qui correspond au p-Laplacien pour $p = 0$.

Presque unicité Nous montrons dans [Cap+09] qu'il y a presque unicité de σ en dimension 2. L'hypothèse la plus importante est qu'il faut que presque partout dans ω, la famille des $(\nabla u_i)_i$ soit génératrice de $\mathbb{R}^2$. De plus il faut rajouter des hypothèses techniques, comme le fait que $|\nabla u_i|^{-1}$ doit être dans $L^2(\omega)$. Sous ces hypothèses, on peut calculer un jeu de direction $d_i(x)$ tel qu'il existe une rotation R (inconnue) telle que $d_i = R\nabla u_i/\|\nabla u_i\|$. En d'autres termes, on connaît les directions des gradients (à une rotation près). Si cette rotation est connue, alors, en réutilisant les données, le vecteur $s_i = \sqrt{\sigma}\nabla u_i$ est connu et ensuite σ lui même est calculable à une constante multiplicative près. Cette preuve d'unicité est directe, au sens où elle se base sur des formules de reconstruction.

Reconstruction La preuve d'unicité n'est pas une preuve de stabilité, elle est même inutilisable en pratique car elle nécessite de dériver les données. Nous nous sommes donc naturellement posé la question dans [Cap+09] d'implémenter un algorithme de reconstruction de σ. Nous sommes partis d'un problème d'identification de moindres carrés et nous comparons deux méthodes

$$\min_\sigma \mathcal{J}_1(\sigma) = \sum_i \int_\omega j(\sigma|\nabla u_i|^2, d_{ii}) \text{ ou } \min_\sigma \mathcal{J}_2(\sigma) = \sum_{i,k}(\int_{\omega_k} j(\sigma|\nabla u_i|^2, d_{ii}))^2$$

où $(a, b) \to j(a, b)$ est un terme d'accroche aux données et u_i vérifie (2.1).

Le problème en $\mathcal{J}_1$ est résolu par une descente de gradient, dans ce cas, nous nous arrêtons sur $j(a, b) = (\sqrt{a} - \sqrt{b})^2$ car la variable a n'est moralement que $L^2(\Omega)$. Concernant la formulation en $\mathcal{J}_2$, nous utilisons une méthode de Gauss-Newton, dans ce cas nous prenons plutôt $j(a, b) = a - b$ et ω_k est un simplexe du maillage. Nous avons étudié ces deux algorithmes et leurs avantages respectifs

Algorithme $\mathcal{J}_1$. Le calcul du gradient de $\mathcal{J}_1$ est fait dans [Cap+09], il nécessite un calcul d'adjoint. Nous avons aussi montré que dans le cas des algorithmes de gradient, la fonctionnelle $\mathcal{J}_1$ était presque strictement convexe aux minimums globaux (dans le cas $a = b$) au sens où, en notant $\tau = \ln(\sigma)$, pour tout a, il existe deux constantes C et c telles que :

$$\sum_i \frac{1}{2}(\int_\omega h_i)^{-1}\left(\int_\omega h_i\delta\right)^2 \le (\frac{\partial^2\mathcal{J}}{\partial\tau^2}\delta, \delta) \le \frac{1}{2}\int_\omega \sum_i h_i\delta^2 \qquad \text{avec } h_i = \sigma|\nabla u_i|^2.$$

On donne un exemple dans [Cap+09] où la Hessienne de $\mathcal{J}_1$ n'est pas définie positive aux minimums globaux, car essentiellement les directions de descente δ sont orthogonales aux h_i. L'algorithme de calcul d'optimisation sur $\mathcal{J}_1$ est un algorithme de gradient classique, il faut faire attention à ne pas optimiser sur σ mais plutôt à faire le changement de variable $\tau = \ln(\sigma)$, cela améliore drastiquement la vitesse de convergence.

Nous testons numériquement notre algorithme sur différents cas, nous rajoutons aussi du bruit pour tester la résistance au bruit additif gaussien dans les données. Nous nous intéressons aussi aux erreurs faites sur la forme du domaine Ω et sur les valeurs de u au bord. Effectivement, dans la pratique, la valeur de u au bord n'est pas connue car il n'y pas d'électrodes sur l'intégralité

1. Mais c'est essentiellement une blague de Y. Capdeboscq.

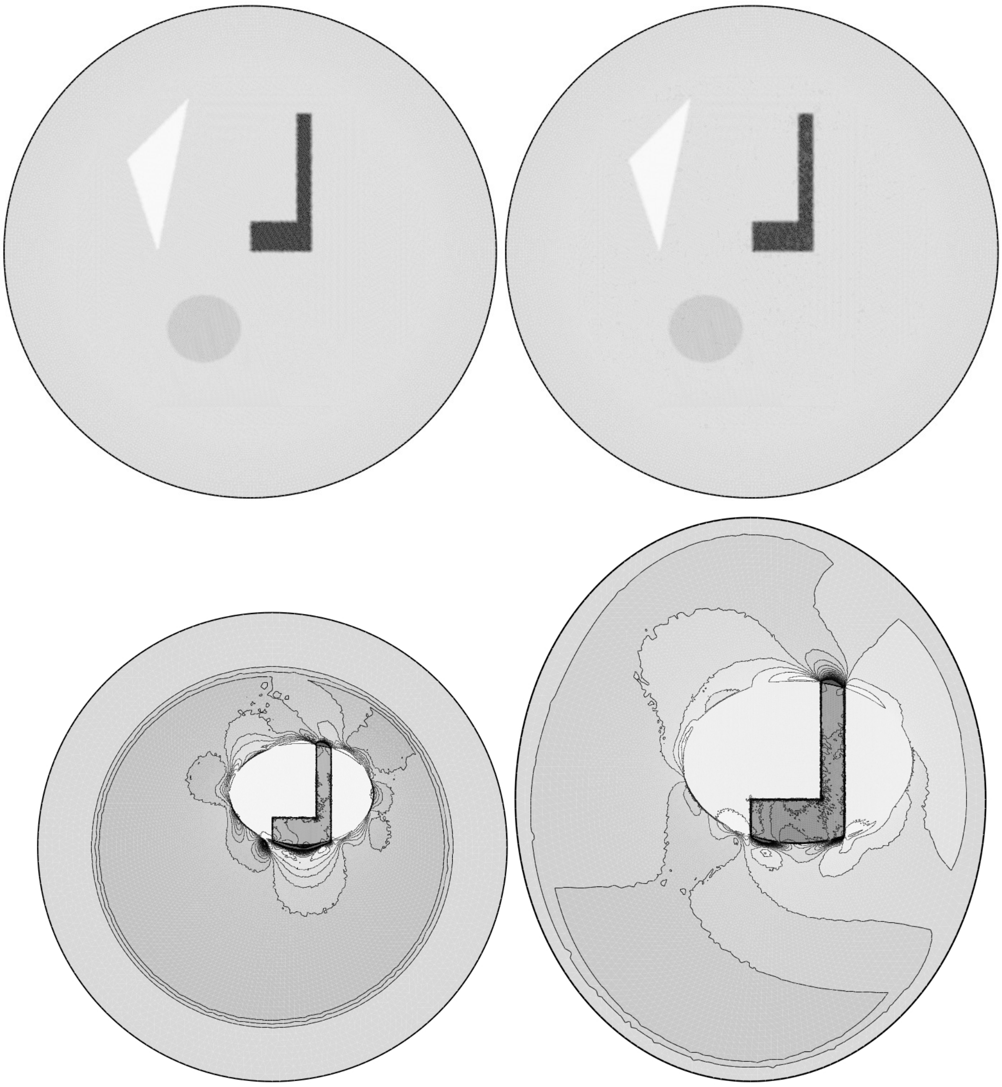


FIGURE 2.1 – Reconstruction d'une conductivité $2d$ par une méthode de gradient sur $\mathcal{J}_1$ dans le cas du 0-Laplacien. En haut à gauche, la configuration de référence, en haut à droite, la reconstruction pour 10% de bruit, l'erreur en norme L^2 est inférieure à 1%. En bas à gauche, reconstruction avec données partielles et en bas à droite, une erreur conséquente est rajoutée sur le bord du domaine. Les données ne sont supposées connues que sur une ellipse où la conductivité est bien retrouvée à un facteur multiplicatif près.

du corps et la modélisation de l'interface peau / air peut changer en fonction des conditions atmosphériques. La physiologie du patient est essentiellement inconnue et non modélisée. Nous montrons empiriquement que le problème est local au sens où il ne dépend pas de la forme extérieure Ω et ne dépend pas des valeurs de u sur $\partial\Omega$. Effectivement l'algorithme réussi à retrouver la conductivité (à un facteur multiplicatif près) à l'intérieur de la zone d'observation. Ces effets sont illustrés dans la Figure 2.1.

Nous essayons aussi de paralléliser l'algorithme d'optimisation par une méthode de recouvrement de domaine à la Schwartz. Nous calculons une itération d'algorithme de gradient pour $\mathcal{J}_1$ dans des sous domaines ω_1 et ω_2 qui nous donnent τ_1 dans ω_1 et τ_2 dans ω_2 et nous posons $\tau = \frac{\tau_1+\tau_2}{2}$ dans $\omega_1 \cap \omega_2$. Nous montrons empiriquement que cet algorithme est parallélisable.

Algorithme $\mathcal{J}_2$. Pour $\mathcal{J}_2$, nous utilisons un algorithme de Gauss-Newton, qui consiste à d'abord faire le changement de variable $\tau = \ln(\sigma)$ et à produire les itérations

$$\tau_{k+1} = \tau_k + d_k \quad \text{avec } (dF^\star \circ dF) d_k = -dF^\star F, \tag{2.2}$$

où dF est la différentielle de $F : \tau \mapsto (F_{ik}(\tau))_{ik}$ avec $1 \leq i \leq s$ et $1 \leq k \leq m$

$$F_{ik}(\tau) = \int_{\omega_k} e^\tau |\nabla u_i|^2 - d_{ii}.$$

Ici, s est le nombre de mesures de potentiel, il est relativement faible et m est le nombre de pixel de discrétisation de τ, il a vocation à être élevé. Pour n'importe direction X ou Y , le calcul de $dF.X$ ou de $dF^\star.Y$ nécessite de calculer un adjoint pour chaque i ce qui fait s calculs du problème direct.

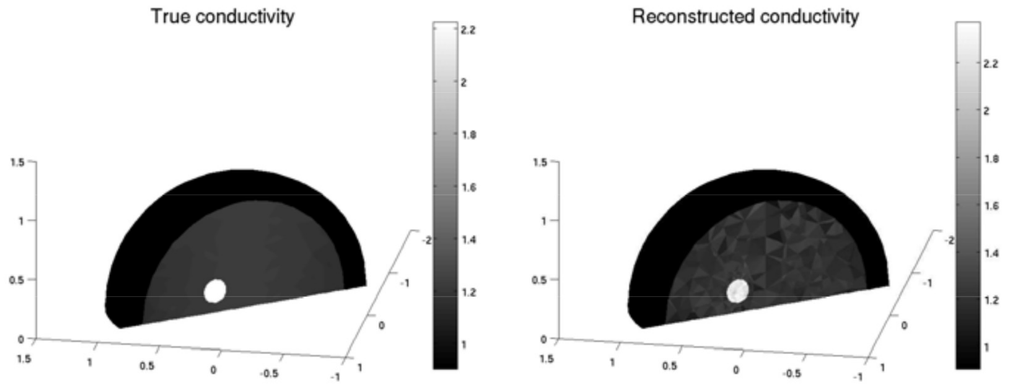


FIGURE 2.2 – Reconstruction d'une conductivité par la méthode de Gauss-Newton (2.2) dans le cas du 0-Laplacien. Les mesures sont obtenues avec 4 électrodes et 2% de bruit. L'erreur en norme L^2 est d'environ 4%.

On utilise ensuite un solveur pour calculer d_k et nous avons deux options. La première est de calculer toute la matrice dF via le produit $dF.X$ pour des vecteurs de la base canonique, ce qui fait ms calculs de problème direct, et utiliser un solveur direct pour trouver d_k. La deuxième option est de faire un calcul iteratif, chaque itération nous coûte $2s$ calculs de problème direct.

Nous avons décidé de faire une méthode multi-résolution en partant de m faible, de choisir la première méthode pour calculer les directions de descente, et de lancer un algorithme de Gauss-Newton. A convergence de l'algorithme de Gauss-Newton, on raffine le maillage de ω_k . Quand m devient trop grand, on décide d'utiliser la deuxième méthode pour calculer les directions de descente dans l'algorithme de Gauss-Newton. Un exemple de reconstruction en $3d$ est montré dans la Figure 2.2.

2.2 [Ber+09] : Retrouver des cylindres fins

Dans [Ber+09], nous nous intéressons à perturber une équation d'électrostatique par un petit cylindre ω_ε dont l'aire de la section tend vers 0 quand ε tend vers 0. Nous nous plaçons dans un contexte très proche de celui du tenseur de polarisation (Proposition 1.4) où ω_ε est contenu dans une boule dont le rayon tend vers 0. D'après [CV03], on sait que si u_ε suit une équation de Laplace à condition de Neumann sur le bord avec conductivité perturbée sur ω_ε :

$$-\,div(\sigma_0 + 1_{\omega_\varepsilon}\sigma_1)\nabla u_\varepsilon = 0 \text{ dans } \Omega \quad \sigma_0 \partial_n u_\varepsilon = g, \int_\Omega u_\varepsilon = 0 \tag{2.3}$$

Alors, si ω_ε tend vers 0 en volume, il existe un tenseur de polarisation $\mathcal{M}$ qui vérifie

$$(u_\varepsilon - u_0)(y) = |\omega_\varepsilon| \int_\Omega (\sigma_1 - \sigma_0)\mathcal{M}\nabla u_0 \nabla N_y d\nu + o(|\omega_\varepsilon|),$$

où M est un tenseur, $d\nu$ est la mesure limite de $|\omega_\varepsilon|^{-1}1_{\omega_\varepsilon}$ et N_y est la fonction de Green du domaine non pertubé qui vérifie

$$-div(\sigma_0 \nabla N_y) = 0 \text{ dans } \Omega \quad \sigma_0 \partial_n N_y = -\delta_y + |\Omega^{-1}|, \int_\Omega N_y = 0.$$

On remarque que grâce à la régularité elliptique du Laplacien, N_y et u_0 sont régulières sur $d\nu$.

La première contribution de [Ber+09] a été de calculer ce tenseur $\mathcal{M}$ dans le cas où ω_ε est un cylindre que l'on définit par sa section d_ε, son axe τ, sa longueur 2ℓ et son point median x_m comme suit, si (e_1, e_2, τ) est un repère orthonormé :

$$\omega_\varepsilon = \{x_m + \varepsilon(ae_1 + be_2) + z\tau, z \in [-\ell, \ell], (a, b) \in d_1\}$$

Dans ce cas le tenseur de polarisation $\mathcal{M}$ est constant et vaut $\tau \times \tau + m$, où m est le tenseur de polarisation bi-dimensionnel (dans $e \times e$) associé à d_1. De plus les inégalités du tenseur d'énergie nous assurent que

$$\begin{cases} \frac{\sigma_0}{\sigma_1} \geq \mathcal{M} \geq Id & \text{dans le cas } \sigma_0 \geq \sigma_1 \\ \frac{\sigma_0}{\sigma_1} \leq \mathcal{M} \leq Id & \text{dans le cas } \sigma_0 \leq \sigma_1. \end{cases}$$

Dans les deux cas, la direction τ est donnée par soit la plus grande, soit la plus petite des valeurs propres de $\mathcal{M}$.

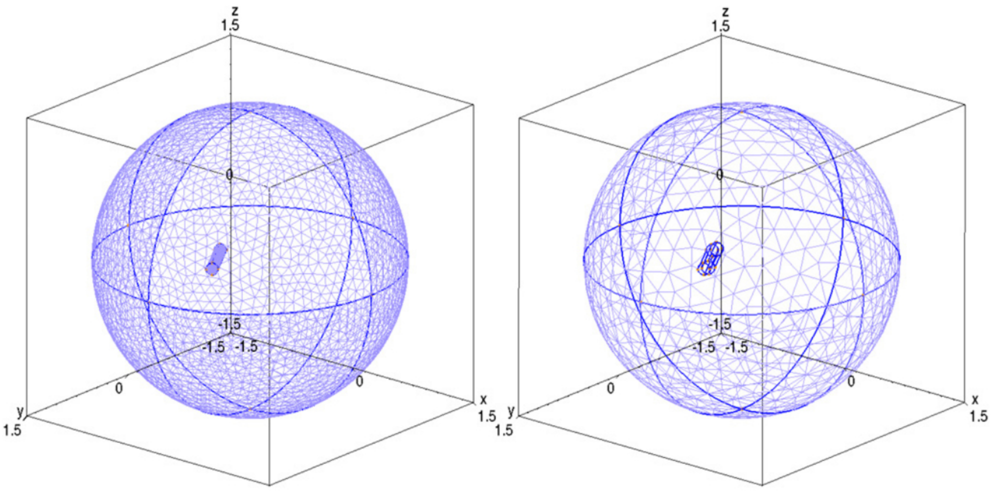


FIGURE 2.3 – Reconstruction d'un cylindre sur lequel une conductivité est perturbée, [Ber+09]. A gauche, le maillage de syntèses des données et le cylindre, à droite la reconstruction.

La deuxième contribution de [Ber+09] a été de donner un algorithme de reconstruction du cylindre dans le cas $\sigma_0 = 1$ si l'on ne connaît que $(u_i)_{i=1..d}$ sur $\partial\Omega$ où u_i vérifie l'équation 2.3 avec $g(x) = e_i \cdot n$. L'algorithme est relativement simple, il s'agit de tester u_i sur le bord de Ω avec differentes fonctions test harmoniques ψ pour obtenir des informations géométriques sur $\mathcal{M}$ et $d\nu$. Par exemple le cas $\psi(x) = (x \cdot e_j)$:

$$\int_{\partial\Omega} u_i(x \cdot e_j) \simeq |\omega_\varepsilon|(1 - \sigma_1)\mathcal{M}e_i \cdot e_j$$

Dans le cas $\sigma_1 \geq 1$, la plus grande valeur propre donne la valeur de $|\omega_\varepsilon|(1-\sigma_1)$ et le vecteur propre associé donne τ. On note ensuite $u_\tau = \sum_i u_i \tau_i$ En utilisant $\psi(x) = (x \cdot \tau)^2 - (x \cdot e_1)^2$, on retrouve $x_m \cdot \tau$. Ensuite en utilisant $\psi_a(x) = |x-a|^{-1}$, on peut retrouver x_m,ℓ et finalement le tenseur m. Ce dernier algorithme est intéressant car on calcule par méthode de dichotomie des maximums globaux sous contrainte de

$$a \mapsto \int_{\partial\Omega} u_\tau(x \cdot e_j),$$

qui correspondent à des paramètres physiques dépendant de x_m, ℓ et m. il s'intéresse à Un exemple en $3d$ est donné dans la Figure 2.3.

2.3 [Cin+10] : Placement optimal d'électrodes

L'article [Cin+10] répond à une question d'électro-chimiothérapie qui consiste à affaiblir localement la barrière phospholipidique des cellules cancéreuses en appliquant un courant électrique. L'affaiblissement de cette barrière, connu sous le nom d'électroporation permet une meilleure absorption des drogues chimiothérapiques. L'objectif du praticien est donc de contrôler spatialement l'intensité du champ électrique pour choisir les cellules qui recevront les drogues [Mar+06].

Nous avons simulé le problème en résolvant une équation de Laplace avec un modèle d'électrodes (dénotées $(V_i)_{i=1..n}$) parfaitement conductrices et une hypothèse de condition de Neumann homogène au bord du domaine.

$$\mathrm{div}(\sigma \nabla u) = 0 \text{ sur } \Omega \setminus V_i, \quad \partial_n u|_{\partial\Omega} = 0 \quad u|_{V_i} = a_i.$$

La carte de conductivité σ est donnée ainsi que $\omega \subset \Omega$ qui est la zone à électroporer. Le critère d'électroporation est un critère à seuil, l'objectif est que $|\nabla u| < E_{rev}$ sur $\Omega \backslash \omega$ et $E_{rev} < |\nabla u| < E_{irr}$ sur ω. Le seuil E_{rev} est le seuil à partir duquel l'électroporation est possible et le seuil E_{irr} est le seuil à partir duquel le courant électrique commence à brûler les cellules. Pour résoudre ce problème, nous minimisons la fonctionnelle objectif suivante

$$\int_{\Omega \backslash V} g(|\nabla u|(x) - 1_\omega(x) \frac{E_{rev} + E_{irr}}{2}, x) dx,$$

où $g(\alpha, x)$ est quadratique en α. Les paramètres d'optimisation sont la position et l'orientation des électrodes (leur forme est supposée fixée) et l'intensité a_i des courants électriques. L'optimisation par rapport aux a_i est de loin la plus facile dans la mesure où u est linéaire par rapport aux a_i et que les a_i vivent dans un espace de dimension faible. L'optimisation par rapport aux électrodes V_i se fait par une méthode de gradient où le calcul de la dérivée par rapport à la position et à l'orientation des électrodes se fait avec les techniques usuelles de dérivée de forme. Par une formule de Leibniz, on peut ensuite facilement en déduire la dérivée par rapport à un paramètre donné.

2.4 [Amm+11] : La version Helmoltz du 0 Laplacien

Dans [Amm+11], nous nous intéressons au même problème que dans [Cap+09] sauf que nous considérons l'équation d'Helmoltz au lieu du Laplacien. Ainsi, soit u_i qui résout :

$$div(\sigma \nabla u_i) + k_i^2 q u_i = 0 \text{ dans } \Omega.$$

Sachant que $d_{ii} = \sigma |\nabla u_i|^2$,$k_i$ et $e_{ii} = q|u_i|^2$ sont connus sur $\omega \subset \Omega$, est-il possible de retrouver σ et q sur Ω ?

Nous nous autorisons plusieurs mesures, pour différentes fréquences k_i et différentes valeurs de u_i au bord. Concernant l'existence et l'unicité, nous donnons des formules de reconstruction exactes dans le cas où $\Omega \subset \mathbb{R}^2$. Pour cela, il faut au moins 3 mesures (possiblement à des fréquences k différentes) et qu'elles soient non seulement non orthogonales, mais que le déterminant de la

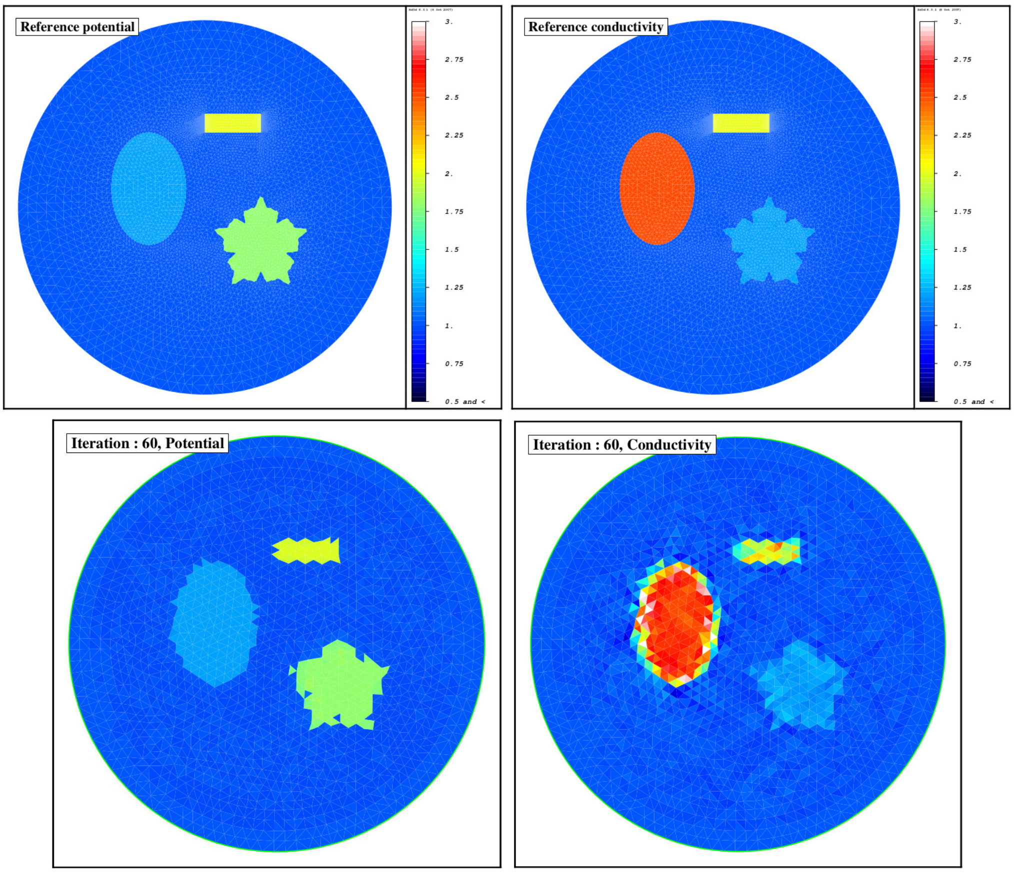


FIGURE 2.4 – Reconstruction du potentiel q et de la conductivité σ en deux dimension pour l'équation d'Helmoltz, [Amm+11] par un algorithme de gradient.

matrice dont les colonnes sont les vecteur $(\partial_x u_i, \partial_y u_i, u_i)$ soit strictement positif en tout point. Ensuite nous proposons un algorithme itératif de reconstruction basé sur une méthode de moindres carrés. Basé sur les idées de [Cap+09], cet algorithme est

$$\min_{\sigma,q} \mathcal{J}(\sigma) = \sum_i \int_\Omega j(\sigma|\nabla u_i|^2, d_{ii}) + j(q|u_i|^2, e_{ii}),$$

avec $j(a,b) = (\sqrt{a} - \sqrt{b})^2$. On montre numériquement que le coefficient q est beaucoup plus facile à retrouver que le coefficient σ. De manière générale, ce problème est beaucoup plus difficile que le 0-laplacien et les études numériques de stabilité au bruit et aux erreurs donnent des résultats bien inférieurs. La raison principale est que la fréquence k peut devenir une fréquence de résonnance (une valeur propre) de l'opérateur quand q et σ varient, ce qui rend l'algorithme d'optimisation beaucoup moins stable. Un exemple de reconstruction est donné en Figure 2.4.

2.5 [EG11] : Tomographie par Impédance électrique

Le problème de la tomographie par impédance électrique est de retrouver le coefficient de diffusion de l'équation de Laplace dans un domaine borné en ne connaissant que la réponse au bord (c'est-à-dire l'opérateur Dirichlet-to-Neumann ou DtN) de l'opérateur. Etant donné ϕ sur le bord d'un domaine Ω, si u résout l'équation de Laplace

$$-\,div(\sigma\nabla u) = 0 \text{ dans } \Omega \text{ et } u = \phi \text{ sur } \partial\Omega, \tag{2.4}$$

alors l'opérateur DtN est l'opérateur $\Lambda : \phi \mapsto \sigma\partial_n u|_\Omega$.

Un des résultats de base est que la connaissance de l'opérateur DtN permet de connaître σ si σ est supposée un peu régulière et surtout si σ est une fonction scalaire, [SU87 ; BT03 ; AP06 ; Buk08]. Si σ est autorisée à être une matrice, le contre-exemple de Tartar montre que pour tout changement de variable F égal à l'identité sur le bord, la matrice σ_F qui correspond à prendre l'image de σ par F donnera le même opérateur DtN.

Dans le cas où σ est une fonction scalaire, la technique usuelle pour montrer l'unicité ou la stabilité correspond à effectuer une transformation de Liouville et à transformer l'équation (2.4) en :

$$-\Delta v + qv = 0 \quad q = \frac{\Delta\sigma^{1/2}}{\sigma^{1/2}} \text{ et } v = \sigma^{1/2}u. \tag{2.5}$$

Ainsi le problème de la tomographie par impédance électrique est souvent posé comme le problème de retrouver un potentiel de Schrödinger q. La clef de l'unicité de q et des estimées de stabilité, est en dimension ≥ 3 de trouver des solutions v qui s'écrivent

$$v(x) = e^{\rho\cdot x}(1 + o(\frac{1}{|\rho|})) \text{ avec } \sum_i \rho_i^2 = 0 \text{ et } \rho \text{ complexe .}$$

De telles solutions sont appelées "solutions CGO" (complex-geometric optics) et peuvent être utilisées comme fonction test pour retrouver les coefficients de Fourier de q.

Nous nous sommes posé dans [EG11] la question de l'unicité de σ une fois que le problème est discrétisé. En d'autres termes, est-ce que la connaissance d'un opérateur Dirichlet-to-Neumann discret permet de retrouver le potentiel discret ? Le problème majeur est que les solutions CGO sont très oscillantes et on ne peut osciller plus finement que la taille d'une maille discrète. Nous avons donc adapté les preuves et la méthodologie du continu dans le cadre discret.

L'existence de fonctions CGO passent par des inégalités de Carleman que nous avons adaptées dans le cadre discret. Les inégalités de Carleman en continu peuvent se démontrer par intégration par partie en utilisant la règle de Leibniz de dérivée d'un produit. Dans le cadre discret, il faut retrouver une notion d'intégration par partie et de produit de fonction compatible avec une règle discrète de dérivation d'un produit. Le cadre naturel dans lequel ces deux notions se retrouvent est la méthode des différences finies. On se fixe un maillage de référence avec des directions unitaires fixées (le stencil du laplacien) $(e_i)_{i=1..n}$ qui ne dépend pas du noeud considéré. En notant d_i l'opérateur de dérivée dans une direction et a_i l'opérateur de moyennage,

$$d_i u(x) = \frac{u(x + \frac{h}{2}e_i) - u(x - \frac{h}{2}e_i)}{h} \text{ et } a_i u(x) = \frac{u(x + \frac{h}{2}e_i) + u(x - \frac{h}{2})}{2},$$

on s'intéresse donc à une discrétisation du type

$$-\sum_i d_i \sigma_i d_i u + qu = 0.$$

Le fait qu'on étudie bien le Laplacien est écrit dans deux paramètres de régularité ε_d et ε_a, supposés petits.

$$\varepsilon_d = \sum_{i,j} \|d_j(\sigma^i)\|_{l^\infty} \quad \varepsilon_a = \|\sum_i a^i(\sigma_i)e_i \otimes e_i - Id\|_{l^\infty}.$$

Le paramètre ε_d contrôle les variations de σ d'une maille à l'autre et le paramètre ε_a contrôle que la matrice σ est proche de l'identité. On définit ensuite le paramètre de régularité du maillage comme étant

$$\varepsilon = \max(\varepsilon_d, \sqrt{\varepsilon_a})$$

Le résultat principal est que si q_1 et q_2 sont deux potentiels supposés à priori bornés en norme L^∞ qui donnent lieu à deux opérateurs dirichlet-to-Neumann discrets Λ_1 et Λ_2, alors en notant

$$\mu = \max\left\{h^{1/2}, |\log(\|\Lambda_1 - \Lambda_2\|)|^{-1}, \varepsilon\right\},$$

on a

$$|q_1 - q_2|_{H^{-r}} \leq C_r \mu^{\frac{2r}{2r+d}}.$$

Ainsi on obtient une estimée de stabilité qui dépend de μ qui lui même est gouverné par la taille du maillage, la différence des opérateurs de Dirichlet to Neumann et de la régularité de la méthode des différences finies adoptée.

2.6 Perspectives

Il faut être honnête, je ne me suis pas intéressé aux problèmes inverses de cette partie récemment. Plus par question de goût que de manque de perspectives sur le sujet. Cependant, il est sorti un article [Lec+23] qui porte sur les égalités de Calderòn discrètes, dans la lignée de [EG11]. Dans cet article, les auteurs regardent le problème inverse de Calderòn avec des données partielles au bord. Ils ré-utilisent le formalisme que nous avons introduit avec Sylvain Ervedoza dans [EG11]. Cependant, ce formalisme se situe dans le cadre des différences finies et n'est pas adapté à la méthode des éléments finis, qui est l'état de l'art dans la résolution numérique de ce genre d'équations. Un point important serait de déployer les méthodes de [EG11] dans le cadre des éléments finis (ou même des volumes finis), c'est l'axe qui m'intéresserait sans doute le plus.

Bibliographie

[Amm+08] Habib Ammari, Eric Bonnetier, Yves Capdeboscq, Mickael Tanter et Mathias Fink. "Electrical impedance tomography by elastic deformation". In : *SIAM Journal on Applied Mathematics* 68.6 (2008), p. 1557-1573.

[Amm+11] Habib Ammari, Yves Capdeboscq, Frédéric de Gournay, Anna Rozanova-Pierrat et Faouzi Triki. "Microwave imaging by elastic deformation". In : *SIAM Journal on Applied Mathematics* 71.6 (2011), p. 2112-2130.

[AP06] Kari Astala et Lassi Päivärinta. "Calderón's inverse conductivity problem in the plane". In : *Annals of Mathematics* (2006), p. 265-299.

[Ber+09] Elena Beretta, Yves Capdeboscq, Frédéric de Gournay et Elisa Francini. "Thin cylindrical conductivity inclusions in a three-dimensional domain : a polarization tensor and unique determination from boundary data". In : *Inverse Problems* 25.6 (2009), p. 065004.

[BT03] Russell M Brown et Rodolfo H Torres. "Uniqueness in the inverse conductivity problem for conductivities with 3/2 derivatives in L p, p> 2n". In : *Journal of Fourier Analysis and Applications* 9.6 (2003), p. 563-574.

[Buk08] Alexander L Bukhgeim. "Recovering a potential from Cauchy data in the two-dimensional case". In : (2008).

[Cap+09] Y Capdeboscq, Jérôme Fehrenbach, Frédéric de Gournay et Otared Kavian. "Imaging by modification : numerical reconstruction of local conductivities from corresponding power density measurements". In : *SIAM Journal on Imaging Sciences* 2.4 (2009), p. 1003-1030.

[CV03] Yves Capdeboscq et Michael S Vogelius. "A general representation formula for boundary voltage perturbations caused by internal conductivity inhomogeneities of low volume fraction". In : *ESAIM : Mathematical Modelling and Numerical Analysis* 37.1 (2003), p. 159-173.

[Cin+10] Nicolae Cindea, Benoît Fabrèges, Frédéric de Gournay et Clair Poignard. "Optimal placement of electrodes in an electroporation process". In : *ESAIM : Proceedings*. T. 30. EDP Sciences. 2010, p. 34-43.

[EG11] Sylvain Ervedoza et Frédéric de Gournay. “Uniform stability estimates for the discrete Calderón problems”. In : *Inverse problems* 27.12 (2011), p. 125012.

[Lec+23] Rodrigo Lecaros, Jaime H Ortega, Ariel Pérez et Luz De Teresa. “Discrete Calderón Problem with Partial data”. In : *Inverse Problems* (2023).

[Mar+06] Michel Marty, Gregor Sersa, Jean Rémi Garbay, Julie Gehl, Christopher G Collins, Marko Snoj, Valérie Billard, Poul F Geertsen, John O Larkin, Damijan Miklavcic et al. “Electrochemotherapy–An easy, highly effective and safe treatment of cutaneous and subcutaneous metastases : Results of ESOPE (European Standard Operating Procedures of Electrochemotherapy) study”. In : *European Journal of Cancer Supplements* 4.11 (2006), p. 3-13.

[SU87] John Sylvester et Gunther Uhlmann. “A global uniqueness theorem for an inverse boundary value problem”. In : *Annals of mathematics* (1987), p. 153-169.

CHAPITRE 3

Le problème de Graetz

3.1 Un peu d'histoire

L'histoire du problème de Graetz commence avec la fin du dix neuvième siècle par le travail [Gra85] du physicien allemand Leo Graetz. Dans ce travail, Graetz considère un fluide caloporteur dans un cylindre axi-symétrique et étudie la convection-diffusion de la température. Les hypothèses simplificatrices sont que le régime est stationnaire et que la convection domine, c'est-à-dire que la diffusion le long du tube est négligée. Ce problème a ensuite été étendu dans [MV74 ; EZ89 ; WKB01 ; LOA02] en prenant en compte la diffusion longitudinale et est nommé le "problème de Graetz étendu". Le cadre des configurations axi-symétriques a été étudié dans [PRL80b ; PRL80a ; PRL81a ; PRL81b], ces modèles prennent en compte des configurations de plus en plus complexes, quoique toujours axisymétriques. Le premier travail sur les configurations non axisymétriques débute dans [PP09] et c'est dans ce cadre que nous travaillons. Considérons pour commencer l'équation de convection-diffusion d'une température T avec une vitesse de convection v :

$$\partial_t T - \operatorname{div}(K\nabla T) + v \cdot \nabla T = 0 \text{ dans } \Omega \times I$$

Où $\Omega \times I$ est un cylindre de section $\Omega \subset \mathbb{R}^2$ et de hauteur $I \subset R$. Le cas $I = \mathbb{R}^+$ est dit "semi-infini". Le terme de convection v vient d'une convection fluide dans un domaine ω. Ainsi le domaine ω est appelé la "phase liquide" et, par opposition, le domaine $\Omega \setminus \omega$ est nommée la phase "solide". Dans $\omega \times \mathbb{R}$, on modélise le fluide comme s'écoulant de manière laminaire et d'amplitude invariante par translation (Loi de Poiseuille) en se donnant l'équation de sa vitesse :

$$v(x, y, z) = h(x, y)e_z,$$

Où h vérifie l'équation

$$\begin{cases} -\Delta h = \alpha \text{ dans } \omega \\ h = 0 \text{ sur } \partial\omega, \end{cases} \tag{3.1}$$

avec $\alpha \in \mathbb{R}$. En séparant la variable z des variables (x, y), on obtient l'équation

$$c\partial_{zz}T + \operatorname{div}(\sigma\nabla T) - h\partial_z T = 0 \text{ on } \Omega \times I, \tag{E}$$

où les coefficients de diffusions $c, \sigma > 0$ sont bornés dans Ω avec une inverse bornée. Les conditions sur le bord de $\partial\Omega$ (LBC) peuvent être du type Neumann, Dirichlet, Robin ou périodiques, respectivement

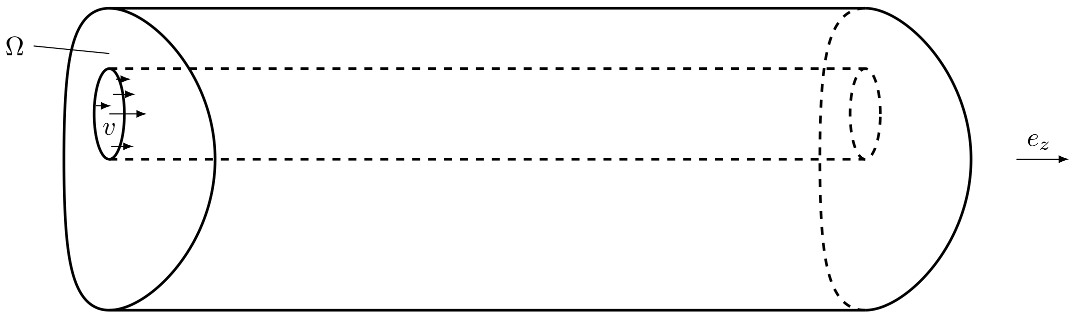


FIGURE 3.1 – Le domain $\Omega \times I$ sur lequel le problème est posé.

sur Γ_N,Γ_D,Γ_R,$\Gamma_\sharp \subset \partial\Omega$ et sont données par :

$$\begin{cases} \sigma\nabla T \cdot n = 0 \text{ sur } \Gamma_N \times I : \text{Neumann, et/ou} \\ T = 0 \text{ sur } \Gamma_D \times I : \text{Dirichlet, et/ou} \\ \sigma\nabla T \cdot n + aT = 0 \text{ sur } \Gamma_R \times I : \text{Robin, et/ou} \\ T \text{ est périodique sur } \Gamma_\sharp \times I : \text{périodique.} \end{cases} \tag{LBC}$$

Les conditions sur le bord de I sont nommées les conditions (I/OBC) (pour Input/Output), nous ne considérons que les conditions du type Dirichlet ou Neumann

$$T = T_D \text{ sur } \Omega_D \text{ et } \partial_z T = S_N \text{ sur } \Omega_N \text{ avec } \Omega_D \cup \Omega_N = \Omega \times \partial I. \tag{I/OBC}$$

En posant le problème de Graetz comme une équation d'évolution en z avec un opérateur de $\mathbb{R}^2$. En notant $\phi = \begin{pmatrix} u \\ s \end{pmatrix}$, cette équation devient :

$$\phi(z) = \begin{pmatrix} \partial_z T \\ T \end{pmatrix} \quad \partial_z \phi(z) = \mathcal{A}\phi(z) \text{ on } \Omega \times I, \text{ avec } \mathcal{A}\begin{pmatrix} u \\ s \end{pmatrix} = \begin{pmatrix} hc^{-1}u - c^{-1} \operatorname{div} \sigma\nabla s \\ u \end{pmatrix}. \tag{3.2}$$

Pour simplifier la présentation de ce document, nous allons adpoter un formalisme qui est arrivé tardivement dans notre étude du problème de Graetz et qui étend des résultats précédents. Effectivement, ce n'est que dans l'étude générale des conditions aux bords [Deb+18] que nous avons trouvé le bon formalisme pour poser le problème, ce qui suit est donc directement tiré de cet article.

Contrôle des contraintes et balancement L'étude de l'opérateur $\mathcal{A}$ dépend des conditions aux limites latérales, on distingue trois cas :

- La mesure de l'ensemble Γ_D ou de l'ensemble Γ_R est non nulle, dans ce cas les **constantes sont contrôlées.**
- Les constantes ne sont pas contrôlées et le débit est non nul, $\int_\Omega h \neq 0$ dans ce cas le problème de Graetz est dit **non-équilibré.**
- Les constantes ne sont pas contrôlées et le débit est nul $\int_\Omega h = 0$, dans ce cas le problème de Graetz est dit **équilibré**.

L'espace naturel sur lequel agit $\mathcal{A}$ est :

$$\mathcal{H} = \{(u, s), u \in L^2(\Omega), s \in H^1(\Omega), \text{ tel que } s = 0 \text{ sur } \Gamma_D \text{ et } s \text{ périodique sur } \Gamma_\sharp\}.$$

auquel on associe la forme bilinéaire

$$\big((u, s) \mid (u', s')\big)_{\mathcal{H}} = \int_\Omega cuu' + \sigma\nabla s \cdot \nabla s' + \int_{\Gamma_R} ass'.$$

On vérifie facilement que $\mathcal{A}$ est symétrique pour cette forme bilinéaire. Quand les constantes sont contrôlées, $\mathcal{H}$ est un espace de Hilbert mais quand les constantes ne sont pas contrôlées il faut quotienter $\mathcal{H}$ par le vecteur $(0,1)$, ce qui rend s défini à une constante près. Une fois que $\mathcal{H}$ est quotienté, on peut étudier le noyau de $\mathcal{A}$ et voir que dans le cas où le problème n'est pas équilibré, le noyau est réduit à $\{0\}$ et que si le problème est équilibré, en résolvant $-\operatorname{div}(\sigma\nabla s^\star) = h$, alors le noyau de $\mathcal{A}$ est l'espace vectoriel engendré par $(1, s^\star)$.

Diagonalisation de $\mathcal{A}$ On montre que $\mathcal{A}$ est un opérateur symétrique à résolvante compacte donc diagonalisable. On note dans la suite les couples de vecteurs propres de $\mathcal{A}$ comme $(\lambda_i, \phi_i)_{i\in M}$, avec $M = \mathbb{Z}$ si $\mathcal{A}$ a un noyau (problème équilibré) et $M = \mathbb{Z}^\star$ sinon.

$$\lambda_{-1} < 0 < \lambda_1 \quad \text{ et } \lambda_i \leq \lambda_j \text{ si } i \leq j.$$

Dans le cas équilibré $\lambda_0 = 0$ et $\phi_0 = (s^\star, 1)$.

Résolution du problème On pose P l'opérateur de projection de $\mathcal{H}$ dans L^2 qui consiste à garder la première composante de $\phi = (u, s)$, ainsi $P(u,s) = u$. Si T résout les équations (E) et (I/OBC), alors il est de la forme :

$$\begin{aligned}
\text{constantes contrôlées : } T(z) &= \sum_{i\in\mathbb{Z}^\star} a_i e^{\lambda_i z} P\phi_i \\
\text{cas non-équilibré : } T(z) &= \sum_{i\in\mathbb{Z}^\star} a_i e^{\lambda_i z} P\phi_i + c \\
\text{cas équilibré : } T(z) &= \sum_{i\in\mathbb{Z}^\star} a_i e^{\lambda_i z} P\phi_i + c + d(z + s^\star), \quad -\operatorname{div}\sigma\nabla s^\star = h,
\end{aligned}$$

où $a_i, c, d \in \mathbb{R}$ pour tout i.

3.2 [Bou+11],[Feh+12] : Le cas Dirichlet semi-infini

Dans [Feh+12], nous nous sommes posés la question de savoir si le problème de Graetz admettait une solution dans le cas où le tube était semi-infini ($I = \mathbb{R}^+$) avec des conditions latérales de Dirichlet (ainsi les constantes sont contrôlées) et s'il existait une méthode adéquate pour calculer les modes de décomposition selon le mode aval. Mathématiquement la question se pose de savoir si étant donné T_0, il est possible de trouver des coefficient a_i tels qu'il existe une décomposition de la forme

$$T(z) = \sum_{i<0} a_i T_i e^{\lambda_i z} \text{ si } T(0) = T_0 \text{ est donné.}$$

Il est important de noter que nous nous limitons ici aux indices i positifs donc aux valeurs propres $\lambda_i < 0$. En d'autre terme, est-il possible étant donné T_0 de trouver $T_1 = \partial_z T(0)$ tel que le vecteur $\phi = (T_1, T_0)$ vérifie $(\phi, \phi_i)_H = 0$ pour tout $i \leq 0$? La réponse de l'existence ne pose *a priori* pas trop de problèmes, effectivement, en supposant h assez régulière l'équation (E) n'est qu'une pertubation compate d'une équation de diffusion $3d$ et l'opérateur $\mathcal{A}$ qu'une perturbation compacte d'un opérateur de type Laplacien. Comme il est facile de trouver le ϕ pour un problème de diffusion pure ($h = 0$) on peut penser que dans le cas $h \neq 0$, on peut résoudre facilement le problème. Cependant la question du calcul numérique est plus compliquée. Effectivement, le problème est un problème spectral d'ordre 2, la donnée de $T_0 = T$ et $T_1 = \partial_z T$ en 0 est nécessaire pour calculer les a_i de la décomposition du vecteur en 0. Le problème semi-infini est donc un problème aux bords (avec $T = 0$ en $+\infty$) et calculer les coefficients a_i revient à trouver T_1 en 0 ce qui revient à résoudre un problème elliptique avec des conditions de type Cauchy, ce qui est su comme étant instable.

La méthode de reconstruction que nous avons proposé consiste tout d'abord à modifier le problème et à chercher $\tilde{T}$ pour que $\Phi = (T_0, \tilde{T})$ vérifie $(\phi, \phi_i)_H = 0$ pour tout $i \leq 0$. Il suffit d'intégrer en z la solution $T(z)$ correspondante pour répondre au problème initial. Ensuite on considère les opérateurs de projection P sur la première variable et π_- sur les modes négatifs définis par

$$P(u,s) = u \quad \pi_- y = \sum_{i\in\mathbb{N}} (y,\phi_i)_H \phi_i$$

En notant $\phi_D = (T_0, 0)$, on cherche Φ tel que $P\Phi = P\phi_D$ et $\phi_-\Phi = \Phi$. Ce qui donne la condition nécessaire

$$\pi_- P^\star P \pi_- \Phi = \pi_- P^\star t$$

Nous montrons ensuite que $\pi_- P^\star P \pi_- = 0$ est bijectif continu, donc que l'opérateur $B_- = (\pi_- P^\star P \pi_-)^{-1}\pi_-$ existe et on obtient comme condition nécessaire $\Phi = B_- P^\star \phi_D$, on vérifie ensuite que $PB_-P^\star = P$. La structure de la preuve réside dans le fait que P est une projection orthogonale et dans l'utilisation des égalités suivantes :

$$(Id - P)\mathcal{A}^{-1}(Id - P) = 0 \quad P\mathcal{A}P = 0 \quad (\mathcal{A}P\phi, P\phi) = \int_\Omega hu^2 \text{ si } \phi = (u,s)$$

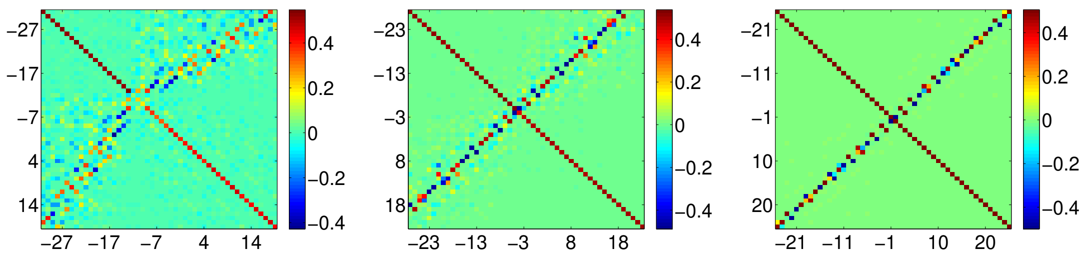


FIGURE 3.2 – La matrice $\pi P^\star P\pi$ pour π pour différentes valeurs du Péclet. Le Péclet est un indice de la norme de la vitesse. De droite à gauche, le Péclet vaut respectivement 10,1,0.1. On calcule les 50 plus petites valeurs propres en valeur absolue, de droite à gauche, il y a respectivement 31,27 et 25 valeurs propres négatifs.

Nous avons aussi étudié l'erreur commise quand tous les modes ne sont pas connus, pour cela, nous considérons la projection π_-^n sur les n premiers modes négatifs, nous montrons que l'opérateur $\pi_-^n P^\star P \pi_-^n$ est bijectif sur l'image de π_-^n et nous considérons l'opérateur restreint

$$B_-^n = (\pi_-^n P^\star P \pi_-^n)^{-1}\pi_-^n.$$

En notant $\Phi = B_- P^\star t$ la vraie solution recherchée et $\Phi^n = B_-^n P^\star t$, la solution calculée en ne prenant en compte que les n premiers vecteurs propres, alors on obtient l'estimation

$$\|\Phi - \Phi^n\| \leq C\|t - \pi_-^n t\|_{\mathcal{H}},$$

Cette estimation assure qu'il est effectivement possible de calculer Φ en se concentrant sur les premiers vecteurs propres négatifs. Il va de soi que cette estimation ne peut être améliorée dans la mesure où le calcul de B_-^n comporte une étape de projection sur π_-^n de t.

Nous donnons aussi dans [Feh+12] des calculs effectifs, on regarde notamment dans la Figure 3.2, la matrice $\pi P^\star P \pi$ pour π la matrice de projection sur les premiers modes calculés (négatifs et positifs), cette matrice est, dans le cas $h = 0$ la matrice $\frac{1}{2}\begin{pmatrix} Id & -Id \\ -Id & Id \end{pmatrix}$. Notre algorithme consiste à n'inverser cette matrice que sur sa partie à mode négatifs.

Dans [Bou+11], nous capitalisons sur ces algorithmes de calcul pour faire l'analyse paramétrique d'échangeurs thermiques quand la phase liquide se compose de deux sous-domaines, l'un avec une pression positive et l'autre avec une pression négative, voir Figure 3.3. Le modèle est dit de convection à contre-courant. L'analyse paramétrique est faite en faisant varier le rayon des cercles ou en faisant varier la pression des tubes.

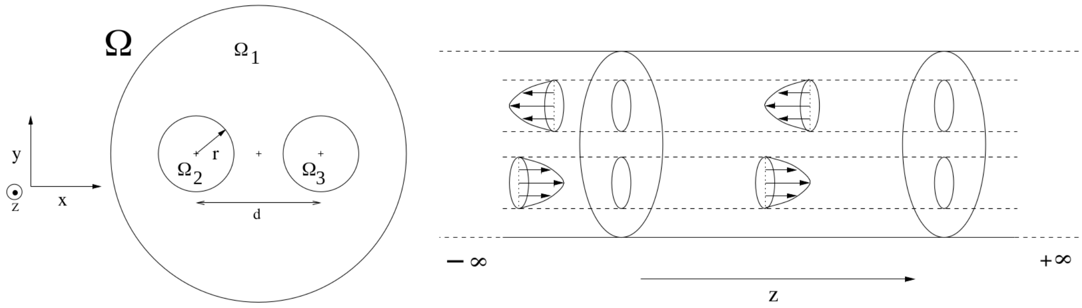


FIGURE 3.3 – Domaines de références pour l'analyse paramétrique faite dans [Bou+11].

3.3 [Deb+18] : D'autres conditions au bord

Dans [Deb+18], nous nous sommes intéressés à généraliser le travail fait dans [Feh+12] à d'autres conditions latérales aux bord que les conditions de Dirichlet. Comme dans [Feh+12], nous nous intéressons, dans le cas semi-infini, à décomposer une température selon ses modes avals. Nous rappelons que nous avons distingué plusieurs cas : tout d'abord s'il existe des conditions latérales de type Dirichlet ou Robin, on dit que les constantes sont contrôlés. Si elle ne sont pas contrôlées, il faut regarder $Q = \int_\Omega h$, le débit total de la vitesse fluide. Si ce débit est nul on dit que le problème est "équilibré", les résultats dépendront aussi du fait que Q soit > 0 ou < 0.

Cas semi-fini En utilisant le formalisme de la Section 3.2, étant donné T_0, nous construisons $\phi_D = (T_0, 0)$. En notant π_+ la projection sur les modes strictement positifs de $\mathcal{A}$, nous nous intéressons à trouver Φ tel que

$$\pi_+\Phi = 0 \text{ et } P\Phi = P\Phi_D \tag{3.3}$$

On introduit l'opérateur $B_=(\pi_- P^\star P \pi_-)^{-1}\pi_-$. On montre que cet opérateur est toujours inversible d'inverse continue. On note

On obtient

Proposition 3.1 Si $\phi = (1, 0)$ et si ϕ_0 le vecteur propre associé à la valeur propre 0 de $\mathcal{A}$. Le vecteur ϕ_0 n'existe que quand les contraintes ne sont pas contrôlées et que $Q = 0$ (le cas est équilibré).

- Si les constantes ne sont pas contrôlées et $Q > 0$, alors il existe une solution Φ à (3.3) si et seulement si
$$(\phi - PB_-\phi|\phi_D) = 0.$$
Cette solution est unique et est donnée par $\Phi = B_- P\phi_D$
- Si les constantes ne sont pas contrôlées et $Q = 0$, alors il existe une unique solution au problème (3.3) donnée par
$$\Phi = B_- P\phi_D + a(B_-\phi - \phi_D) \text{ avec } a = \frac{(\phi - PB_-\phi|\phi_D)_\mathcal{H}}{(\phi - PB_-\phi|\phi)_\mathcal{H}}$$
- Si les constantes sont contrôlées ou si $Q < 0$, il existe une unique solution $\Phi = B_- P\phi_D$ au problème.

Quand on transforme ce résultat en résultat sur T, et en n'oubliant pas que l'on a quotienté par les constantes, on obtient que si T_0 est fixé, il existe une unique solution de l'équation qui ait une croissance sous-exponentielle quand z tend vers l'infini et elle s'écrit de la forme

$$T_0 = \sum_{i<0} a_i e^{\lambda_i z} P\phi_i + r(z)$$

Où r dépend du cas considéré et caractérise le comportement de la solution à l'infini.

- Si les constantes sont contrôlées $r = 0$ et les a_i sont donnés par $a_i = (B_-\phi_D|\phi_i)_{\mathcal{H}}$
- Si les constantes ne sont pas contrôlées et $Q \neq 0$, alors $r = T_\infty$ est une constante. Si $Q < 0$, T_∞ est une inconnue du problème et doit être donnée. Si $Q > 0$, alors T_∞ est donné par

$$T_\infty = \frac{(\phi - PB_-\phi|\phi_D)_{\mathcal{H}}}{(\phi - PB_-\phi|\phi)_{\mathcal{H}}}.$$

  Les coefficients a_i sont donnés par $a_i = (B_-(\phi_D - T_\infty\phi)|\phi_i)_{\mathcal{H}}$
- Si les constantes ne sont pas contrôlées et $Q = 0$, alors $r(z) = T_\infty + c_2(z + P\phi_0)$, la constante c_2 est une inconnue du problème et quand c_2 est fixée, alors

$$T_\infty = \frac{(\phi - PB_-\phi|\phi_D - c_2\phi_0)_{\mathcal{H}}}{(\phi - PB_-\phi|\phi)_{\mathcal{H}}}$$

  Les coefficients a_i sont donnés par $a_i = (B_-(\phi_D - T_\infty\phi - c_2\phi_0)|\phi_i)_{\mathcal{H}}$. Dans la pratique, la constante c_2 est fixée à 0 pour empêcher la température d'avoir une croissance linéaire à l'infini.

Ce résultat confirme l'intuition physique du problème qui consiste à dire que la température à l'infinie est nulle dans le cas où il existe des conditions de Robin et/ou Dirichlet qui entraîne de la dissipation de température. Dans les autres cas, la température à l'infini doit être fixée quand le débit (qui est la vitesse totale) est < 0, c'est à dire qu'il existe en moyenne de l'advection de fluide qui provient de $+\infty$. Pour le cas équilibré, il faut choisir si on autorise une rampe linéaire de la température à l'infini. Dans le cas de Robin, on effectue une étude du paramètre de Robin qui permet de passer continuement d'un cas non contrôlé (Neumann) au cas contrôlé de Dirichlet, voir Figure 3.4.

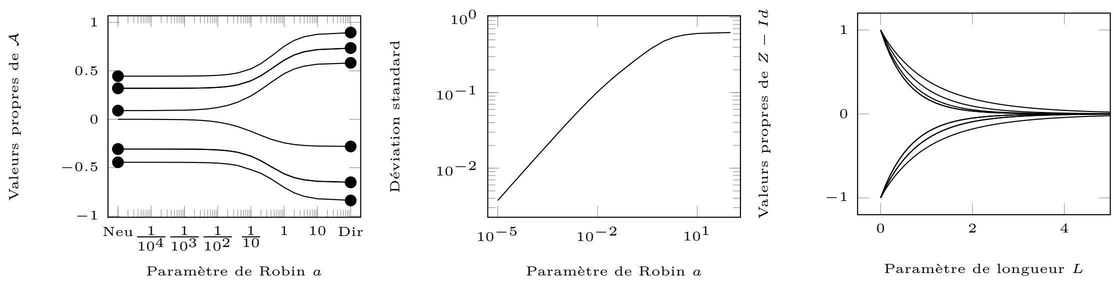


FIGURE 3.4 – (Gauche) : Évolution des valeurs propres de $\mathcal{A}$ de plus petite amplitude en fonction du paramètre de Robin. Les valeurs propres de Dirichlet et de Neumann sont notées avec des points. (Milieu) : Évolution de la norme de la différence entre le vecteur propre de $\mathcal{A}$ de plus grande valeur propre négative de et sa moyenne en fonction du paramètre de Robin. (Droite) : Dans le cas d'un débit > 0 et d'un paramètre de Robin fixé, évolution des valeurs propres de $Z - Id$ (3.4) en fonction de la longueur L de l'échangeur.

Cas fini On s'intéresse aussi à la résolution du cas fini. Nous ne discuterons ici que du cas où les constantes sont contrôlées, les autres cas sont gérés dans [Deb+18] mais ne feraient qu'alourdir la discussion. On suppose donc que la température T est fixée en $\pm L$, on note $T(\pm L) = T_\pm$. Pour résoudre le problème, on utilise la linéarité pour mettre $T_+ = 0$, on résout le tube semi-infini en supposant que $T(-L) = T_-$, on récupère l'erreur qui est faite en $T(L)$ par rapport à la valeur voulue qui est 0. On re-propage l'erreur en $z = L$ par un tube semi-infini dans le sens des $z < 0$ et on récupère encore une erreur en $T(-L)$, que l'on retro-propagera dans le sens des $z > 0$, etc... Comme les évolutions dans le tube semi-infini de la température sont exponentiellement décroissantes, on a bon espoir qu'un tel schéma fonctionne. Cela nous amène à considérer les opérateurs $M_\pm = Pe^{\mp 2L\mathcal{A}}B_\pm$. On reconnaît $B_\pm$ les opérateurs de résolution des cas semi-infinis (dans le sens des $z > 0$ et $z < 0$) et on reconnaît $e^{\mp 2L\mathcal{A}}$ qui est un opérateur d'évolution de la solution le long du tube fini. Il existe une solution au problème fini si et seulement si l'opérateur Z défini par

$$Z = \begin{pmatrix} Id & M_- \\ M_+ & Id \end{pmatrix}, \tag{3.4}$$

est inversible. Dans [Deb+18], nous montrons que cet opérateur est bien inversible pour les petites et les grandes valeurs de $L > 0$ mais on ne sait rien faire pour les valeurs intermédiaires de L. Voir Figure 3.4 (droite) pour une représentation des valeurs propres de Z. Néanmoins nous donnons les formules de résolution dans le cas fini, j'en épargne la lecture ici, ce sont des problèmes linéaires à faible nombre de variables (de l'ordre du nombre de vecteurs propres calculés).

3.4 [GFP14] : Optimisation de formes

Nous nous sommes naturellement portés sur l'optimisation de formes du problème de Graëtz dans [GFP14]. L'objectif est d'optimiser la première valeur propre positive de Graëtz par rapport à la forme de ω, la forme du tube dans lequel passe le fluide caloporteur. Nous commençons par noter toutes les variables qui dépendent de ω en mettant explicitement leur dépendence par rapport à ω.

- Les coefficients de diffusion de la température. Effectivement le fluide caloporteur diffuse la température différemment que la phase solide. Ainsi se fixe des coefficients de diffusion dans la phase solide et liquide et les coefficients c_ω et σ_ω sont définis par des formules $c_\omega = c_0 + c_1 \mathbb{1}_\omega, \sigma_\omega = \sigma_0 + \sigma_1 \mathbb{1}_\omega$.
- La vitesse d'advection de la température $h_\omega = \alpha_\omega v_\omega, \alpha_\omega \in \mathbb{R}$ où v_ω vérifie l'équation de Poiseuille :

$$-\Delta v_\omega = 1 \quad \text{sur } \omega \quad v = 0 \quad \text{sur } \partial\omega$$

  et h_ω et renormalisé par un coefficient α_ω qui peut lui aussi dépendre de ω, les trois types de renormalisation sont :

$$\text{"Flot" : } \int_\omega h_\omega = 1 \text{ ou "Travail" : } \int_\omega -\Delta h_\omega = 1 \text{ ou "Dissipation" : } \int_\omega |\nabla h_\omega|^2 = 1$$

- L'opérateur de Graetz $\mathcal{A}_\omega$ et le produit scalaire $\mathcal{B}_\omega$ sur $\mathcal{H}$ [1]

$$\mathcal{A}_\omega \begin{pmatrix} u \\ s \end{pmatrix} = \begin{pmatrix} h_\omega c_\omega^{-1} u - c_\omega^{-1} \operatorname{div} \sigma_\omega \nabla s \\ u \end{pmatrix}.$$

$$\left(B_\omega\left(u,s\right),\left(u',s'\right)\right) = \int_\Omega c_\omega uu' + \sigma_\omega \nabla s \cdot \nabla s' + \int_{\Gamma_R} ass'.$$

- Et finalement le premier mode positif λ_ω définit par un quotient de Rayleigh :

$$\lambda_\omega = \max_\Phi \frac{(B_\omega \Phi, \Phi)}{(\mathcal{A}_\omega \Phi, \Phi)}$$

Dans le cadre de l'analyse de sensibilité par rapport aux variations de domaines, chacune des dépendances est relativement bien comprise. Les dérivées de forme de la vitesse v_ω et donc du paramètre de renormalisation α_ω sont une application directe de techniques usuelles [AS07 ; HP06]. Les dérivées des opérateurs $\mathcal{A}_\omega$ et B_ω sont typiques des dérivées avec sauts dans les coefficients [Pan05 ; BP03] et les dérivées de valeur propres sont similaires aux résultats de [HR80 ; SLO94 ; Gou06]. Cependant mettre en musique les différentes techniques est ardu, nous démontrons dans [GFP14] la différentiabilité de la première valeur propre par rapport à ω. Il existe une difficulté supplémentaire dans l'établissement de la différentiablité, qui est qu'il n'y a pas de saut de

1. A l'époque nous n'avions pas le même formalisme que celui pris dans ce document qui est tiré de [Deb+18]. Donc les choix des opérateurs $\mathcal{A}$ et $\mathcal{B}$ sont un peu différents dans [GFP14].

dérivabilité entre l'opérateur $\mathcal{A}$ et $\mathcal{B}$ et que la compacité de la résolvante de $\mathcal{A}$ est "cachée". Même en supposant l'espace propre associé à λ_1 de dimension 1, l'existence de la dérivée doit être faite à la main en montrant la convergence des vecteurs propres. Dans [GFP14], nous montrons l'existence de dérivées directionnelles de λ_1 quand l'espace propre est de dimension arbitraire. Je n'écrirais pas la formule de la dérivée de forme ici, qui est trop complexe à mon goût, nous noterons qu'elle est quadratique par rapport aux vecteurs propres de E_1, l'espace associé à la première valeur propre de $\mathcal{A}$ et qu'elle s'écrit

$$d\lambda_1(\theta) = \min_{\phi \in \mathcal{E}_1, (B\phi, \phi)=1} \int_{\partial\omega} (\theta \cdot n)\lambda_1 \nabla h \nabla p + e(\phi),$$

où e est une forme quadratique explicite, où h est la vitesse d'advection et p est un état adjoint nécessaire pour relever la dépendance de $\mathcal{A}$ par rapport à h. Cet état adjoint p est défini pour tout $\phi = (\lambda_1 T, T) \in E_1$ par

$$\int_\omega \nabla p \nabla \psi = \int_\omega T^2 \psi \quad \forall \psi$$

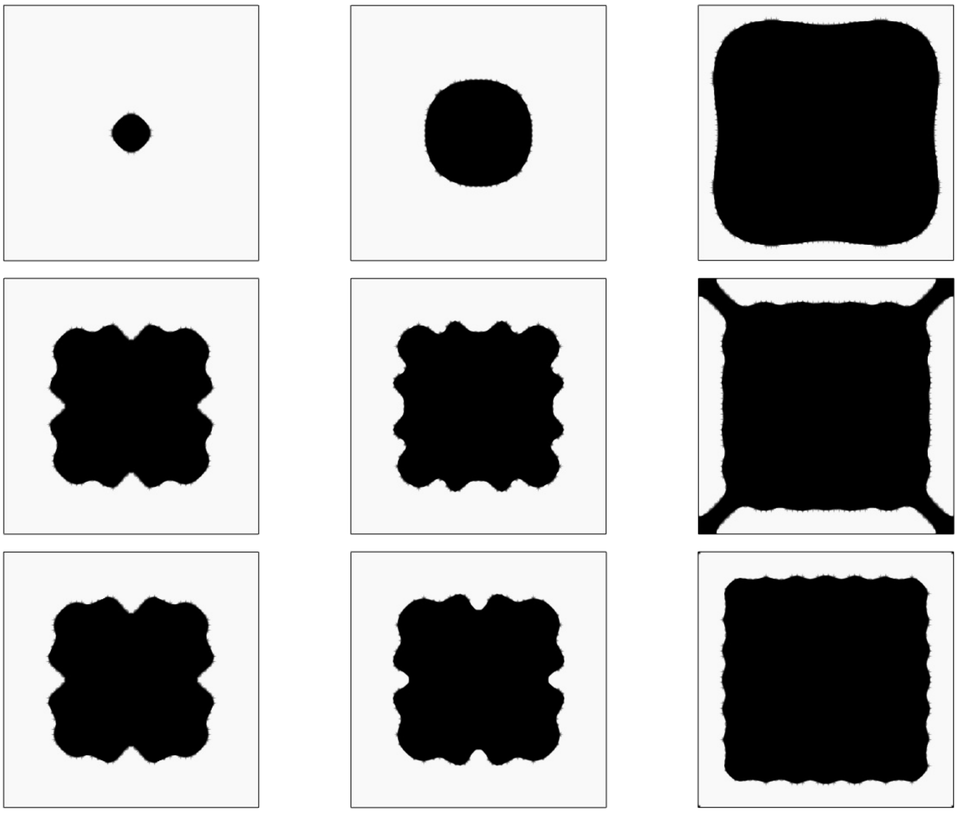


FIGURE 3.5 – Les formes optimales pour la méthode de level-set et le flot fixé. La conductivité de la phase solide évolue de haut en bas et prend les valeurs $1, 5, 10$. De gauche à droite, le flot total prend les valeurs : $F = 0.1, 1, 10$

L'état adjoint est bien quadratique par rapport aux vecteurs propres, étant donné une base de vecteurs propres de E_1 de taille n, il est nécessaire de calculer $\frac{n(n+1)}{2}$ états ajoints pour calculer la forme quadratique $\phi \mapsto \lambda_1 \nabla h \nabla p + e(\phi)$. Des techniques d'optimisation développées dans [Gou06] peuvent être ensuites appliquées. Nous donnons des résultats numériques d'optimisation par la méthode de level-set.

Nous nous sommes ensuites aperçus que les formes optimales présentent une structure non lisse sur le bord, ceci sans doute afin de maximiser l'échange entre la partie fluide ω et la partie

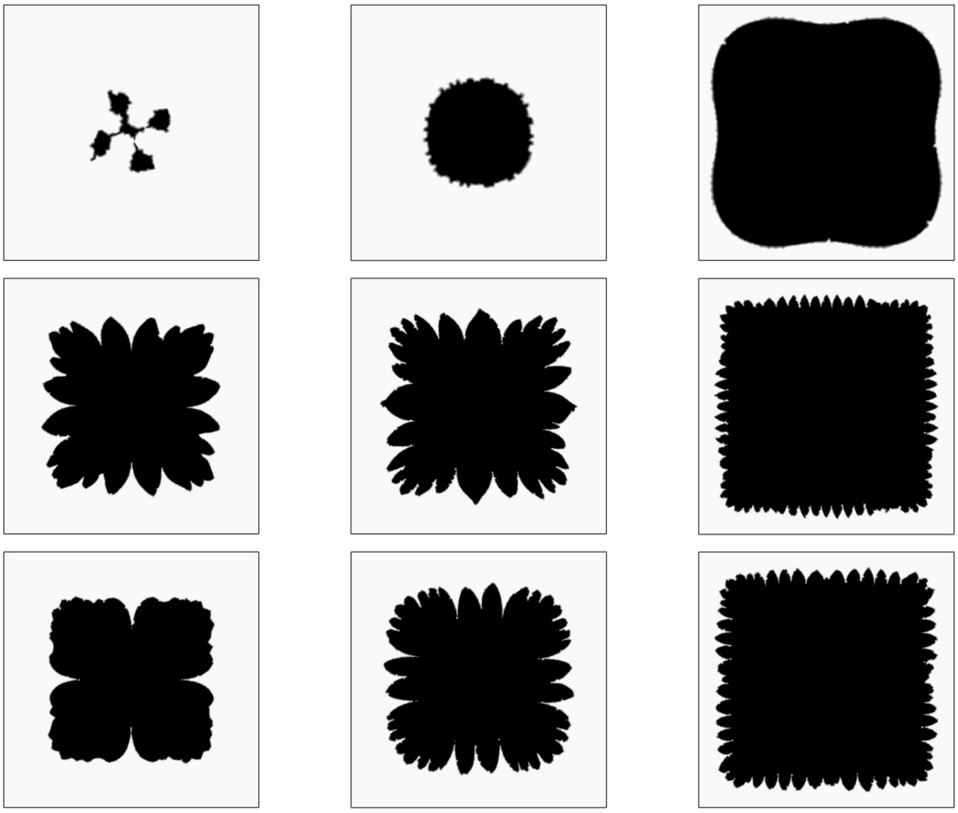


FIGURE 3.6 – Les formes optimales pour la méthode d'advection de points de la frontière et le flot fixé. La conductivité de la phase solide évolue de haut en bas et prend les valeurs $1, 5, 10$. De gauche à droite, le flot total prend les valeurs : $F = 0.1, 1, 10$

solide $\Omega \setminus \omega$. Ainsi nous avons aussi implémenté des algorithmes de remaillage de forme de la frontière pour capter ce comportement. Ces algorithmes de remaillage prennent comme paramètre les points de la frontière de $\partial\omega$, les advectent et remaillent depuis ces nouveaux points l'intégralité du domaine Ω. Ces algorithmes sont beaucoup plus sensibles que la méthode level-set car ils ont tendance à replier la frontière sur elle-même, rendant le domaine topologiquement impossible à mailler. Cependant, correctement initialisé par une méthode de courbe de niveaux, ils donnent d'excellent résultats, si on se concentre sur de la minimisation locale. Nous avons aussi implémenté des régularisations par pénalisation du périmètre, les résultats n'étant pas si probants que cela, nous n'avons pas continué à chercher dans cette voie.

3.5 [Pie+14] : Des échangeurs

Dans [Pie+14], nous nous sommes interessés à recoller entre eux des problèmes de Graëtz pour arriver à simuler des échangeurs. La forme classique d'un échangeur est donnée en Figure 3.7, il consiste ici en deux tubes infinis dans la direction e_z de section Ω_1 et Ω_2. Ces deux tubes échangent de la température sur la partie $z \in [0, L]$. Sont imposées les températures T_i à l'infini aux seuls endroits où la vitesse "rentre" dans le tube, l'objectif est de calculer les températures de sorties T_o.

Nous nous sommes posés la question dans [Pie+14] de calculer numériquement ces températures de sortie. Pour cela nous calculons 3 problèmes aux valeurs propre de Graëtz (les deux tubes infinis et la section d'échange) et nous découpons l'espace en 5 tubes, deux semi-infinis situés en $z < 0$, deux autres en $z > L$ et l'échangeur à proprement parler en $z \in [0, L]$. Dans ces 5 tubes, la

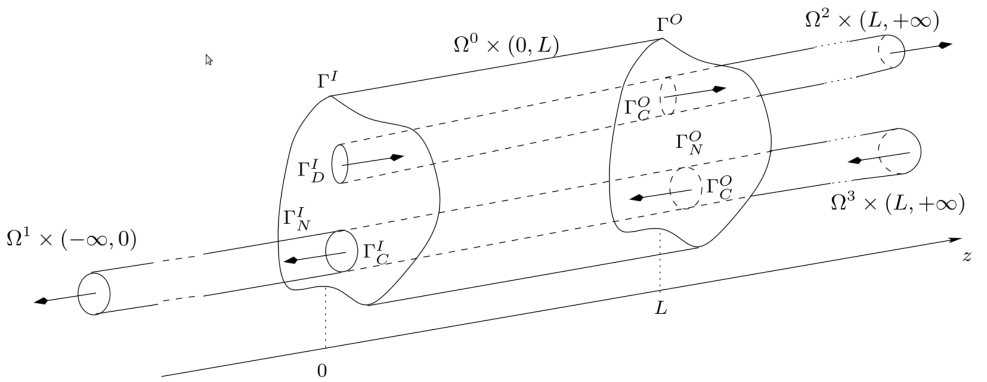


FIGURE 3.7 – Exemple typique d'échangeur avec 5 domaines et 3 problèmes de Graetz.

température est définie via une des formules :

$$
\begin{aligned}
\text{constantes controllées : } T &= \sum_{i \in \mathbb{Z}^\star} a_i e^{\lambda_i z} P \phi_i \\
\text{cas non-équilibré : } T &= \sum_{i \in \mathbb{Z}^\star} a_i e^{\lambda_i z} P \phi_i + c \\
\text{cas équilibré : } T &= \sum_{i \in \mathbb{Z}^\star} a_i e^{\lambda_i z} P \phi_i + c + d(z + s^\star), \quad -\operatorname{div} \sigma \nabla s^\star = h,
\end{aligned}
$$

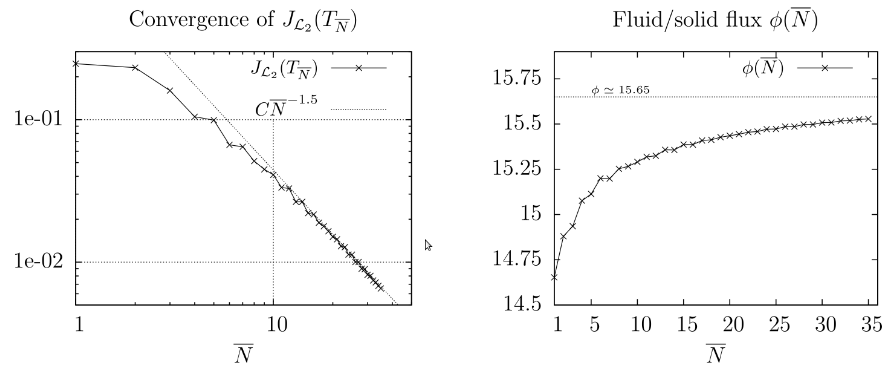


FIGURE 3.8 – Convergence de l'erreur d'approximation sur les interfaces (gauche) vers 0 et convergence du flux de température sur l'interface solide/fluide (droite) en fonction du nombre N_{modes} de valeurs propres en échelle bi-logarithmique. Sur la figure de droite, la valeur théorique pour $N \to +\infty$

où les a_i, c, d sont les inconnues dans chaque tube et sont différent dans chaque tube et les (λ_i, ϕ_i) sont calculés comme des valeurs propres des problèmes de Graetz dont dépendent les tubes. Pour trouver ces inconnues nous résolvons ensuite un problème de minimisation de moindre carré qui consiste à minimiser, en $z = 0$ et $z = L$, le saut de la température et de sa dérivée normale en norme L^2. Après un travail très technique en programmation et en notations, on se retrouve à construire une grosse matrice que l'on doit inverser. Nous avons vérifié notre algorithme en comparant dans des cas précis les valeurs calculées à des solutions analytiques connues [PP20], voir Figure 3.8.

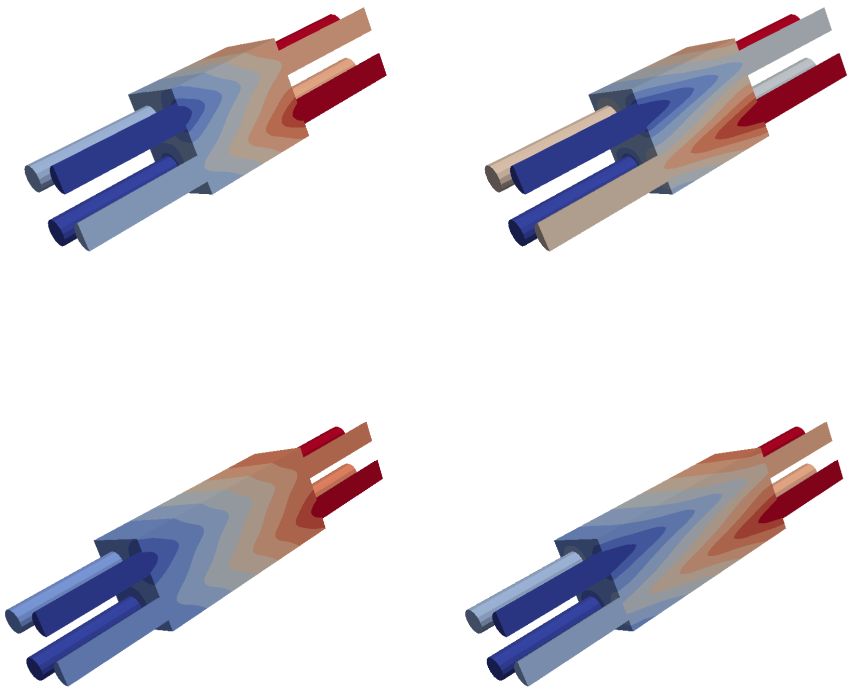


FIGURE 3.9 – Quatres différentes solutions d'échangeur à condition latérales périodique en faisant varier la longueur de l'échangeur ou le débit de la phase fluide. La taille de l'échangeur est 2 fois plus grande en bas qu'en haut. Le débit du fluide est 3 fois plus grand à droite qu'à gauche. Ces résultats sont issus de [Deb+18].

3.6 [DGP17],[DDP19] : Des applications

Nous avons utilisé le code développé dans [Pie+14] ainsi que la gestion des conditions limites quelconques faites dans [Deb+18] pour faire plusieurs analyse paramétrique des échangeurs. La première publication [DGP17] est à destination de la communauté thermique. Dans cette analyse, une structure type a été implémentée, elle consiste en des échangeurs à 5 tubes et 3 problèmes de Graetz à section circulaire. Les conditions latérales sont de type Robin.

Différents paramètres physiques ont été modifiés pour faire une analyse paramétrique : le Péclet qui est un indice de la vélocité du fluide, le nombre de Biot qui caractérise les conditions de Robin latérales, le ratio des conductivités de température entre les phases solides et liquides et finalement le rayon des tubes dans lesquels le fluide évolue. Le paramètre observé est l'efficacité thermique qui est défini comme

$$\epsilon = \frac{T_H^{-\infty} - T_H^{+\infty}}{T_H^{-\infty} - T_C^{+\infty}}$$

Où T_H et T_C représentent la température dans le tube chaud ou froid respectivement et l'indice $\pm\infty$ représente le fait qu'on s'intéresse à la température quand $z \to \pm\infty$. Notons que dans ce cas là, $T_H^{-\infty}, T_C^{-\infty}$ sont des données du problème et seule $T_H^{+\infty}$ est une inconnue.

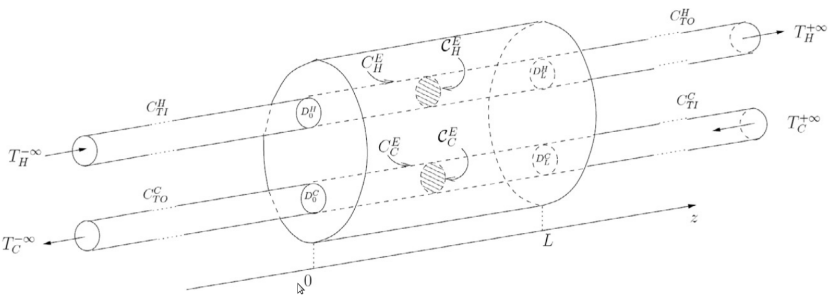


FIGURE 3.10 – Configuration de l'échangeur dans les tests numériques de [DGP17]

Dans [DDP19], nous avons aussi étudier des échangeurs avec des conditions aux bords périodiques et avons appliqué notre méthode de calcul à une modélisation biologique. Nous voulons modéliser les échanges thermiques entre l'artère qui contient du sang chaud et les veines qui contiennent du sang plus froid [Des37] [2].

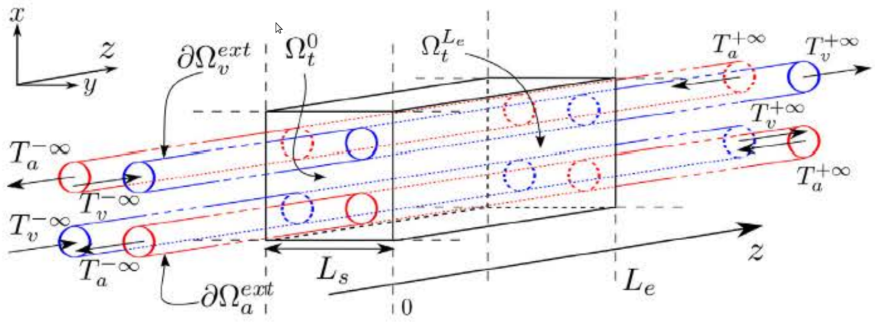


FIGURE 3.11 – Modélisation simple du réseau sanguin utilisé dans [DDP19]

Une analyse paramétrique est faite, nous nous sommes concentrés ici sur le diamètres des vaisseaux sanguins, sur leur densité (contrôlée par la taille de la boîte extérieure) et sur le ratio des pressions aortiques et veineuses. Ce sont les paramètres qui ont le plus d'influence sur la capacité d'échange thermique du réseau sanguin.

3.7 Perspectives

Les deux applications [DGP17] et [DDP19] montrent l'importance du problème de Graetz et de sa modélisation pour d'autres sciences, en thermique ou en biologie. Il reste des questions mathématiques en suspens, notamment concernant l'unicité de la solution dans les échangeurs. Une grosse difficulté est que le code de calcul a été construit de manière historique, en agglomérant les résultats et les demandes de plusieurs années. Il faudrait surtout prendre le temps de faire un nettoyage profond de ce code pour pouvoir l'utiliser de manière plus efficace.

2. Dans cette publication l'auteur estime même que la différence de température est à l'origine de la circulation sanguine ! ! Cela ne l'a cependant pas empêché de fonder les sciences modernes.

Bibliographie

[AS07] Grégoire Allaire et Marc Schoenauer. *Conception optimale de structures*. T. 58. Springer, 2007.

[BP03] Ch Bernardi et O Pironneau. “Sensitivity of Darcy’s law to discontinuities”. In : *Chinese Annals of Mathematics* 24.02 (2003), p. 205-214.

[Bou+11] Julien Bouyssier, Jérôme Fehrenbach, Frédéric de Gournay, Charles Pierre et Franck Plouraboué. “Analyse de la convection-diffusion entre deux tubes parallèles plongés dans un domaine cylindrique”. In : 20 ème congrès français de mécanique. 2011.

[Deb+18] Valentin Debarnot, Jérôme Fehrenbach, Frédéric de Gournay et Léo Martire. “The Case of Neumann, Robin, and Periodic Lateral Conditions for the Semi-infinite Generalized Graetz Problem and Applications”. In : *SIAM Journal on Applied Mathematics* 78.4 (2018), p. 2227-2251.

[Des37] René Descartes. *Discours de la méthode, Pour bien conduire sa raison, et chercher la vérité dans les sciences*. 1637.

[DDP19] Jules Dichamp, Frédéric De Gournay et Franck Plouraboué. “Thermal significance and optimal transfer in vessels bundles is influenced by vascular density”. In : *International Journal of Heat and Mass Transfer* 138 (2019), p. 1-10.

[DGP17] Jules Dichamp, Frédéric de Gournay et Franck Plouraboué. “Theoretical and numerical analysis of counter-flow parallel convective exchangers considering axial diffusion”. In : *International Journal of Heat and Mass Transfer* 107 (2017), p. 154-167.

[EZ89] MA Ebadian et HY Zhang. “An exact solution of extended Graetz problem with axial heat conduction”. In : *International Journal of Heat and Mass Transfer* 32.9 (1989), p. 1709-1717.

[Feh+12] Jérôme Fehrenbach, Frédéric de Gournay, Charles Pierre et Franck Plouraboué. “The generalized Graetz problem in finite domains”. In : *SIAM Journal on Applied Mathematics* 72.1 (2012), p. 99-123.

[Gou06] Frédéric de Gournay. “Velocity extension for the level-set method and multiple eigenvalues in shape optimization”. In : *SIAM journal on control and optimization* 45.1 (2006), p. 343-367.

[GFP14] Frédéric de Gournay, Jérôme Fehrenbach et Franck Plouraboué. “Shape optimization for the generalized Graetz problem”. In : *Structural and Multidisciplinary Optimization* 49.6 (2014), p. 993-1008.

[Gra85] von L Graetz. “Über die wärmeleitungsfähigkeit von flüssigkeiten”. In : *Annalen der Physik* 261.7 (1885), p. 337-357.

[HR80] Edward J Haug et Bernard Rousselet. “Design sensitivity analysis in structural mechanics. II. Eigenvalue variations”. In : *Journal of Structural Mechanics* 8.2 (1980), p. 161-186.

[HP06] Antoine Henrot et Michel Pierre. *Variation et optimisation de formes : une analyse géométrique*. T. 48. Springer Science & Business Media, 2006.

[LOA02] J. Lahjomri, A. Oubarra et A. Alemany. “Heat transfer by laminar Hartmann flow in thermal entrance region with a step change in wall temperatures : the Graetz problem extended”. In : *Int. J. Heat Mass Transfer* 45.5 (2002), p. 1127-1148.

[MV74] ML Michelsen et John Villadsen. “The Graetz problem with axial heat conduction”. In : *International Journal of Heat and Mass Transfer* 17.11 (1974), p. 1391-1402.

[Pan05] Olivier Pantz. “Sensibilité de l’équation de la chaleur aux sauts de conductivité”. In : *Comptes Rendus Mathematique* 341.5 (2005), p. 333-337.

[PRL81a] E. Papoutsakis, D. Ramkrishna et H-C. Lim. “Conjugated Graetz problems. Pt.1 : general formalism and a class of solid-fluid problems.” English. In : *Chemical Engineering Science* 36.8 (1981), p. 1381-1391.

[PRL81b] E. Papoutsakis, D. Ramkrishna et H-C. Lim. “Conjugated Graetz problems. Pt.2 : Fluid-Fluid problem”. English. In : *Chemical Engineering Science* 36.8 (1981), p. 1393-1399.

[PRL80a] E. Papoutsakis, D. Ramkrishna et H. C. Lim. “The extended Graetz problem with Diriclet wall boundary conditions”. In : *Appl. Sci. Res.* 36 (1980), p. 13-34.

[PRL80b] E. Papoutsakis, D. Ramkrishna et H. C. Lim. “The extended Graetz problem with prescribed wall flux”. In : *AIChE J.* 26 (1980), p. 779-787.

[PP09] C. Pierre et F. Plouraboué. “Numerical analysis of a new mixed-formulation for eigenvalue convection-diffusion problems”. In : *SIAM Appl. Math.* 70.3 (2009), p. 658-676.

[Pie+14] Charles Pierre, Julien Bouyssier, Frédéric de Gournay et Franck Plouraboué. “Numerical computation of 3D heat transfer in complex parallel heat exchangers using generalized Graetz modes”. In : *Journal of Computational Physics* 268 (2014), p. 84-105.

[PP20] Charles Pierre et Franck Plouraboué. “Generalised graetz problem : analytical solutions for concentric or parallel configurations”. In : *Meccanica* 55.8 (2020), p. 1545-1559.

[SLO94] Alexander P Seyranian, Erik Lund et Niels Olhoff. “Multiple eigenvalues in structural optimization problems”. In : *Structural optimization* 8.4 (1994), p. 207-227.

[WKB01] B Weigand, M Kanzamar et H Beer. “The extended Graetz problem with piecewise constant wall heat flux for pipe and channel flows”. In : *International journal of heat and mass transfer* 44.20 (2001), p. 3941-3952.

Chapitre 4

Parcimonie en optimisation

4.1 [Boy+19] :Théorème de représentation

Un théorème de représentation est une caractérisation de l'ensemble des minimiseurs d'une fonctionnelle. Considérons par exemple un algorithme de reconstruction. Etant donné E un espace de Banach, on se donne $\Phi : E \to \mathbb{R}^m$ un opérateur de mesure linéaire et on se donne un fonction coût $f : \mathbb{R}^m \to \mathbb{R}$. Par exemple pour une prédiction u, $f(\Phi u)$ peut représenter l'écart avec des mesures réellement observées. Ainsi la fonction f est dite une *fonction d'attache aux données.* On se donne aussi un régulariseur $\mathcal{R} : E \to \mathbb{R}$ et on s'intéresse à

$$\min_{u \in E} f(\Phi u) + \mathcal{R}(u) \tag{4.1}$$

Nous utiliserons un régulariseur qui est la jauge d'un ensemble convexe C.

$$\mathcal{R}(u) = \inf \{\lambda \text{ tel que } u \in \lambda C, \lambda \geq 0\}$$

Notons que les normes sont des jauges de leurs boule unité. Le régulariseur a deux avantages.

- Il permet de poser correctement le problème, car l'opérateur de mesure Φ a un noyau important si le nombre des mesures m est faible devant la dimension de E.
- De plus l'opérateur $\mathcal{R}$ peut favoriser certaines solutions [Cha+12] aux détriment d'autres. L'art de l'optimiseur est de trouver un régularisateur $\mathcal{R}$ qui va favoriser des solutions que l'on veut *a priori* obtenir.

Par exemple il est connu qu'un régularisateur issu de la norme L^1 favorisera des solutions parcimonieuses. Dans le problème de l'*inpainting* par exemple, on se donne une image de référence u_0 dont on cache certains pixels. Cette opération linéaire Φ correspond à mutliplier les pixels cachés par 0 et les pixels connus par 1 et on essaie de retrouver u_0. Pour cela on peut utiliser le formalisme de (4.1) en prenant $f(y) = (y - \Phi u_0)^2$ et prendre comme régulariseur

$$\mathcal{R}(u) = \|\nabla u\|_{\mathcal{M}},$$

Où la norme $\mathcal{M}$ est la masse totale d'une mesure si u est suposée BV (Bounded Variation), la norme $\mathcal{M}$ devient la norme L^1 classique dès que la fonction est plus régulière (la fonction u doit être $W^{1,1}$ dans notre exemple). Ce choix de régulariseur permet de privilégier les images constantes par morceaux (avec des aplats). Nous nous somme intéressés dans [Boy+19] à essayer de caractériser, pour un $\mathcal{R}$ donné, quels sont les solutions qui sont favorisées.

4.1.1 Résultats

Nous allons donner ici les deux résultats principaux de [Boy+19], mais il faut tout d'abord rappeler quelques définitions. Si C est un ensemble convexe linéairement fermé (toute intersection de C avec une droite est fermée pour la topologie de $\mathbb{R}$, ce qui est l'hypothèse de fermeture la plus faible qui soit compatible avec la notion de convexité), on définit C_L (le **linéal** de C), un sous-espace vectoriel de E et qui est défini comme l'ensemble des lignes contenues dans C. On peut ensuite quotienter C en $C = C_B + C_L$ où C_B est un ensemble convexe ne contenant aucune ligne. Ne contenant aucune ligne, on peut définir sur C_B les points extrémaux $\text{ext}(C_B)$ et les rayons extrémaux $\text{rext}(C_B)$. Nous appelons "atomes" des points extrémaux de C ou des points sur les rayons extrémaux de C.

La contribution principale de [Boy+19] est :

Proposition 4.1 On suppose qu'il existe v une solution de (4.1). On note δ qui vaut

$$\delta = 1 \text{ si } \mathcal{R}(v) > \inf_{u \in E} \mathcal{R}(u) \quad \delta = 0 \text{ sinon.}$$

et on note d la dimension de $\Phi(C_L)$ alors il existe une solution de (4.1), notée $u^\star$ qui s'écrit :

$$u^\star = u_L + \sum_{k=1}^{r} \alpha_k e_k,$$

Où $u_L \in C_L$, $\alpha_k \geq 0$ et les e_k sont les atomes de C_B, c'est-à-dire soit des points extrémaux de C_B soit dans des rayons extrémaux de C_B. On donne une borne sur r qui est

$$\begin{aligned} r &\leq m - d + 1 - \delta \text{ dans le cas général} \\ r &\leq m - d - \delta \text{ si un des } e_k \text{ est dans un rayon extrémal} \end{aligned} \tag{4.2}$$

Le paramètre δ peut être vu comme un multiplicateur de Lagrange qui s'intéresse à savoir si la contrainte de régularisation $\mathcal{R}$ est active au minimum, en d'autres termes si l'introduction du régularisateur a un effet sur le problème de minimisation original.

Démonstration

Notons que le cas $\delta = 0$ est simple à montrer car s'il existe une solution v, alors tout $v + \phi$ avec $\phi \in ker\Phi$ est solution et quitte à quotienter par C_L, c'est un espace vectoriel de codimension $m - d$. Un théorème de Carathéodory permet de conclure. La contribution de [Boy+19] s'est finalement avérée être une réduction du nombre de points extrémaux pour décrire un ensemble convexe dans le cas $\delta = 1$.

Dans le cas où f est convexe, alors l'ensemble des solutions $S^\star$ devient un ensemble convexe, en quotientant par C_L, on obtient $S^\star = S_B + S_L$ avec $S_L = \ker(\Phi) \cap C_L$ et nous avons un résultat plus fort qui consiste à pouvoir représenter tout $p \in S_B$ dont la face est de dimension j comme

$$p = u_L + \sum_{k=1}^{r+j} \alpha_k e_k,$$

avec les bornes sur r données par (4.2).

4.1.2 Applications

Nous généralisons avec ce théorème un certain nombre de résultats obtenus pour différents régularisateurs, il suffit de calculer pour un ensemble convexe donné son ensemble linéal et ses atomes. Par exemple, si le régularisateur $\mathcal{R}$ est la norme ℓ^1 des vecteurs, alors sa la boule unité n'a pas d'ensemble linéal et ses atomes sont les vecteurs de la base canonique. Ainsi, nous montrons qu'il existe une solution m-sparse, c'est la notion de base de l'échantillonnage compressé [CRT06]. Cette remarque s'applique aussi quand on optimise sur l'orthant positif, ici $\mathcal{R}$ est l'indicatrice de l'orthant positif C, alors C n'admet que 0 comme point extrémal et ses rayons extrémaux sont $\{\alpha e_i, \alpha \geq 0\}$ où e_i est un vecteur de la base canonique. Dans ce cas, s'il existe une solution, il en existe une $m-1$ parcimonieuse, voir [DT05].

Si $E = \mathcal{M}$ l'ensemble des mesures de Radon sur Ω et si on choisit

$$\mathcal{R}(u) = \begin{cases} -\infty \text{ si } u \text{ non positive} \\ u(\Omega) \text{ si } u \text{ positive} \end{cases},$$

alors $\mathcal{R}$ est la masse de la mesure u et on minimise sur les mesures positives. Les atomes sont les masses de Dirac et certaines solutions sont combinaisons d'au plus m masses de Dirac, voir [CF14 ; DP15]. Finalement dans le cas où E est l'ensemble des fonctions BV et $\mathcal{R}(u) = \|\nabla u\|_{\mathcal{M}}$, alors les points extrémaux de C sont les ensembles simples (connexes et à complémentaire connexe) et la minimisation BV favorise les solutions s'écrivant avec m atomes. Ceci explique le phénomène de "staircaising".

4.1.3 Pour aller plus loin

Le théorème de [Boy+19] est un peu plus explicite que la forme énoncée ici. Il dit que les points extrémaux de $S_B^\star$ vivent dans une face de dimension au plus $m-1$ de C_B et la situation peut se révéler plus favorable qu'une application directe du théorème. Par exemple, concernant C le cône des matrices semi-définies positives. Les points extrémaux sont les matrices de rang 1 et en appliquant le théorème, on retrouve le fait qu'il existe des solutions qui s'écrivent comme $m-1$ somme de matrice de rang 1 et donc des matrices de rang $m-1$. Cependant les faces de dimension m du cône sont composées de matrice de rang $O(\sqrt{m})$ [Dat10, Sec.2.8.2.2], donc la borne sur le rang est en $\sqrt{m}$. Similairement, concernant la norme TV, si deux ensembles simples appartiennent à la même face, alors leur intersection aussi ainsi que toutes les composantes connexes de cette intersection. Ainsi les faces de dimension 2 ont pour points extrémaux des ensembles soit disjoints soit inclus l'un dans l'autre. Il existe d'autres règles basées sur le passage au complémentaire, mais ce qui est important de retenir est que dans le cas $m = 3$, les atomes ont des propriétés entre eux qui nécessite l'étude fine de la structure des faces.

L'autre intérêt de ce théorème est de se limiter dans le problème initial (4.1) à rechercher des solutions qui vérifient le problème de représentation. En supposant que C n'admette pas de ligne et en notant $\mathcal{A}$ l'ensemble des atomes, on obtient

$$\min_{\alpha \in \mathbb{R}^{+m}, e \in \mathcal{A}} f(\sum_i \alpha_i \Phi e_i) + \mathcal{R}(\sum_i \alpha_i e_i),$$

De manière générique, on peut s'attendre à $\mathcal{R}(\sum_i \alpha_i e_i) = \sum_i a_i$. Si f est convexe, on décompose le problème en un problème d'optimisation convexe standard sur les α et un problème posé sur les atomes. Cette démarche est utilisée dans [FW19] et est prometteuse d'applications.

4.2 [FGW20],[FGW19] : Programmation semi-infinie

Dans [FGW20] nous nous intéressons aux problèmes du type

$$\min_{\mu \in \mathcal{M}(\Omega)} f(\Phi\mu) + |\mu|(\Omega), \tag{4.3}$$

où $\mathcal{M}$ est l'ensemble des mesures de Radon et Ω un sous-ensemble de $\mathbb{R}^n$ et Φ est linéaire à valeurs dans $\mathbb{R}^m$. En se donnant des mesures cibles y et en prenant le choix $f(x) = 0$ si $x = y$ et $+\infty$, on retrouver le problème basis pursuit [CDS01] et en prenant $f(x) = \frac{\tau}{2}\|x-y\|^2$, on retrouve une variante du Lasso, le Beurling-lasso [Tib96 ; DG12]. En utilisant [Boy+19], nous savons qu'il existe une solution écrite sous la forme d'au plus m mesures de Dirac :

$$d\mu^\star = \sum_{i=1}^{m} \alpha_i^\star \delta_{x_i^\star}, \quad s \le m. \tag{4.4}$$

On peut donc couper ce problème en deux parties et à chaque itération k, se fixer Ω^k un domaine et chercher uniquement le vecteur $\alpha_k^\star$ solution du problème d'optimisation avec régularisation ℓ^1 :

$$\min_{\alpha\in\mathbb{R}^n} f\Big(\sum_{x_i\in\Omega^k} \alpha_i \Phi \delta_{x_i}\Big) + \|a\|_{\ell^1}. \tag{4.5}$$

Parmi les choix d'algorithmes, on peut choisir de fixer Ω^k à l'avance, comme une grille cartésienne fixe, ce qui rend le problème souvent trop complexe [TBR13 ; DP15]. Dans [BP13], les auteurs rroposent d'enrichir Ω^k au fur et à mesure des itérations. Dans [TA20], les auteurs proposent une méthode de gradient sur les variables couplées (a, x) et finalement, dans [BSR17 ; Den+19] les auteurs proposent d'optimiser sur les variables couplées (a, x) puis d'augmenter Ω^k au fur et à mesure. Dans [FGW20] nous proposons une stratégie mixte de raffinement et d'optimisation en montrant un taux de convergence linéaire.

4.2.1 Stratégie de raffinement

Sous des hypothèses minimales, le problème (4.3) admet un **problème dual** qui est un problème équivalent et qui est donné par

$$\sup_{\|\Phi^\star q\|_{L^\infty(\mathcal{O})}\le 1} -f^\star(q), \tag{4.6}$$

où $f^\star$ est la transformée de Legendre de f.

Démonstration

Il s'agit d'un développement classique qui suit les lignes suivantes

$$\begin{aligned}
\inf_{\mu\in\mathcal{M}(\mathcal{O})} f(\Phi\mu) + |\mu|(\mathcal{O}) &= \inf_{\mu\in\mathcal{M}(\mathcal{O})} \underbrace{\sup_q \langle q, \Phi\mu\rangle - f^\star(q)}_{=f(\Phi\mu)} + |\mu|(\mathcal{O}) \\
&= \sup_q \underbrace{\inf_{\mu\in\mathcal{M}(\mathcal{O})} \langle \Phi^\star q, \mu\rangle + |\mu|(\mathcal{O})}_{=\begin{cases} 0 \text{ if } \|\Phi^\star q\|_{L^\infty(\mathcal{O})} \le 1 \\ -\infty \text{ if } \|\Phi^\star q\|_{L^\infty(\mathcal{O})} > 1 \end{cases}} - f^\star(q)
\end{aligned}$$

Sous des hypothèses faibles de régularité de f, il existe un seul maximiseur au problème (4.6). On note $q^\star$ la solution de (4.6) pour $\mathcal{O} = \Omega$ et q_k celle pour Ω^k. Afin de comprendre la stratégie de raffinement, nous donnons la proposition suivante qui permet de déterminer si la solution du problème pour Ω^k est aussi solution du problème pour Ω.

Proposition 4.2 — certificat dual. Soit $\Omega_k \subset \Omega$, si q_k est l'unique solution de (4.6) avec $\mathcal{O} = \Omega^k$, alors forcément $\|\Phi^\star q_k\|_{L^\infty(\Omega^k)} \le 1$. Si en plus on a $\|\Phi^\star q_k\|_{L^\infty(\Omega)} \le 1$, alors $q_k = q^\star$ et la solution de (4.5) est l'unique solution de (4.3).

Démonstration

Comme $\Omega^k \subset \Omega$, le problème (4.6) a moins de contraintes quand $\mathcal{O} = \Omega^k$ que quand $\mathcal{O} = \Omega$. Ainsi la valeur du supremum de (4.6) est plus grande pour $\mathcal{O} = \Omega^k$ que pour $\mathcal{O} = \Omega$ et ainsi $-f^\star(q^k) \ge -f^\star(q^\star)$. Comme q^k est admissible pour $\mathcal{O} = \Omega$, et que $q^\star$ est le supremum de $-f^\star$, on a $-f^\star(q^\star) \ge -f^\star(q^k)$, par suite il y a égalité $-f^\star(q^k) = -f^\star(q^\star)$ et par unicité du minimiseur, on a $q^k = q^\star$.

La stratégie de raffinement est tirée de la Proposition 4.2. Si l'hypothèse $\|\Phi^\star q_k\|_{L^\infty(\Omega)} \leq 1$ est vraie, alors on a trouvé la solution. Si elle est fausse, alors elle nous donne une indication sur les points qu'il faut rajouter à Ω^k :

Algorithme

A chaque itération k :

- Si Ω^k est donné, on calcule q^k, solution de (4.6) avec $\mathcal{O} = \Omega^k$.
- On calcule X^k, qui sont tous les maximums locaux de $x \mapsto \Phi^\star q^k(x)$ qui sont strictement plus grand que 1. Ce sont les points où l'hypothèse $\|\Phi^\star q_k\|_{L^\infty(\Omega)} \leq 1$, nécessaire dans la Proposition 4.2, est la plus contredite.
- On rajoute à Ω^k les points de X^k au sens où $\Omega^{k+1} = \Omega^k \cup X^k$.

Nous montrons dans (4.3) qu'après un nombre d'itérations finies, l'algorithme rajoute au plus s points à chaque itération et nous montrons qu'il existe un taux de convergence linéaire pour

$$d(x^\star, \Omega^k) = \sup_i \inf_{x \in \Omega^k} \|x_i^\star - x\|. \tag{4.7}$$

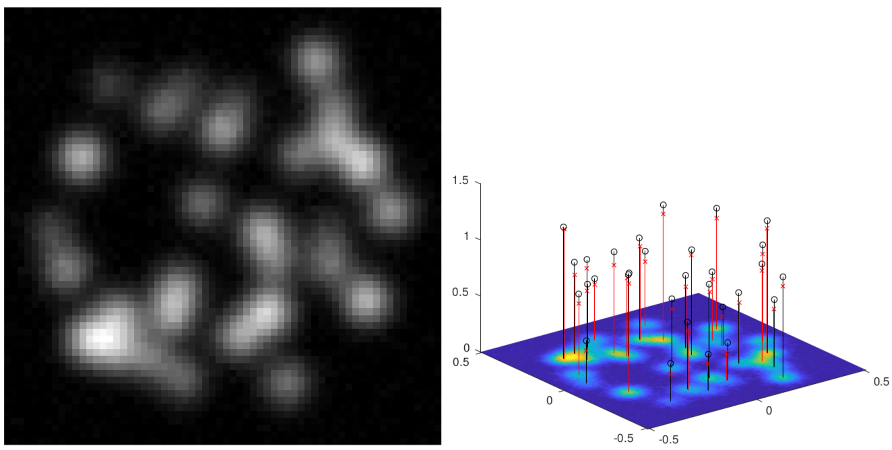


Figure 4.1 – Illustration de [FGW20], à gauche les données qui sont des mesures convolées avec des Gaussiennes auxquelles on a rajouté du bruit. A droite, le résultat de l'algorithme de reconstruction. Les barres verticales bleues avec les cercles représentent les positions et les amplitutes des données synthétiques, les barres rouges avec des croix représentent la reconstruction. A part un léger biais en amplitude dû à la reconstruction ℓ_1, les données sont quasiment retrouvées exactement.

Pour obtenir ces résultats, nous devons supposer que f est convexe avec un gradient L-Lipchitz et nous devons supposer que Φ est assez régulière au sens où il existe des fonctions $a_i \in C^2(\Omega)$ telles que

$$(\Phi d\mu)_i = \int_\Omega a_i d\mu.$$

Nous avons de plus besoin d'une condition similaire à la "non-degenerate source condition" : Si $q^\star$ est le point optimum du problème dual de (4.3), alors la fonction $x \mapsto |\Phi^\star q^\star|(x)$ ne doit atteindre son maximum qu'aux points $x_i^\star$ et doit être localement strictement concave en ces points.

La démonstration en elle-même de la convergence de l'algorithme de raffinement est relativement complexe, elle est similaire à un algorithme de boot-strap où l'on peut déduire que q^k est proche de $q^\star$ et $d(x^\star, \Omega^k)$ est faible, grâce à la régularité de Φ et à la stricte concavité de $\Phi q^\star$ autour de $x_i^\star$. Ainsi les maximums locaux de $\Phi^\star q_k$ seront proches des maximums locaux de $\Phi^\star q^\star$. En retour, comme ces maximums locaux sont ajoutés à Ω^k alors $d(x^\star, \Omega^{k+1})$ se réduit et aussi q^{k+1} se rapprochera de $q^\star$. On doit cependant utiliser un argument de recouvrement fini pour assurer que $d(x^\star, \Omega^k)$ décroît à la bonne vitesse.

4.2.2 Algorithme final

Nous nous intéressons aussi à l'algorithme de gradient en (a, x) et montrons l'existence d'un bassin d'attraction. Finalement nous proposons un algorithme hybride qui propose le meilleur des deux mondes. Nous faisons une étape de raffinement qui nous rajoute de nouveaux points, puis nous lançons une méthode de gradient sur les points que nous venons de rajouter, nous attendons la convergence et nous recommençons.

4.3 [FGW23] : Super-résolution hors-grille résolue avec une grille

Le problème majeur de [FGW20] est que si des points trop proches sont rajoutés dans Ω_k alors le problème (4.5) commence à être mal conditionné. De plus, quand le nombre de points dans Ω_k augmente, le problème (4.5) est difficile à résoudre numériquement. Troisièmement, la recherche systématique des maximiseurs locaux de $x \mapsto |\Phi^\star(x)q^k|$ est coûteuse. Elle est résolue par une méthode de gradient et peut être sensible aux minimas locaux. Nous désirons résoudre ces problèmes en nous intéressant à des algorithmes multi-résolution sur des grilles.

L'algorithme est le suivant

Algorithme

Pour un $J \in \mathbb{N}$ donné :

- **Problème discret :** Nous résolvons le système discret (4.6) sur une grille cartésienne $\mathcal{O} = \Omega_k$.
- **Calcul du critère :** Pour chaque cellule de la grille de $\Omega_k \subset \mathbb{R}^d$, nous calculons un critère qui détermine si la cellule doit être raffinée par dichotomie (i.e., découpée en 2^d cellules).
- **Restriction sur le raffinage :** Parmi les cellules marquées comme devant être raffinées, nous ne raffinons que les plus grosses. Cela donne un nouveau maillage Ω_{k+1}.
- **Critère d'arrêt** : L'algorithme s'arrête dès qu'il raffine une cellule de diamètre plus petit que 2^{-J}.

Les résultats de convergence dépendent du choix de critère de raffinement. Nous proposons un choix qui assure les propriété de convergence suivante.

Proposition 4.3 En choisissant correctement le critère de raffinement, on montre qu'il des constantes $(c_i)_i$ telles que l'algorithme présenté ici a les propriétés suivantes :

- **Convergence linéaire :** Il s'arrête en $k^\star = c_1 + c_2 J$ itérations
- **Contrôle de la complexité :** Le nombre de points dans $\Omega_{k^\star}$ est plus petit que $c_3 J$.
- **Convergence des itérérs :** Chaque point $x^\star$ est à distance au plus $c_4 2^{-J}$ d'un point de $\Omega_{k^\star}$.
- **Convergence des certificats :** On a $\|q_{k^\star} - q^\star\| \leq c_5 2^{-J}$.

Toute la difficulté de [FGW20] est de comprendre quel critère il est pertinent d'implémenter pour que l'algorithme optimise "le plus vite". Par "le plus vite", on entend que les bonnes cellules doivent être raffinées mais que le nombre de cellule n'explose pas, sinon la résolution du système discret devient trop complexe.

- Si le nombre de cellules raffinées n'importait pas (critère de **Contrôle de la complexité**), la stratégie de raffinement uniforme fonctionnerait et serait optimale.

- Nous montrons dans [FGW20], qu'un critère qui raffine les cellules qui contiennent un maximum local de $|A^\star q_k|$ dont la valeur critique dépasse 1 est un critère qui permet d'assurer la convergence. Ce critère correspond à l'étude faite dans [FGW20] où nous rajoutions à chaque itération les maximums locaux dont la valeur critique dépasse 1.
- Afin de limiter le nombre de cellules raffinées à chaque itération nous imposons que l'utilisateur garantisse l'existence d'une constante c telles que les cellules ω qui vérifient

$$\|\Phi^\star q_k\|_{L^\infty(\omega)} < 1 - c|\omega|^2, \tag{4.8}$$

ne sont pas raffinées. Ici $|\omega|$ représente le diamètre de la cellule. Naturellement, le critère que l'on utilise est un critère basé sur des développements de Taylor d'ordre 2.

Démonstration

La **restriction sur le raffinage** est la technique qui permet le **contrôle de la complexité**. Grâce à cette hypothèse, si on raffine une cellule, alors forcément les cellules qui contiennent un maximum local de $|\Phi^\star q_k|$ qui dépasse 1 (qui par définition doivent valider le critère de raffinage) sont forcément plus petites. Donc forcément la valeur maximale de $|\Phi^\star q_k|$ ne peut pas être trop grande (par régularité de $x \mapsto |\Phi^\star q_k|(x)$). Ainsi q_k n'est pas trop éloigné de $q^\star$ et comme $\Phi^\star q^\star$ est strictement concave autour de ses maximums locaux qui dépassent 1, alors la cellule qui est raffinée doit être proche de ces maximums locaux de $\Phi^\star q^\star$. Ainsi on ne raffine que dans un voisinage de ces maximums locaux et on peut garantir de cette sorte un contrôle sur la complexité.

4.4 Perspectives

Les perspectives principales sont des améliorations de l'algorithme de [FGW23] et sont au nombre de trois. Premièrement, on pourrait vouloir abandonner certaines cellules qui sont trop grosses et trop loin des maximums locaux, elles sont inutiles pour l'optimisation. Cependant il faut concevoir un critère facile à implémenter qui puisse permettre de retirer ces cellules. Un autre axe d'amélioration est d'étendre cet algorithme à d'autres critères de régularisation que la variation totale. Puis, pour proposer à la communauté un code viable, il faut que l'on retravaille sur le solveur discret qui, pour l'instant, n'est pas adapté (il n'a notamment, à l'itération k, pas de mémoire du calcul à l'itération $k-1$.)

Bibliographie

[BSR17] Nicholas Boyd, Geoffrey Schiebinger et Benjamin Recht. "The alternating descent conditional gradient method for sparse inverse problems". In : *SIAM Journal on Optimization* 27.2 (2017), p. 616-639.

[Boy+19] Claire Boyer, Antonin Chambolle, Yohann de Castro, Vincent Duval, Frédéric de Gournay et Pierre Weiss. "On representer theorems and convex regularization". In : *SIAM Journal on Optimization* 29.2 (2019), p. 1260-1281.

[BP13] Kristian Bredies et Hanna Katriina Pikkarainen. "Inverse problems in spaces of measures". In : *ESAIM : Control, Optimisation and Calculus of Variations* 19.1 (2013), p. 190-218.

[CF14] Emmanuel J Candès et Carlos Fernandez-Granda. "Towards a mathematical theory of super-resolution". In : *Communications on pure and applied Mathematics* 67.6 (2014), p. 906-956.

[CRT06] Emmanuel J Candès, Justin Romberg et Terence Tao. "Robust uncertainty principles : Exact signal reconstruction from highly incomplete frequency information". In : *IEEE Transactions on information theory* 52.2 (2006), p. 489-509.

[Cha+12] Venkat Chandrasekaran, Benjamin Recht, Pablo A Parrilo et Alan S Willsky. “The convex geometry of linear inverse problems”. In : *Foundations of Computational mathematics* 12.6 (2012), p. 805-849.

[CDS01] Scott Shaobing Chen, David L Donoho et Michael A Saunders. “Atomic decomposition by basis pursuit”. In : *SIAM review* 43.1 (2001), p. 129-159.

[Dat10] Jon Dattorro. *Convex optimization & Euclidean distance geometry.* Lulu. com, 2010.

[DG12] Yohann De Castro et Fabrice Gamboa. “Exact reconstruction using Beurling minimal extrapolation”. In : *Journal of Mathematical Analysis and applications* 395.1 (2012), p. 336-354.

[Den+19] Quentin Denoyelle, Vincent Duval, Gabriel Peyré et Emmanuel Soubies. “The sliding Frank–Wolfe algorithm and its application to super-resolution microscopy”. In : *Inverse Problems* 36.1 (2019), p. 014001.

[DT05] David L Donoho et Jared Tanner. “Sparse nonnegative solution of underdetermined linear equations by linear programming”. In : *Proceedings of the national academy of sciences* 102.27 (2005), p. 9446-9451.

[DP15] Vincent Duval et Gabriel Peyré. “Exact support recovery for sparse spikes deconvolution”. In : *Foundations of Computational Mathematics* 15.5 (2015), p. 1315-1355.

[FGW19] Axel Flinth, Frédéric de Gournay et Pierre Weiss. “Eventual linear convergence rate of an exchange algorithm for superresolution”. In : SPARS. 2019.

[FGW23] Axel Flinth, Frédéric de Gournay et Pierre Weiss. “Grid is Good : Adaptive Refinement Algorithms for Off-the-Grid Total Variation Minimization”. 2023.

[FGW20] Axel Flinth, Frédéric de Gournay et Pierre Weiss. “On the linear convergence rates of exchange and continuous methods for total variation minimization”. In : *Mathematical Programming* (2020), p. 1-37.

[FW19] Axel Flinth et Pierre Weiss. “Exact solutions of infinite dimensional total-variation regularized problems”. In : *Information and Inference : A Journal of the IMA* 8.3 (2019), p. 407-443.

[TBR13] Gongguo Tang, Badri Narayan Bhaskar et Benjamin Recht. “Sparse recovery over continuous dictionaries-just discretize”. In : *2013 Asilomar Conference on Signals, Systems and Computers.* IEEE. 2013, p. 1043-1047.

[Tib96] Robert Tibshirani. “Regression shrinkage and selection via the lasso”. In : *Journal of the Royal Statistical Society : Series B (Methodological)* 58.1 (1996), p. 267-288.

[TA20] Yann Traonmilin et Jean-François Aujol. “The basins of attraction of the global minimizers of the non-convex sparse spike estimation problem”. In : *Inverse Problems* 36.4 (2020), p. 045003.

Chapitre 5

Transport Optimal

5.1 Présentation du transport optimal

Étant donné deux mesures de probabilité μ et ν sur $\mathbb{R}^d$, la distance de Wasserstein (pour la norme 2) entre ces deux mesures, qui est une distance de transport optimal est donnée par :

$$W^2(\mu,\nu) = \inf_{\gamma\in\Pi} \int_{\mathbb{R}^{2d}} \|x-y\|^2 d\gamma(x,y),$$

Où $\|\bullet\|$ représente la norme euclidienne et Π est l'ensemble des mesures sur $\mathbb{R}^{2d}$ dont les marginales sont respectivement μ et ν. En d'autres termes γ appartient à Π si et seulement si, pour tout ensemble mesurable A,

$$\gamma \in \Pi \iff \left(\forall A \quad \gamma(\mathbb{R}^d, A) = \nu(A) \text{ et } \gamma(A, \mathbb{R}^d) = \mu(A)\right). \tag{5.1}$$

Par nature, le calcul de la distance de Wasserstein est un problème de programmation linéaire en γ. Un bon moyen de visualiser le problème de programmation linéaire est de supposer que les deux mesures μ et ν sont des sommes finies de mesures de Dirac de positions respectives $(x_i)_i$ et $(y_j)_j$ associées à des poids respectifs $(m_i)_i$ et $(n_j)_j$. Dans ce cas $\gamma \in \Pi$ est une somme finie de mesures de Dirac de positions $((x_i, y_j))_{ij}$ associées à des poids $(\gamma_{ij})_{ij}$. En supposant que les contraintes sur les positions $((x_i, y_j))_{ij}$ sont vérifiées, alors γ est vérifie les contraintes de marginale si et seulement si :

$$\gamma \in \Pi \Leftrightarrow \sum_i \gamma_{ij} = n_j \quad \text{ et } \quad \sum_j \gamma_{ij} = m_i. \tag{5.2}$$

Et le problème de Wasserstein s'écrit

$$W^2(\mu,\nu) = \inf_{\gamma\in\Pi} \sum_{i,j} \|x_i - y_j\|^2 \gamma_{ij}. \tag{5.3}$$

En tant que problème de programmation linéaire, le calcul de la distance de Wasserstein admet une formulation duale qui est donnée par

Proposition 5.1 — Formulation duale. Pour tout ϕ, on pose $\phi^c(y) = \inf_x \|x-y\|^2 - \phi(x)$. Dans

ce cas
$$W^2(\mu,\nu) = \sup_\phi \int_{\mathbb{R}^d} \phi^c d\nu + \int_{\mathbb{R}^d} \phi d\mu$$

Démonstration

On suit les lignes classiques de la formulation duale en posant ϕ et ψ, deux multiplicateurs de Lagrange pour les contraintes de marginales :
$$\begin{aligned} W^2(\mu,\nu) = \inf_{\gamma\geq 0} \sup_{\phi,\psi} \quad & \int_{\mathbb{R}^{2d}} \|x-y\|^2 d\gamma(x,y) + \int_{\mathbb{R}^d} \psi(y)(d\nu(y) - d\gamma(\mathbb{R}^d, y)) \\ & + \int_{\mathbb{R}^d} \phi(x)(d\mu(x) - d\gamma(x, \mathbb{R}^d)) \end{aligned}$$
En échangeant l'infimum et le suprémum, et en résolvant la minimisation en γ sous la contrainte $\gamma \geq 0$, on trouve
$$W^2(\mu,\nu) = \sup_{\phi(x)+\psi(y)\leq\|x-y\|^2} \int_{\mathbb{R}^d} \psi d\nu + \int_{\mathbb{R}^d} \phi d\mu$$
À ϕ fixé, on résout ensuite en ψ ce qui donne comme $\psi = \phi^c$ comme ψ optimal, avec $\phi^c(y) = \inf_x \|x-y\|^2 - \phi(x)$ et on obtient le résultat recherché.

Transport optimal semi-discret Dans le transport optimal semi-discret, on suppose que μ est une somme finie de mesures de Dirac. Dans ce cas, il est intéressant d'introduire les **cellules de Laguerre**.

Proposition 5.2 Supposons que ν est absolument continue par rapport à la mesure de Lebesgue et que $\mu = \sum_i m_i \delta_{x_i}$ est une somme finie de mesures de Dirac de positions $(x_i)_i$ associées aux masses $(m_i)_i$. Pour tout $\phi = (\phi_i)_i$, on partitionne $\mathbb{R}^d$ en cellules de Laguerre, définies par :
$$\mathcal{L}_i(\phi) = \left\{ x : \|y - x_i\|^2 - \phi_i \leq \|y - x_j\|^2 - \phi_j \quad \forall j \right\}.$$
De sorte que $\phi^c(y) = \sum_i \mathbb{1}_{\mathcal{L}_i(\phi)} \left(\|y-x_i\|^2 - \phi_i \right)$. Dans ce cas, le problème de transport optimal devient :
$$W^2(\mu,\nu) = \sup_\phi g(\phi) \text{ avec } g(\phi) = \sum_i \int_{\mathcal{L}_i(\phi)} (\|y-x_i\|^2 - \phi_i) d\nu + m_i\phi_i, \tag{5.4}$$

Démonstration

Les calculs sont assez simples, la subtilité est que ν doit être absolument continue par rapport à la mesure de Lebesgue. Effectivement les cellules de Laguerre ne forment pas réellement une partition de $\mathbb{R}^d$ dans la mesure où l'intersection de deux cellules de Laguerre peut être non vide, ce sont des polytopes de dimension $d-1$. Avec l'hypothèse d'absolue continuité de ν, on obtient une partition à un ensemble de mesure vide (pour ν). Le reste des calculs se comprend au sens ν-presque partout.

Comme g est une fonction duale, elle est concave et son gradient est donné par
$$\frac{\partial g}{\partial \phi_i} = m_i - \int_{\mathcal{L}_i(\phi)} d\nu.$$
Aux points critiques de g, la masse de ν transportée depuis la cellule de Laguerre $\mathcal{L}_i(\phi)$ est bien égale à m_i, la masse présente au point x_i pour μ.

Interprétation physique des cellules de Laguerre L'interprétation physique de ϕ est celle d'un multiplicateur de Lagrange pour la contrainte $\gamma(A, \mathbb{R}^d) = \mu(A)$, c'est-à-dire que la masse transportée vers le point x_i doit être égale à m_i. L'interprétation de $\mathcal{L}_i(\phi)$ est la zone de l'espace où de la masse de ν est enlevée pour être transportée vers le point x_i. Le transport optimal classique consiste à fixer les mesures μ et ν et à trouver ϕ point critique de g. Cependant on peut aussi pour se donner ν, les position $(x_i)_i$ et ϕ et chercher le vecteur de masse $(m_i)_i$ tel que le transport optimal entre μ et ν soit réalisé pour ϕ. Il suffit pour cela d'annuler le gradient de g, c'est-à-dire de fixer

$$m_i = \nu(\mathcal{L}_i(\phi)).$$

Ainsi, pour tout ϕ fixé, les cellules de Laguerre représentent un plan de transport optimal, il faut juste modifier la distribution de masse de μ [Aur87].

Diagrammes de Voronoï On peut remarquer que dans le cas $\phi = 0$, les cellules de Laguerre $(\mathcal{L}_i(0))_i$ sont les cellules classiques de Voronoï. On remarque que

- La variable ϕ est la variable duale sur les masses à mettre aux points x_i. Mettre une variable duale à 0 consiste à relaxer la contrainte.
- Si on fixe ν et la position $(x_i)_i$ et $\phi = 0$, la discussion du paragraphe précédent nous affirme que le choix de masse
$$m_i = \nu(\mathcal{L}_i(0)),$$
est celui qui minimise la distance de transport optimale entre μ et ν.

Ainsi le problème de minimisation

$$\inf_m W^2(\sum_i m_i \delta x_i, \nu)$$

est résolu en traçant le diagramme de Voronoï des x_i et en posant m_i comme la masse de ν sur la cellule de Voronoï. Il est intéressant de remarquer que le plan de transport, donné par le diagramme de Voronoï est indépendant de la mesure cible ν. En termes géométriques ce résultat se comprend de la manière suivante : si vous devez transporter de manière économe une mesure ν vers une mesure discrète μ pour laquelle vous n'avez qu'une contrainte de position en x (et toute latitude pour choisir la masse aux positions fixées), votre meilleure stratégie sera d'affecter chaque unité de masse de la mesure ν aux positions x_i les plus proches. Le plan de transport ainsi crée suit exactement le diagramme de Voronoï.

5.2 [Gou+17] : Choix de la méthode de résolution

Dans [Gou+17], nous nous sommes posé la question de savoir quelle méthode de calcul de la distance de Wasserstein était la plus facile à utiliser entre deux mesures dont

- l'une μ sera issue d'une discrétisation d'une courbe. On intéressera notamment au cas où μ sera portée par des mesures de Dirac qui représentent un échantillonnage fin de la courbe.
- l'autre ν est issue d'une discrétisation d'une image. La discrétisation peut être réalisée par des mesures de Dirac centrées en chaque pixel ou par une densité constante par pixel (ou une approximation bilinéaire...).
- Une fois discrétisées, les mesures μ et ν doivent avoir un nombre de variables importantes (de l'ordre du million)
- La distance de Wasserstein doit pouvoir être différentiable par rapport aux paramètres de discrétisation des mesures (position ou masse si une mesure est discrétisée par une somme de mesures de Dirac).

Un travail similaire avait déjà commencé dans [Cha+17] avec des distances définies comme

$$d(\mu, \nu) = \|h \star (\mu - \nu)\|_{L^2},$$

qui appellent un traitement par transformation de Fourier. Cependant comme μ n'a pas de structure cartésienne, il faut utiliser des outils de transformée de Fourier non uniforme, voir [PS03]. Nous voulions voir si on pouvait calculer des distances de Wasserstein avec des coûts similaires.

Programmation linéaire La première idée de calcul est de discrétiser ν comme une somme de mesures de Dirac et de revenir à un problème de programmation linéaire. Si $\mu = \sum_i m_i \delta_{x_i}$ et $\nu = \sum_j n_j \delta_{y_j}$ le problème devient

$$W^2(\mu,\nu) = \inf_{\gamma, \sum_i \gamma_{ij} = n_j, \sum_j \gamma_{ij} = m_i} \sum_{i,j} \|x_i - y_j\|^2 \gamma_{ij}. \tag{5.5}$$

Les problèmes linéaires sont relativement faciles à résoudre, cependant le problème de cette formulation est son manque de différentiabilité aux paramètres et le nombre important de variable à prendre en considération (γ est un plan de transport, qui a environ 10^{12} variables ! !).

Algorithme de Sinkhorn Pour circonvenir aux problèmes de la programmation linéaire, les auteurs de [Cut13 ; Ben+15] proposent de rajouter un terme de régularisation de style entropique au problème linéaire, ce qui rend la fonction coût strictement convexe et de résoudre

$$W^2(\mu,\nu) = \inf_{\gamma, \sum_i \gamma_{ij} = n_j, \sum_j \gamma_{ij} = m_i} \sum_{i,j} \|x_i - y_j\|^2 \gamma_{ij} - \varepsilon \gamma_{ij} \left(log(\gamma_{ij}) - 1 \right). \tag{5.6}$$

Dans ce cas, en passant en variable duale des contraintes linéaires, et en utilisant un algorithme de divergence de Bregman, on obtient l'algorithme de Sinkhorn qui se résume à trouver les vecteurs duaux u et v par l'algorithme suivant :

$$u_i \leftarrow \frac{m_i}{(\xi v)_i} \quad v_j \leftarrow \frac{n_j}{(\xi^T u)_j}, \tag{5.7}$$

où ξ est la matrice $\xi_{ij} = e^{-\|x_i - y_j\|^2/\varepsilon}$. Le calcul de produits matrice /vecteur ξv et $\xi^T u$ est très simple si les points x_i et y_j sont sur une grille et se fait par transformée de Fourier indépendamment dans chaque direction. Cependant cette méthode contient deux défauts qui sont rédhibitoires pour l'usage que l'on veut faire de la distance de Wasserstein :

- Comme nous ne sommes pas capable de stocker la matrice ξ, il faut calculer le produit matrice vecteur à chaque itération. Cependant comme μ n'a pas de structure cartésienne, la meilleure manière de calculer ce produit matrice-vecteur est de calculer des transformées de Fourier non uniforme, ce que nous voulions éviter.
- La régularisation entropique a tendance à vouloir étaler la mesure de couplage γ et son influence est contrôlée par le paramètre ε qui est subtil à calibrer, on doit s'assurer que $\varepsilon^{1/2}$ doit être petit par rapport aux distances caractéristiques du problème. Le souci majeur est que la régularisation entropique diffuse γ uniformément dans toutes les directions et dans l'usage que nous prévoyions, qui est celle de points alignés sur une courbe, il y a deux distances caractéristiques différentes : la première est la distance entre deux points de la discrétisation de la courbe et la seconde est la distance de transport optimal des points de la courbe vers les points de ν. En notant α_1, la distance entre les points et α_2, la distance à laquelle un point envoie sa masse par le transport optimal, une estimation rapide de α_2 est obtenue en supposant que tous les points s'envoient sur un rectangle de tailles $\alpha_2 \times \alpha_1$ et on obtient donc $n\alpha_2\alpha_1 \simeq 1$. Comme α_1 est choisi d'ordre $1/n$, α_2 est d'un ordre supérieur à α_1. Ainsi on ne peut choisir un paramètre ε de régularisation adéquat dans (5.7).

Algorithme de Laguerre Nous avons donc développé une méthode issue de [Goe+12 ; Lév15 ; KMT16] qui discrétise la mesure de fond ν comme étant une mesure continue (constante ou bilinéaire par pixel) et qui utilise la formulation duale du problème de Kantorovitch (5.4). Notre apport à cette méthode est une manière efficace d'évaluer la fonction coût g et ses dérivées en utilisant la structure cartésienne sous-jacente de ν. Nous utilisons CGAL [Fab+00] pour calculer les cellules de Laguerre et une fois ces cellules obtenues, il faut encore calculer g et $\nabla_\phi g$ (et éventuellement $\partial^2_{\phi^2} g$), ce qui suppose de calculer différents moments de ν sur les polygones $\mathcal{L}_i$. Nous nous sommes ensuite intéressés à l'algorithme le plus rapide et le plus efficace pour calculer le transport optimal. Nous nous sommes arrêtés à une méthode de Levenberg-Marquardt avec calcul de la Hessienne de g qui donne comme itération sur ϕ

$$\phi^{k+1} = H^{-1}\nabla_\phi g(\phi^k) \quad \text{avec } H = \partial^2_{\phi^2} g(\phi^k) + cId$$

Où le paramètre de Levenberg-Marquardt $c \geq 0$ est calculé à chaque itération pour assurer la condition d'Armijo de décroissance de la fonction coût.

5.3 [GKL19b] : Régularité de la fonctionnelle de coût et optimisation des positions

Dans [GKL19b], nous nous sommes posés la question de savoir quelles hypothèses la fonctionnelle g définie dans (5.4) était différentiable et nous étudions des algorithmes d'optimisation du transport optimal.

Conditions de différentiabilité Nous regardons les variations de g par rapport au multiplicateur de Lagrange ϕ, par à x, la position des mesures de Dirac de la mesure μ. Des résultats de différentiabilité première et seconde par rapport à la variable ϕ avaient déjà été obtenus dans [KMT16] et les résultats de dérivés secondes par rapport à la variable x sont connus mais les démonstrations de ces derniers résultats restent formelles. Les résultats de [GKL19b] sont dans un cadre plus général que la distance de Wassertein euclidienne (on s'autorise un coût de transport plus général que $\|x-y\|^2$) cependant dans le cadre de la distance de Wassertein euclidienne, ils sont résumés par :

- La fonctionnelle g est directionnellement différentiable par rapport à ϕ et x.
- La fonctionnelle g est différentiable par rapport aux variables ϕ et x si la mesure ν ne charge pas les intersections des cellules de Laguerre au sens où $\nu(\mathcal{L}_i \cap \mathcal{L}_j) = 0$ si $i \neq j$.
- La fonctionnelle g est différentiable deux fois par rapport aux variables ϕ et x si la mesure ν admet une densité (par rapport à la mesure de Lebesgue) continue et $W^{1,1}$.

Les valeurs des différentielles d'ordre 1 sont, si $\hat{m}_i = \int_{\mathcal{L}_i} d\nu$ est la masse de la $i^{\text{ème}}$ cellule de Laguerre par rapport à la mesure ν et si $\hat{x}_i = (\int_{\mathcal{L}_i} x d\nu)/\hat{m}_i$ est son barycentre :

$$\frac{\partial g}{\partial x_i} = 2(x_i - \hat{x}_i)\hat{m}_i \quad \text{et } \frac{\partial g}{\partial \phi_i} = m_i - \hat{m}_i \tag{5.8}$$

Les formules de dérivations d'ordre 2 sont les formules attendues et déjà démontrées de manière formelle.

Bruit bleu et pointillation Nous nous intéressons dans [GKL19b] à deux problèmes d'optimisation en distance de Wasserstein. Etant donné une mesure de fond ν donnée, on cherche une mesure μ portée par un nombre fini de mesures de Dirac qui soit la plus proche de ν au sens de Wasserstein. Dans le problème du **Bruit bleu** (Blue Noise), seules les positions des mesures de Dirac de μ sont des variables d'optimisation tandis que dans la **pointillation** (Stippling), les positions des mesures de Dirac ET leur masses sont des variables d'optimisation.

Les algorithmes usuels de résolution du Bruit bleu et de la pointillation sont les suivants, dans le cas de la pointillation, il porte le nom de Algorithme de LLoyd :

Algorithme

- Etant donnés les points $(x_i)_i$, calcul des cellules de Laguerre $(\mathcal{L}_i(\phi))_i$. Pour la pointillation, on utilise $\phi = 0$ (cellules de Voronoï) et pour le bruit bleu, il faut utiliser calculer le ϕ qui réalise le transport optimal entre μ et ν.
- On calcul $\hat{x}_i$ le barycentre de la cellule de Laguerre pour la mesure ν. On pose $x_i = \hat{x}_i$ et on recommence.

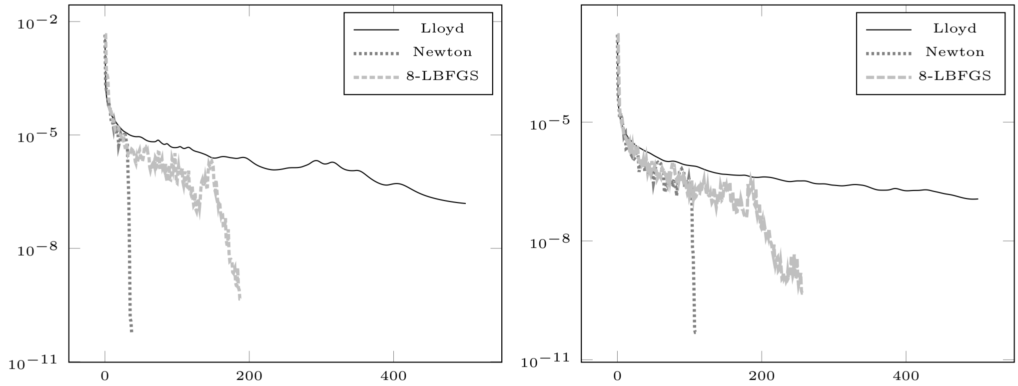


FIGURE 5.1 – Illustration de la Section 5.3 : Evolution de $\nabla_x g$ dans le problème du *Blue Noise*, à gauche pour 1000 points et à droite pour 10000 points.

Au vu de la formule 5.8, ces algorithmes consistent à modifier x_i en

$$x_i - \frac{1}{2\hat{m}_i}\frac{\partial g}{\partial x_i}.$$

Ces algorithmes sont donc des algorithmes de gradient pondérés à la Newton en utilisant la matrice dont la diagonale est $2\hat{m}_i$. Ces algorithmes qui consistent à placer le point au barycentre ont une autre interprétation. Supposons le plan de transport fixé à l'itération k et intéressons nous à la position des points x_i qui minimisent le coût de transport :

$$\min_x \sum_i \int_{\mathcal{L}_i(\phi)} \|x - x_i\|^2 d\nu(x) \quad \text{avec } \phi \text{ fixé}$$

Les points qui résolvent le problème ci-dessus sont les barycentres $\hat{x}_i$. Ainsi l'algorithme de Lloyd (ou sa variante Blue-Noise) peut être vu comme un problème de minimisation alternée du transport optimal en minimisant alternativement entre le plan de transport et la position des points.

Bien que satisfaisant sur le plan géométrique et garantissant la décroissance de la fonction coût, aucune des deux interprétations ci-dessus ne donnent des vitesses de convergence vers des minimums qui soient rapides, nous avons essayé une méthode de Newton et une méthode BFGS similaire à [LL10] sur ce problème et avons accéléré la convergence dans le bassin d'attraction (Voir Figure 5.1).

5.4 [Leb+19] : Le curvling (la courburation)

Dans [Leb+19], nous nous sommes intéressés à approximer au mieux une densité ν par une mesure μ portée par une courbe définie par un paramétrage de vitesse et d'accélération contrôlées. L'objectif est d'optimiser des schémas d'échantillonnage en IRM où il nous est indiqué une densité ν dans l'espace de Fourier que l'opérateur aimerait échantillonner lors de l'acquisition de l'image. Cependant des contraintes techniques font qu'il n'est capable que d'échantillonner le long d'une courbe $\gamma : [0,1] \mapsto \mathbb{R}^d$ avec des contraintes de vitesse et d'accélération données. Nous modéliserons cela en supposant que la mesure μ est une somme de n mesures de Dirac aux points $x_i = \gamma(\frac{i}{n-1})$ et de même masse $\frac{1}{n}$. Les contraintes de vitesse (A_1) et d'accélération (A_2) sont traduites numériquement par

$$\|(A_1 x)_i\|_2 \leq \alpha_1 \quad \|(A_2 x)_i\|_2 \leq \alpha_2 \qquad \text{avec } (A_1 x)_i = x_{i+1} - x_i \text{ et } A_2 = A_1^T A_1.$$

Ces contraintes se réécrivent $Ax \in Y$, en notant A la matrice (A_1, A_2) et Y l'ensemble des vecteurs de points de taille $2n$ dont les n premiers points ont une norme euclidienne inférieure à α_1 et les n derniers une norme inférieure à α_2.

FIGURE 5.2 – Illustration de la Section 5.4 : Applications du curvling avec $256K$ points de discrétisation des courbes. A gauche l'image originelle et de gauche à droite des bornes de plus en plus importantes sur la longueur totale de la courbe. La borne sur la courbure est la même sur tout les cas tests.

L'algorithme choisi est un algorithme de gradient projeté. On note Π, l'opérateur de projection pour la norme euclidienne d'un ensemble de points z sur un ensemble x admissible qui respecte les contraintes de vitesse et d'accélération.

L'algorithme de calcul de Π que nous avons choisi est un algorithme d'ADMM qui consiste à ajouter une variable supplémentaire $y = Ax$, à traiter la contrainte $y = Ax$ avec un Lagrangien augmenté de multiplicateur λ, et à minimiser de manière alternée en x puis en y la fonctionnelle :

$$\min_{(x,y), y \in Y} \|x - z\|_{L^2} + \frac{\beta}{2} \|Ax - y - \lambda\|_{L^2}.$$

A chaque étape, le multiplicateur de Lagrange λ est modifié en lui rajoutant la valeur actuelle de $Ax - y$.

Notre algorithme de calcul introduit une méthode de multi-résolution sur le nombre de points n. La discrétisation de la courbe est raffinée par un facteur 2 dès que l'algorithme semble avoir convergé.

L'algorithme est résumé par :

Algorithme

On se donne n points qui vérifient les contraintes et de masses $\frac{1}{n}$. On se donne ν une mesure cible.

- On construit μ la mesure portée par $(x_i)_i$ de poids $\frac{1}{n}$. On calcule le transport optimal μ et ν. avec les cellules de Laguerre.
- On calcule $\hat{x}_i$ les barycentres des cellules de Laguerre.
- On calcule $\tilde{x}_i = \Pi(\hat{x}_i)$, un minimiseur de $\min_{z, Az \in Y} \|\tilde{z} - \hat{x}_i\|^2$. Puis on pose $x_i = \tilde{x}_i$ et on recommence au début jusqu'à convergence.
- On multiplie n, le nombre de points par 2, en rajoutant les points du milieu sur la courbe et on recommence au début jusqu'à convergence.

5.5 [LGK20] : Le transport optimal 3/4 discret

Dans [LGK20], nous nous sommes intéressés au transport optimal 3/4-discret qui est le transport optimal entre une mesure discrète μ et une mesure portée par des segments ν. Le nom 3/4-discret vient d'une "interpolation" entre le transport discret (deux mesures discrètes) et le transport semi-discret (une mesure discrète et une mesure absolument continue par rapport à la mesure de Lebesgue). L'objectif est de calculer ces transports 3/4-discrets puis d'optimiser la distance de Wasserstein 2 par rapport à la mesure ν, c'est-à-dire par rapport aux paramètres des segments définissant le support de ν.

Nous avons déjà démontré dans [GKL19b] que le transport optimal est différentiable par rapport à ν à condition qu'aucun des segments ne soit contenu dans une intersection de cellules de Laguerre $\mathcal{L}_i \cap \mathcal{L}_j$. Sinon la mesure ν chargerait la frontière $\mathcal{L}_i \cap \mathcal{L}_j$. Les formules de dérivation du transport optimal par rapport à ν sont ensuite calculées par rapport aux paramètres des segments définissant ν.

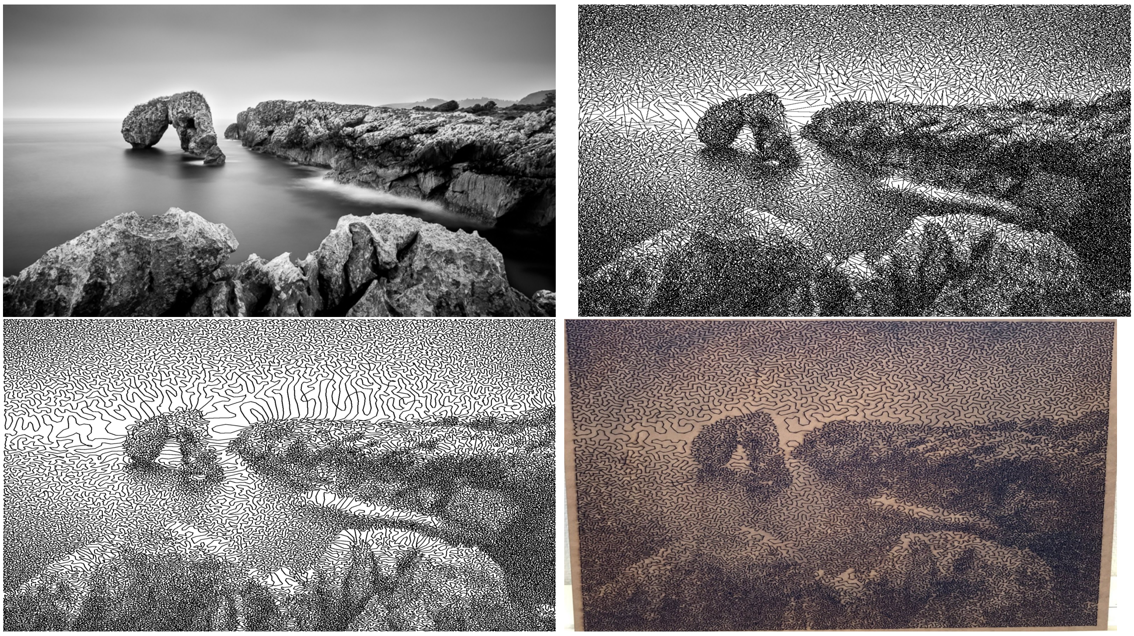


Figure 5.3 – Illustration de transport 3/4-discret. En haut à gauche,l'image originelle. En haut à droite une approximation par 400K segments sans contrainte. En bas à gauche, approximation par 250K segments avec des contraintes cinématiques. Finalement en bas à droite résultat après gravage sur bois final avec 80K segments. Pour tous les tests, la mesure μ est composée de 320K mesures de Dirac.

Concernant la partie numérique, le calcul de la fonction g et de ses différentielles nécessite le calcul des moments de ν sur chaque cellule de Laguerre. Nous utilisons le fait que ν soit portée par des segments pour identifier dans quelle cellule de Laguerre le segment commence et nous sommes capable de calculer le point de sortie ainsi que la cellule de Laguerre de sortie du segment. Ensuite nous calculons les moments de ν sur chaque cellule de Laguerre en avançant sur les segments. Cet algorithme se parallélise facilement sur les segments.

Concernant les choix d'optimisation pour le calcul de transport optimal, c'est à dire l'optimisation de g par rapport à ϕ, le problème principal est que la masse de ν sur les cellules de Laguerre, $\nu(\mathcal{L}_i(\phi))$, est souvent nulle au cours des itérations, contrairement au cas semi-discret où on peut s'arranger pour que ce cas soit une exception. Dans ce cas la Hessienne de g par rapport à ϕ est non-inversible. Nous nous sommes donc arrêtés sur une méthode mixte qui part d'un $L - BFGS$ et qui utilise des itérations de Newton dès que $\nu(\mathcal{L}_i(\phi)) \neq 0$ pour tout i. Cette méthode est couplée avec une recherche linéaire de pas de style Wolfe. Pour l'optimisation par rapport aux paramètres du segment, nous avons implémenté une simple méthode d'ordre 1 avec recherche de pas. Nous avons appliqué ce résultat à la recherche de structure de filaments dans une soupe de points, telle que l'agencement de galaxies en $3d$ ou à l'approximation d'images en $2d$, voir Figure 5.3.

5.6 [GKL19a] : Approximation de courbes

Dans [GKL19a] nous nous intéressons aux approximations de courbes régulières par une suite de points. Dans la continuité de [Leb+19], nous voulons étudier comment approximer, au sens de la distance de Wasserstein une courbe qui obéit à plusieurs contraintes géométriques par une suite de points qui obéit à des contraintes discrètes et vice-versa. Nous calculons explicitement les projecteurs d'un ensemble vers un autre et nous donnons des bornes en fonction du nombre de points qui donnent la vitesse de convergence de ces approximations. Les courbes sont représentées par des fonctions $f : [0, 1] \to \mathbb{R}^d$ périodiques [1] et pour tout vecteur $\alpha \in \mathbb{R}^m$, nous introduisons W^α, l' α multi-boules de Sobolev qui est l'ensemble des fonctions qui vérifient

$$\forall 1 \leq r \leq m, \|f^{(r)}\|_{L^q([0,1])} \leq \alpha_r,$$

où $1 \leq q < +\infty$ est fixé à l'avance. Une α multi-boule est donc l'intersection de m boules de rayon α_r dans $W^{r,q}$. Nous définissons de même pour des suites périodiques de points, P^α, l'α multi-boules de points en remplaçant dans la définition de W^α les dérivées de f par des dérivées numériques des points.

Pour $p = (p_i)_i \in P^\alpha$, on appelle "spline d'ordre 0" la mesure $\frac{1}{n}\sum_{i=1}^n \delta_{p_i}$ qui est la mesure portée par les points $(p_i)_i$. De même on appelle "spline d'ordre 1" la mesure de probabilité portée uniformément sur les segments $[p_{i-1}, p_i]$. Selon le choix des splines (d'ordre 0 et d'ordre 1), on peut ainsi créer deux distances entre des éléments $p \in P^\alpha$ et de $f \in W^\alpha$ qui sont les distances de Wasserstein entre la spline de p et la mesure image de $[0, 1]$ par f. On s'intéresse donc au calcul de la distance de Hausdorff entre P^α et W^α et à trouver des projections de P^α sur W^α et de W^α sur P^α.

Les résultat principaux sur les distances de Hausdorff sont :
- Si $m \geq 1$, en considérant les splines d'ordre 0, la distance de Hausdorff entre W^α et P^α est bornée par $O(n^{-1})$.
- Si $m \geq 2$, en considérant les splines d'ordre 1, la distance de Hausdorff entre W^α et P^α est bornée par $O(n^{-2})$.

Concernant les projections, elles sont construites de la même façon pour les splines d'ordre 0 et les splines d'ordre 1. Elles vérifient les bornes des distances de Hausdorff et sont définies par :
- Si $f \in W^\alpha$ la projection sur P^α est faite en prenant un échantillonnage uniforme de f, ainsi $p_i = f(\frac{i}{n})$
- Si $p \in P^\alpha$, la projection sur W^α est faite en prenant une interpolation par une certaine spline d'ordre m puis en la normalisant pour qu'elle appartienne à W^α.

C'est surtout ce dernier point qui mérite l'attention, car même si les splines sont connues pour optimiser certaines normes de Sobolev, à notre connaissance personne ne s'était intéressé à reconstruire dans des α multi-boules, c'est à dire à considérer plusieurs normes de Sobolev en même temps. Dans la preuve très technique de ce papier, nous avons eu besoin d'une relation de récurrence sur les nombres Eulériens, que nous (en fait Jonas Kahn) avons démontré.

5.7 Perspectives

Récemment, la librairie CGAL a mis à disposition un code de calcul de cellules de Laguerre en dimension arbitraire. Une perspective simple consiste donc à transposer les codes actuels, qui sont limités à la dimension 2 vers la dimension quelconque. Une autre direction de recherche est de comprendre l'intérêt des cellules de Laguerre dans la méthode des volumes finis. Les cellules de Voronoï et leurs duales, les triangulations de Delaunay ont un intérêt majeur dans la méthode des volumes finis et des éléments finis. Les cellules de Laguerre partagent des qualités avec les cellules de Voronoï, par exemple les bords des cellules sont orthogonaux avec la triangulation duale. L'avantage des cellules de Laguerre est que l'on peut manipuler leur masse à notre guise. Peu d'articles ont cependant, à ma connaissance, utilisé les cellules de Laguerre en modélisation. On donnera [Qu+22] comme exemple récent.

1. le cas non-périodique est aussi traité dans [GKL19a] mais est beaucoup plus complexe à détailler ici.

Bibliographie

[Aur87] Franz Aurenhammer. “Power diagrams : properties, algorithms and applications”. In : *SIAM Journal on Computing* 16.1 (1987), p. 78-96.

[Ben+15] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna et Gabriel Peyré. “Iterative Bregman projections for regularized transportation problems”. In : *SIAM Journal on Scientific Computing* 37.2 (2015), A1111-A1138.

[Cha+17] Nicolas Chauffert, Philippe Ciuciu, Jonas Kahn et Pierre Weiss. “A projection method on measures sets”. In : *Constructive Approximation* 45.1 (2017), p. 83-111.

[Cut13] Marco Cuturi. “Sinkhorn distances : Lightspeed computation of optimal transport”. In : *Advances in neural information processing systems.* 2013, p. 2292-2300.

[Fab+00] Andreas Fabri, Geert-Jan Giezeman, Lutz Kettner, Stefan Schirra et Sven Schönherr. “On the design of CGAL a computational geometry algorithms library”. In : *Software : Practice and Experience* 30.11 (2000), p. 1167-1202.

[Goe+12] Fernando de Goes, Katherine Breeden, Victor Ostromoukhov et Mathieu Desbrun. “Blue noise through optimal transport”. In : *ACM Transactions on Graphics (TOG)* 31.6 (2012), p. 171.

[GKL19a] Frédéric de Gournay, Jonas Kahn et Léo Lebrat. “Approximation of curves with piecewise constant or piecewise linear functions”. 2019.

[GKL19b] Frédéric de Gournay, Jonas Kahn et Léo Lebrat. “Differentiation and regularity of semi-discrete optimal transport with respect to the parameters of the discrete measure”. In : *Numerische Mathematik* 141.2 (2019), p. 429-453.

[Gou+17] Frédéric de Gournay, Jonas Kahn, Léo Lebrat et Pierre Weiss. “Approches variationnelles pour le stippling : distances L^2 ou transport optimal ?” In : GRETSI. 2017.

[KMT16] Jun Kitagawa, Quentin Mérigot et Boris Thibert. “Convergence of a Newton algorithm for semi-discrete optimal transport”. In : *arXiv preprint arXiv :1603.05579* (2016).

[LGK20] Léo Lebrat, Frédéric de Gournay et Jonas Kahn. “3/4-Discrete Optimal Transport”. In : *SIAM Journal on Scientific Computing* 42.4 (2020), A2088-A2107.

[Leb+19] Léo Lebrat, Frédéric de Gournay, Jonas Kahn et Pierre Weiss. “Optimal transport approximation of 2-dimensional measures”. In : *SIAM Journal on Imaging Sciences* 12.2 (2019), p. 762-787.

[Lév15] Bruno Lévy. “A numerical algorithm for L2 semi-discrete optimal transport in 3D”. In : *ESAIM : Mathematical Modelling and Numerical Analysis* 49.6 (2015), p. 1693-1715.

[LL10] Bruno Lévy et Yang Liu. “L^p Centroidal Voronoi Tessellation and its applications”. In : *ACM Transactions on Graphics (TOG)* 29.4 (2010), p. 1-11.

[PS03] Daniel Potts et Gabriele Steidl. “Fast summation at nonequispaced knots by NFFT”. In : *SIAM Journal on Scientific Computing* 24.6 (2003), p. 2013-2037.

[Qu+22] Ziyin Qu, Minchen Li, Fernando De Goes et Chenfanfu Jiang. “The power particle-in-cell method”. In : *ACM Transactions on Graphics (TOG)* 41.4 (2022), p. 1-13.

CHAPITRE 6

Travaux récents

Dans ce chapitre, je vais mettre les récents travaux qui ne forment pas encore une section à part entière. Ce sont des travaux qui ont tous en commun de traiter d'intelligence artificielle.

6.1 [Leb+21] : Déformation de maillage avec des réseaux de neurones

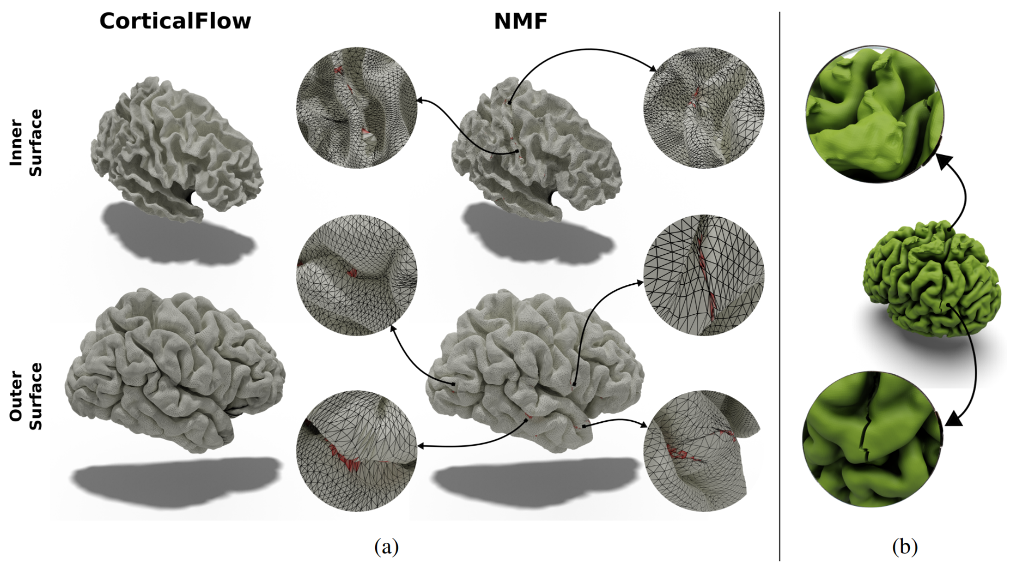


FIGURE 6.1 – a : Surface du cortex prédites par CorticalFlow [Leb+21](gauche) et NMF (droite), les faces qui s'auto-intersectent dans NMF sont représentées en rouge. b : Illustration des problèmes générés par les algorithmes de correction de topologie par DeepCSR.

Dans cet article nous nous intéressons à la segmentation de cerveau, pour cela, on s'intéresse à retrouver à partir d'images $3D$ IRM les surfaces internes et externes du cortex. Ces surfaces sont irrégulières et difféomorphes à des hémisphères. Pour un champ de vecteur v autonome donné, on calcule le flot $\phi_x(t)$ par $\phi_x'(t) = v(\phi_x(t))$ et $\phi_x(0) = x$. On se donne aussi $\mathcal{C}$ un cerveau de référence, et $\mathcal{C}_v$ l'image à l'instant $t=1$ de ce cerveau par le flot ϕ. Cette image est donnée par

$$\mathcal{C}_v = \{\phi_x(1), \text{ tel que } x \in \mathcal{C}\}$$

L'objectif est de créer un programme $P_\Theta : I \to v$ qui prend en entrée des images IRM et qui rend un champ de vecteur v tel que $\mathcal{C}_v$ soit le plus proche possible d'une segmentation de référence de l'image de I. Ici Θ représente le jeu de paramètres du programme P_Θ qui est appris pour des I dans une base de donnée d'entraînement. L'intérêt majeur de [Leb+21] réside dans une méthode multi-échelle (à 3 échelles) où P_Θ est une succession d'algorithmes qui calculent v à des échelles de plus en plus fines. L'optimisation des paramètres Θ se fait aussi de manière multi-échelle, en commençant par l'algorithme le plus grossier. On montre dans [Leb+21] que cette méthode multi-échelle, quand elle est correctement implémentée, permet d'améliorer l'état de l'art sur un grand nombre de critères, par exemple l'algorithme DeepCSR [Cru+21] ou NMF [GC20]. La figure 6.1 permet de se rendre compte des avancées permises par CorticalFlow sur la qualité de la topologie de la segmentation.

6.2 [GG22] : Calibrage du pas d'optimisation par différentiation automatique

Dans cet article, nous valorisons un travail que j'avais initié pour comprendre la différentiation automatique. La question qui m'intéressait et à laquelle je n'ai pas encore de réponse définitive est la signification exacte des termes calculés dans le calcul de rétropropagation de gradient.

6.2.1 Différentiation automatique

Pour poser le problème de la différentiation automatique, nous proposons dans [GG22] le cadre suivant : Soit $\Theta = (\theta_s)_{0\le s\le n}$ des paramètres par rapport auxquels nous voulons dériver. Chaque θ_s appartient à un espace de Hilbert $\mathcal{G}_s$. On se donne un programme séquentiel qui calcule des données $X = (x_s(\Theta))_{0\le s\le n+1}$ avec la formule de récurrence suivante

$$x_{s+1}(\Theta) = \mathcal{F}_s(x_s(\Theta), \theta_s) \quad \forall\, 0 \le s \le n, \tag{6.1}$$

où $\mathcal{F}_s$ est un sous-programme élémentaire régulier. On suppose que chaque x_s appartient à un espace de Hilbert $\mathcal{H}_s$. L'objectif est de calculer le gradient de $\Theta \mapsto x_{n+1}(\Theta)$ si $\mathcal{H}_{n+1} = \mathbb{R}$. Le calcul de différentiation automatique est donné par la proposition suivante

Proposition 6.1 — Rétro-propagation du gradient. Soient $(\partial_x \mathcal{F}_s)^\star$ et $(\partial_\theta \mathcal{F}_s)^\star$ les adjoints des différentielles de $\mathcal{F}_s$ par rapport à x et respectivement θ. Soit $\hat{X} = (\hat{x}_s)_{0\le s\le n+1}$ et $\hat{\Theta} = (\hat{\theta}_s)_{0\le s\le n}$ les adjoints donnés par les récurrences inversées (rétro-propagation) suivantes :

$$\begin{cases} \hat{x}_s = (\partial_x \mathcal{F}_s)^\star \hat{x}_{s+1} & \text{et } \hat{x}_{n+1} = 1 \\ \hat{\theta}_s = (\partial_\theta \mathcal{F}_s)^\star \hat{x}_{s+1}, \end{cases} \tag{6.2}$$

Alors $\hat{\Theta}$ est le gradient de $\Theta \mapsto x_{n+1}(\Theta)$ pour l'espace de Hilbert produit des $\mathcal{G}_s$. C'est-à-dire que pour toute direction $\dot{\Theta} = (\dot{\theta}_s)_{0\le s\le n}$, on a

$$dx_{n+1}.\dot{\theta} = \sum_{s=0}^{n} \langle \dot{\theta}_s, \hat{\theta}_s \rangle_{\mathcal{G}_s}$$

Démonstration

Pour prouver cette proposition, on se fixe une direction $\dot{\Theta} = (\dot{\theta}_s)_s$ quelconque et on regarde $\dot{X} = (\dot{x}_s)_s$ la dérivée directionnelle des données (aussi appelé le tangent). Il se calcule en dérivant

(6.1) :

$$\dot{x}_{s+1} = \partial_x \mathcal{F}_s \dot{x}_s + \partial_\theta \mathcal{F}_s \dot{\theta}_s \quad \text{et} \quad \dot{x}_0 = 0. \tag{6.3}$$

Alors la définition de $\hat{x}_s$ et $\hat{\theta}_s$ nous donne

$$\langle \hat{x}_s, \dot{x}_s \rangle_{\mathcal{H}_s} + \langle \hat{\theta}_s, \dot{\theta}_s \rangle_{\mathcal{G}_s} = \langle \hat{x}_{s+1}, \dot{x}_{s+1} \rangle_{\mathcal{H}_{s+1}}.$$

En sommant ces inégalités de $s = 0$ à n, et en utilisant $\hat{x}_{n+1} = 1$ et $\dot{x}_0 = 0$, on obtient $\dot{x}_{n+1} = \sum_s \langle \hat{\theta}_s, \dot{\theta}_s \rangle$ et ainsi $\hat{\Theta}$ est bien le gradient de $\Theta \mapsto x_{n+1}(\Theta)$.

Si l'interprétation de $\hat{\Theta}$ est claire, celle de $\hat{X}$ l'est beaucoup moins. Nous avons démontré dans [GG22] que $\hat{X}$ intervenait dans le calcul de la Hessienne de x_{n+1} via la formule suivante :

Proposition 6.2 Pour tout s, on note $\nabla^2 \mathcal{F}_s : \mathcal{H}_s \times \mathcal{G}_s \to \mathcal{H}_{s+1}$ la fonction bilinéaire symétrique de la différentiation seconde. Elle vérifie/est définie par la formule de Taylor suivante, vraie pour tout $\dot{x}_s$ et $\dot{\theta}_s$:

$$\mathcal{F}(x_s + \dot{x}_s, \theta_s + \dot{\theta}_s) = \mathcal{F}(x_s, \theta_s) + (\partial_x \mathcal{F})\dot{x}_s + (\partial_\theta \mathcal{F})\dot{\theta}_s + \frac{1}{2}\nabla^2 \mathcal{F}_s \colon (\dot{x}_s, \dot{\theta}_s) \otimes (\dot{x}_s, \dot{\theta}_s) + o\left(\|\dot{x}_s\|^2 + \|\dot{\theta}_s\|^2\right)$$

Pour toute direction $\dot{\Theta} = (\dot{\theta}_s)_s$, si le tangent $\dot{x}_s$ est calculé par la formule (6.3), la Hessienne de $\Theta \mapsto x_{n+1}(\Theta)$ vérifie

$$\nabla^2 x_{n+1} \colon \dot{\Theta} \otimes \dot{\Theta} = \sum_s \left\langle \nabla^2 \mathcal{F}_s \colon (\dot{x}_s, \dot{\theta}_s) \otimes (\dot{x}_s, \dot{\theta}_s), \hat{x}_s \right\rangle,$$

Démonstration

Si nous appelons $\ddot{x}_s$ la dérivée d'ordre 2 dans la direction $\dot{\Theta}$ (2 fois), alors la dérivation de (6.3) donne

$$\ddot{x}_{s+1} = \partial_x \mathcal{F}_s \ddot{x}_s + \frac{1}{2}\nabla^2 \mathcal{F}_s \colon (\dot{x}_s, \dot{\theta}_s) \otimes (\dot{x}_s, \dot{\theta}_s)$$

Alors la définition de $\hat{x}_s$ et $\hat{\theta}_s$ nous donne

$$\langle \hat{x}_s, \ddot{x}_s \rangle_{\mathcal{H}_s} + \frac{1}{2}\langle \hat{x}_s, \nabla^2 \mathcal{F}_s \colon (\dot{x}_s, \dot{\theta}_s) \otimes (\dot{x}_s, \dot{\theta}_s)\dot{\theta}_s \rangle_{\mathcal{H}_s} = \langle \hat{x}_{s+1}, \ddot{x}_{s+1} \rangle_{\mathcal{H}_{s+1}}.$$

En sommant ces inégalités pour les différents s et en utilisant $\hat{x}_{n+1} = 1$ et $\ddot{x}_0 = 0$, on obtient la formule recherchée.

Nous appelons *courbure directionnelle* le terme suivant

$$c(\Theta, \dot{\Theta}) = \frac{\nabla^2 x_{n+1}(\Theta) : \dot{\Theta} \otimes \dot{\Theta}}{\|\dot{\Theta}\|^2} \tag{6.4}$$

Si Θ et $\dot{\Theta}$ sont choisis, la Proposition 6.2 nous assure que ce terme de courbure c est relativement facile à calculer, il suffit de faire trois récurrences, la première calcule X, la deuxième $\hat{X}$ et $\hat{\Theta}$ et la troisième calcule $\dot{X}$ et la courbure. Par rapport au temps de calcul d'un gradient qui nécessite deux passes (calcul de X puis de $(\hat{X}, \hat{\Theta})$), le temps calcul est multiplié par 1.5. L'empreinte mémoire est multipliée par 2. On montre dans [GG22] que l'empreinte mémoire peut être conservée si on accepte de multiplier le temps calcul par 2.

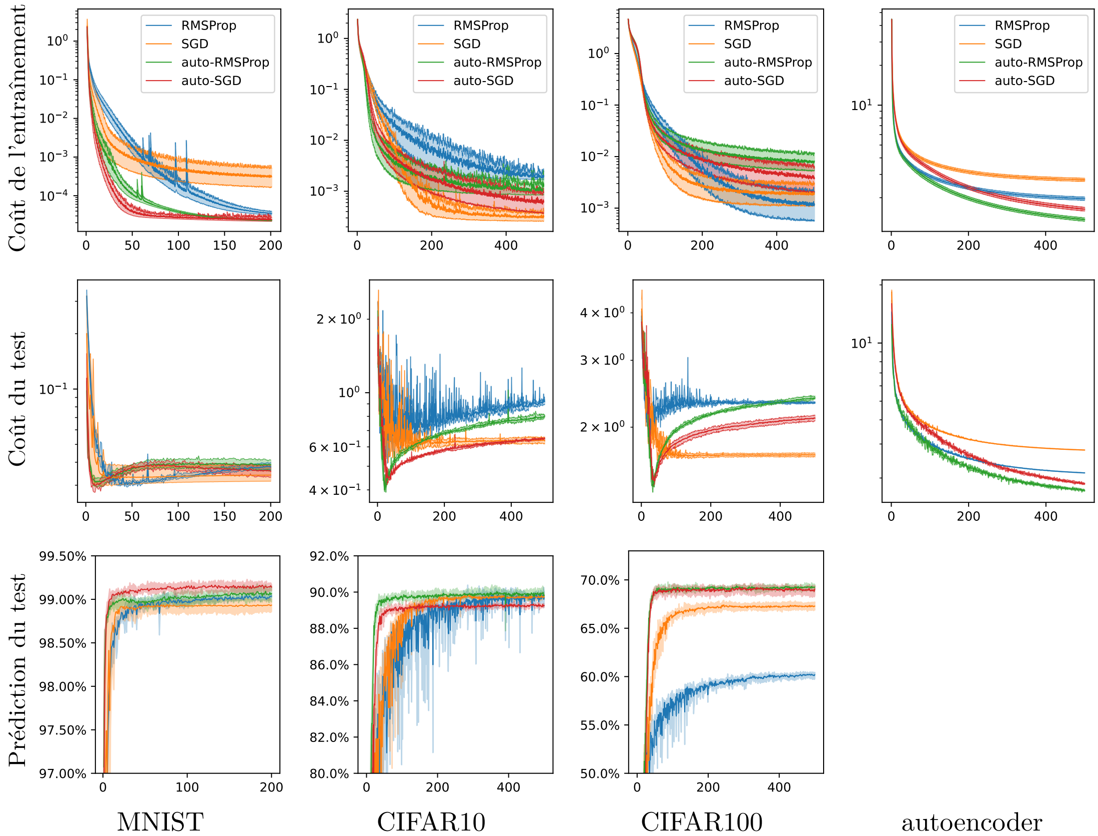


FIGURE 6.2 – Méthode de calibrage [GG22] (resp. vert et rouge) comparée avec les méthodes optimisées manuellement (resp. bleu et orange) pour 4 problèmes standards (classification MNIST,CIFAR10 et CIFAR100 et un auto-encodeur). Les méthodes optimisées manuellement le sont pour le critère *Coût de l'entraînement* pour lequel les méthodes de calibrages restent compétitives. Les méthodes de calibrages sont meilleures sur les autres critères.

6.2.2 Choix du pas avec la courbure

Si la courbure (6.4) est connue, on peut estimer la constante de Lipschitz locale d'une fonction. Effectivement, une formule de Taylor-Young montre rapidement que :

$$x_{n+1}(\Theta - \tau\dot{\Theta}) \leq x_{n+1}(\Theta) - \tau\langle \nabla x_{n+1}(\Theta), \dot{\Theta}\rangle + \tau^2 \frac{L}{2}\|\dot{\Theta}\|^2$$

avec $L = \max_{s\in[0,\tau]} c(\Theta - s\dot{\Theta}, \dot{\Theta})$.

L'objectif est d'estimer L en ayant connaissance de la courbure $c(\Theta, \dot{\Theta})$. La connaissance de L est primordiale en optimisation, car elle dicte le choix du pas. Le choix $\tau = \frac{2}{L}$ permet d'assurer que la fonction ne croît pas (choix d'exploration) et le choix $\tau = \frac{1}{L}$ permet sa décroissance rapide (choix de convergence). Le premier choix permet de tester différents jeux de paramètres en contrôlant le comportement de la fonction coût ainsi *d'explorer* l'ensemble des paramètres. Le deuxième choix permet de *converger* plus vite vers un minimum local mais ne permet pas de sortir des bassins d'attraction. En suivant cette idée, nous proposons dans un cadre stochastique dans [GG22] un algorithme qui nous permet de nous affranchir du choix du pas et de sa décroissance. Nous comparons notre algorithme avec des versions où le pas et sa décroissance ont été optimisés à la main, nous obtenons des résultats compétitifs sur les critères où les algorithmes concurrents ont été optimisés (coût de l'entraînement dans la Figure 6.2) et des résultats meilleurs sur les autres critères (critères dits de "tests").

Le principal problème de l'approche proposée dans cette section est qu'elle ne fonctionne que pour des directions de descente. Les algorithmes les plus performants actuellement en optimisation stochastique sont des algorithmes avec inertie, qui gardent une mémoire des précédents gradients. Nous travaillons actuellement à étendre les idées de cette section aux algorithmes à inertie.

6.3 [GGW20] : On essaie d'optimiser des schémas pour l'IRM

Etant donné une image (une fonction à valeurs réelles et à support compact) à reconstruire u, une mesure d'IRM est modélisée par un opérateur $A[\Xi]$ qui est un échantillonnage en Fourier. Si $\Xi = (\xi_m)_{0 \leq m \leq M-1}$ est une famille de M vecteurs de $\mathbb{R}^3$, alors $A[\Xi]$ est un opérateur linéaire de l'ensemble des images dans $\mathbb{C}^M$ tel que pour tout $m = 0 \ldots M-1$:

$$(A[\Xi]u)_m = \int_{x \in \mathbb{R}^3} e^{i\xi_m \cdot x} u(x) dx.$$

Les points d'échantillonnages $\Xi = (\xi_m)_{0 \leq m \leq M-1}$ doivent respecter des contraintes physiques. Il existe un certain nombre de trajectoires ("shots" en anglais) qui ont une accélération et une vitesse bornée et les ξ_m sont un échantillonnage uniforme en temps de ces trajectoires. On note X_{ad} l'ensemble des Ξ admissibles. L'objectif de l'optimisation de schémas d'échantillonnage pour l'IRM est de donner "le meilleur" schéma possible qui va permettre de retrouver avec le moins d'erreurs possible l'image u initiale. Il existe plusieurs modélisation de ce problème, nous en donnons une ci-dessous :

Définition 6.3 — Problème type d'optimisation de schéma. Soit $\mathcal{U} = (u_p)_p$ une famille d'images à reconstruire, soit w un bruit et $A[\Xi]$ un opérateur d'acquisition. La mesure IRM est y_p, donnée par

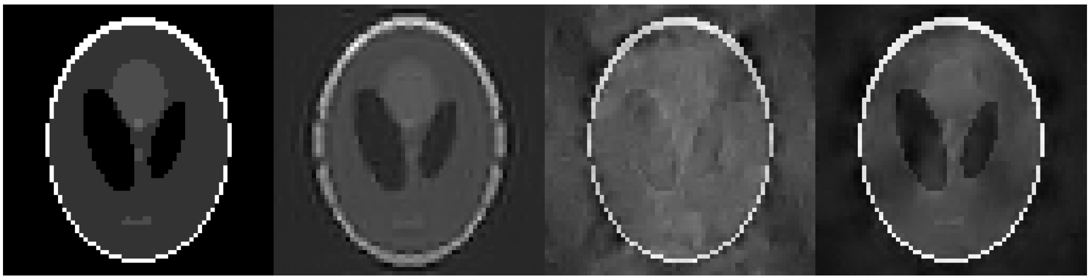
$$y_p = A[\Xi]u_p + w \quad \forall \quad p$$

Si on se donne $\mathcal{R}$ un "reconstructeur", c'est à dire une fonction qui dépend de Ξ et y telle que $\mathcal{R}(\Xi, y) \simeq u$, le problème d'optimisation de schéma est un problème d'optimisation en Ξ donné par

$$\min_{\Xi \in X_{\text{ad}}} \sum_{u \in \mathcal{U}} \mathbb{E}_w \left(\|\mathcal{R}(\Xi, A[\Xi]u + w) - u\|^2 \right)$$

Le problème d'échantillonnage optimal donné par la définition 6.3 a connu un regain d'intérêt dû notamment à la libération de la puissance de calcul GPU (Graphical Processing Unit) et à l'utilisation massive de la différentiation automatique. Parmi les références on citera [Goz+18 ; JUY19 ; She+20 ; BDS19 ; ZHR21 ; Wan+21 ; Wan+21 ; Lok+21 ; Pen+22 ; AJ20] ainsi que [Wei+19] qui sont des collaborateurs.

Dans [GGW20], qui est un proceedings pour la conférence ITWIST, nous étudions l'optimisation directe du problème d'optimisation d'échantillonnage donné par la Définition 6.3 sans aucune contrainte X_{ad}. Nous utilisons pour cela un algorithme de type BFGS à mémoire limitée et nous avons donc besoin du gradient par rapport aux paramètres. Ce gradient est calculé par rétropropagation. Dans la majorité des cas, le reconstructeur $\mathcal{R}$ est défini de manière implicite, par un problème de minimisation. il y a dans ce cas deux choix pour différentier le reconstructeur, soit on suppose que l'algorithme est arrivé au point minimum et on différentie l'équation d'Euler via un théorème des fonctions implicites, soit on garde en mémoire toutes les étapes de calcul et on les différentie par rapport aux paramètres. Dans [GGW20], nous appliquons cette deuxième technique. Un exemple d'application est donné en Figure 6.3d.



(a) Image originelle (b) BF, PSNR= 44.3dB (c) DV, PSNR = 44.5dB (d) Points optimisés, PSNR= 51.8dB

FIGURE 6.3 – Reconstruction de l'image originelle (Figure 6.3a) par plusieurs méthodes : Seules les basses fréquences sont conservées dans la Figure 6.3b, un échantillonnage uniforme des fréquences dans la Figure 6.3c, les points d'échantillonnages sont optimisés dans la Figure 6.3d. Le PSNR (Peak Signal to Noise Ratio) est un critère de qualité de la reconstruction, il doit être le plus élevé possible.

6.4 [GGWss] : On ne peut pas optimiser de schéma pour l'IRM

Dans toutes les références de la section précédente, les auteurs expriment (plus ou moins explicitement) des problèmes de convergence. Nous nous sommes attelés dans [GGWss] à l'étude des minimiseurs et nous montrons qu'il existe un nombre combinatoirement élevé de tels minimiseurs.

Cadre de travail de [GGWss] : Dans [GGWss], nous nous limitons à l'étude de la dimension 1 (les signaux sont $1d$) et à des reconstructeurs $\mathcal{R}$ linéaires. Les trois types de reconstructeurs que nous considérons sont, si $A[\Xi]^\star$ représente l'adjoint de $A[\Xi]$:

$$R(\Xi, y) = A[\Xi]^\star y \text{ ou } R(\Xi, y) = (A[\Xi]^\star A[\Xi] + \lambda Id)^{-1} y \text{ ou } R(\Xi, y) = A[\Xi]^\dagger y,$$

Où $A[\Xi]^\dagger$ est la pseudo-inverse de $A[\Xi]$. Ces trois reconstructeurs sont fréquemment utilisés, le premier est exact si les Ξ sont situés sur une grille (dans ce cas, $A[\Xi]$ est un opérateur de séries de Fourier) et s'appelle "méthode de rétro-projection". Le deuxième reconstructeur vient du problème de moindre carré avec régularisation de Tikhonov suivant :

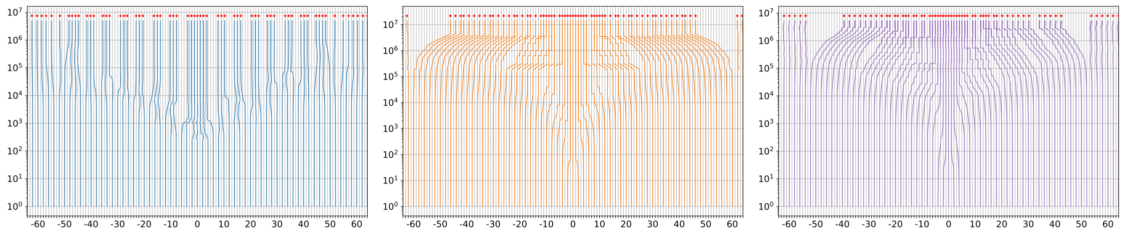
$$R(\Xi, y) = \arg\min_f \frac{1}{2}\|A(\Xi)^* f - y\|_2^2 + \frac{\lambda}{2}\|f\|_2^2. \tag{6.5}$$

Le troisième reconstructeur est la limite quand λ tend vers 0 du deuxième. Finalement, pour simplifier l'analyse mathématique, on se limite au où cas $\mathcal{U}$ ne contient qu'une image. Et où il n'y a aucune contrainte X_{ad} sur l'ensemble des Ξ admissibles.

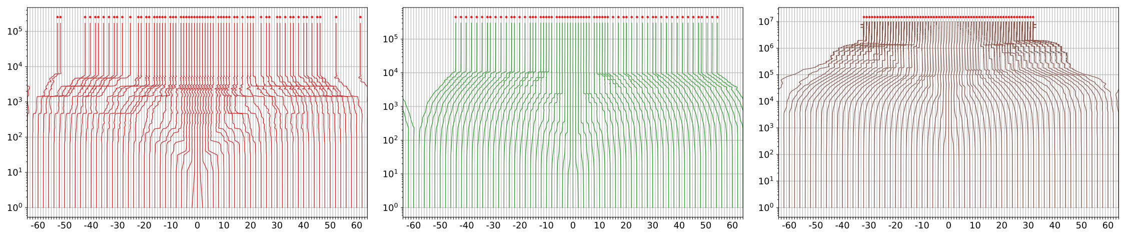
Nombre combinatoires de minimums locaux : Nous montrons que pour ces reconstructeurs, il existe un nombre combinatoires de minimums locaux du problème de la définition 6.3. Si le signal u est constant par pixel et composé de N pixels et que l'on procède à $M \simeq \sqrt{N}$ mesures, alors il existe des images pathologiques où il existe $M!e^{c\sqrt{N}}$ minimiseurs locaux. Le terme $M!$ est juste le facteur multiplicatif obtenu par permutation des points de mesures dans Ξ. Les grandes étapes de la démonstration sont résumées dans l'encadré ci-dessous :

Démonstration

Pour construire l'exemple, on se donne une fonction u dont la densité spectrale de Fourier, i.e. la fonction $\xi \mapsto |(Fu)(\xi)|^2$ possède un grand nombre, S, de maximiseurs locaux bien espacés. Les maximiseurs locaux sont situés sur une grille $\mathcal{G}$. Si les points d'échantillonnages sont choisis sur la grille, i.e. $\Xi \subset \mathcal{G}$, alors $A^\star[\Xi] = A^{-1}[\Xi]$. Comme les points de la grille $\mathcal{G}$, sont assez espacés, les interactions entre deux ξ différents dans Ξ sont faibles. Ainsi, on contrôle localement autour des points de la grille le comportement de la fonctionnelle à minimiser. On peut ainsi montrer que si on choisit M points parmi les S de la grille, il existe un minimum local de la fonctionnelle proche de ces M points. Le nombre de minimums locaux est donné par le nombre de combinaison de M points parmi S. On construit une fonction avec $S \simeq \sqrt{N}$, qui est le meilleur compromis entre un grand nombre de minimiseurs et des minimiseurs bien espacés.



(a) Descente de gradient, $P = 1$ (b) Descente de gradient, $P = 1000$ (c) Gradient stochastique, $P = 1000$



(d) L-BFGS, $P = 1000$ (e) Descente de gradient à métrique variable, $P = 1000$ (f) Gradient stochastique à métrique variable, $P = 1000$

FIGURE 6.4 – Trajectoires des Ξ (en dimension 1) lors de l'optimisation. Les itérations sont représentées sur l'axe vertical et l'axe horizontal représente la distribution des ξ. Le paramètre P correspond aux nombres d'images dans la banque de donnée $\mathcal{U}$. L'algorithme naïf de gradient avec une image est celui de la Figure 6.4a, notre algorithme recommandé est celui de la Figure 6.4f.

Nous montrons aussi que le gradient de la fonctionnelle s'écrase pour les hautes fréquences au sens où la norme de sa dérivée par rapport à une mesure ξ est bornée par $\|\xi\|^{-\alpha}$ si l'image u à reconstruire est $C^{1+\alpha}$. A cause de cette décroissance, les points d'échantillonnage situés aux hautes fréquences ne sont pas advectés par une méthode usuelle de gradient.

Proposition de remède : Nous proposons 3 remèdes à ces problèmes, premièrement, en augmentant la taille de $\mathcal{U}$, on montre que le nombre de minimums locaux diminue et que leur bassin d'attraction devient plus petit. Deuxièmement, on compense la décroissance de la norme du gradient pour les hautes fréquences en post-multipliant le gradient par un facteur, cela correspond à prendre une métrique variable dans l'optimisation . Troisièmement, on montre empiriquement que le gradient stochastique, qui a une tendance à tourner autour des minimums parvient à s'extraire des minimas locaux.

6.5 [GGW22] : Finalement on peut optimiser des schémas pour l'IRM

Dans [GGW22], nous proposons une approche bayésienne pour optimiser le problème de la définition 6.3. Parce que ce problème peut avoir un grand nombre de minima locaux comme montré dans [GGWss], les algorithmes de gradient peuvent avoir tendance à être inefficaces. Effectivement la direction donnée par le gradient sera probablement la direction du plus proche minimum local et ne peut avoir aucun intérêt pour de l'optimisation globale. Nous déployons donc une méthode d'optimisation bayésienne qui est une méthode d'ordre 0. Ces méthodes fonctionnent mieux en petite dimension, donc nous procédons à une réduction de dimension au préalable. Pour ce faire, nous changeons deux fois de variable d'optimisation. Le premier changement est de ne plus considérer les trajectoires Ξ comme variables d'optimisation mais la densité des points d'échantillonnage qui est une mesure de probabilité. Le deuxième changement consiste à paramétriser l'ensemble des mesures de densité par un ensemble convexe de faible dimension. En amont de l'algorithme d'optimisation nous utilisons deux algorithmes supplémentaires : un Générateur de densité et un échantillonneur. Les différentes composantes du programme sont détaillées ci-dessous

- **Générateur de densité :** On génère aléatoirement une famille de densités. On effectue une analyse en composantes principales de cette famille, on obtient $L+1$ vecteurs représentatifs $(\mu_\ell)_{\ell=0\ldots L}$. Le vecteur μ_0 étant de densité constante, tous les autres sont de moyenne nulle. On construit ensuite l'ensemble convexe $\mathcal{C} \subset \mathbb{R}^L$ de vecteurs $z = (z_1, \ldots, z_L)$ tels que la fonction $\mu_0 + \sum_\ell z_l \mu_l$ soit la densité d'une probabilité. Il suffit que cette fonction soit positive ici. Le générateur de densité est un algorithme qui prend un vecteur de $z \in \mathcal{C}$ et qui rend une densité de probabilité ρ.
- **Échantillonneur :** Cet algorithme est situé après le générateur de densité. Son rôle est de construire un schéma $\Xi \in X_{\text{ad}}$ à partir d'une densité ρ donnée par le générateur. Nous utilisons l'échantillonneur de [Boy+16] qui résout le problème de minimisation
$$\operatorname*{arg\,min}_{\Xi \in X_{\text{ad}}} \|h \star (\Xi - \rho)\|_{L^2}^2,$$
où h est un noyau de convolution/régularisation nécessaire car Ξ est une somme de mesures de Dirac et n'appartient pas à L^2.
- **Calcul de la fonction coût :** Etant donné un schéma Ξ, on simule une expérience IRM sur un échantillon de la banque d'image FastMRI, [Zbo+18] en utilisant une transformée de Fourier hors-grille [Shi+21], puis on lance un algorithme de reconstruction, soit un moindre carré avec terme de régularisation TV soit un réseau de neurones (un unrolled-ADMM avec le débruiteur de [Zha+21]). Finalement on calcule la fonction coût globale.
- **Un algorithme d'optimisation bayésienne :** Vu rapidement, ces algorithmes sont d'ordre 0 (pas de calcul du gradient), ils évaluent la fonction coût en des points donnés et l'"interpolent" [1] avec un *a priori*. Ensuite il suffit de minimiser une certaine fonction coût, on recalcule la fonction en ces nouveaux points candidats à être minimum et on reconstruit l'interpolant. La difficulté ici est que nous travaillons sur un convexe $\mathcal{C}$. Pour initialiser l'algorithme, il faut un algorithme d'échantillonnage de convexes, nous utilisons [DW22]. Ensuite, il nous faut re-écrire l'algorithme de minimisation de la fonction coût (ici EI :Expected Improvment avec un noyau de Matern) pour le transformer en algorithme de minimisation sous la contrainte $z \in \mathcal{C}$.

Il a fallu comprendre les différents choix qui s'offraient à nous dans la conception des algorithmes. Au final, la méthode d'optimisation bayésienne permet de gagner en temps calcul (on gagne un facteur 20 par rapport aux méthodes concurrentes) et peut être utilisée avec moins d'images dans la banque de données. Cependant il y a une perte de qualité qui est assez prononcée quand le reconstructeur est un réseau de neurones. Notre interprétation est que le générateur de densité ne permet pas d'atteindre la densité optimale, voir la Figure 6.5. Notre opinion est qu'une meilleure compréhension de la densité optimale dont a besoin le réseau de neurones est nécessaire pour améliorer la méthode.

6.6 Perspectives

Les différentes perspectives sur ces sujets de recherche sont nombreuses. Nous citerons ici

1. Le mot "interpolation" ne rend pas du tout hommage à l'optimisation bayésienne, mais je vais l'utiliser ici pour simplifier les chose.

(a) Optimisation bayésienne

(b) Optimisation jointe des trajectoires et des paramètres du réseau de neurones.

Figure 6.5 Méthode d'optimisation bayésienne contre les meilleures méthodes de l'état de l'art. La densité sous-jacente de la méthode de droite est complexe et n'est pas une image acceptable du générateur de densité.

Premièrement, les contre-exemples de [GGWss] qui montrent l'existence d'un grand nombre de minima locaux ne sont pas satisfaisants E ectivement, ils sont construits en montrant qu'il existe autant de minimiseurs locaux que de maxima de la power spectral density function (psdf, carré du module de la transformée de Fourier). Cependant les tests numériques de [GGWss ; GGW22] montrent que les oscillations sont sous-pixeliques. Il faut arriver à comprendre ce phénomène pour arriver à le contrer de manière e cace.
Deuxièmement les solutions apportées dans [GGW22] montrent vite leurs limites car la famille de densités à partir de laquelle on génère l'analyse en composantes principales est donnée *a priori* . C'est-à-dire que nous demandons à l'algorithme de choisir la meilleure densité par rapport à une famille donnée, cette famille n'est clairement pas adaptée dans le cas de réseau de neurones.
Les tests de [GGW22], ont été fait avec la méthode Sparkling [Laz+19] pour reconstruire une trajectoire depuis une densité. On peut aussi utiliser la méthode que nous avons développée [GKL19]. Le seul souci de cette méthode est que les algorithmes de di érentiation automatique ne sont pas disponibles.
Finalement, l'article [GG22] est une méthode de calcul automatique de pas pour les méthodes de direction de descente. Il serait intéressant de l'étendre aux méthodes à inertie à la Polyak ou aux méthodes accélérées à la Nesterov.

Bibliographie

[AJ20] Hemant Kumar Aggarwal et Mathews Jacob . J-MoDL : Joint model-based deep learning for optimized sampling and reconstruction . In : IEEE journal of selected topics in signal processing 14.6 (2020), p. 1151-1162.

[BDS19] Cagla Deniz Bahadir , Adrian V Dalca et Mert R Sabuncu . Learning-based optimization of the under-sampling pattern in MRI . In : International Conference on Information Processing in Medical Imaging. Springer. 2019, p. 780-792.