

EDPs numériques pour l'analyse d'images

Jean-Marie Mirebeau*

February 10, 2021

Abstract

L'objectif du cours est de présenter des outils et méthodes permettant la résolution numérique efficace d'équations aux dérivées partielles, en vue d'applications au traitement de l'image.

Contents

1	Introduction	2
1.1	Pertinence des EDPs en traitement d'image	2
1.2	Les EDPs considérées dans ce cours	2
2	Définition et théorie	5
2.1	Rappels: différentielle et gradient	6
2.2	Préliminaire: coût d'un produit matriciel	7
2.3	Trois approches de la différentiation automatique	8
2.4	Dérivées d'ordre deux et supérieur	10
3	Discrétisation de la diffusion anisotrope et non-linéaire	12
3.1	Flot gradient dans un espace abstrait	12
3.2	L'équation de la chaleur en tant que flot gradient	14
3.3	Schéma aux différences finies	16
3.4	Décomposition matricielle de Selling	18
3.5	Non-linéarité	19
4	Chaleur et distances géodésiques	21
4.1	Distance riemannienne	22
4.2	Transformation logarithmique	25
4.3	Solution de viscosité	26
4.4	Transport optimal entropique	29
5	Schémas monotones	31
5.1	Principe de comparaison, existence d'une solution	32
5.2	Exemples	33
5.3	Itérations d'Euler et de Newton	35
5.4	Itérations de Jacobi	37

*Université Paris-Saclay, ENS Paris-Saclay, CNRS, Centre Borelli, F-91190 Gif-sur-Yvette, France.

1 Introduction

Les équations aux dérivées partielles (EDPs) sont l'un des formalismes mathématiques permettant de passer du local - en décrivant un comportement à l'échelle infinitésimale - au global - par la résolution d'un système d'équations couplées définies en chaque point d'un domaine. Elles sont incontournables dans certains domaines, comme la physique des milieux continus. Nous donnons dans cette section leur pertinence dans le cadre du traitement de l'image, et un aperçu des exemples et méthodes qui seront traités dans le cours.

1.1 Pertinence des EDPs en traitement d'image

De nombreuses approches mathématiques sont pertinentes dans le traitement de l'image et du signal, comme les statistiques et probabilités, l'optimisation et l'apprentissage, ou encore différentes branches de l'analyse. Voici certaines des raisons qui justifient l'intérêt des EDPs dans ce contexte.

- *Physique des dispositifs d'acquisition.* De nombreux dispositifs d'imagerie sont fondés sur la physique des milieux continus (vibrations, absorption lumineuse, ...) qui est décrite de manière directe par des EDPs. En particulier les méthodes de *tomographie*, qui consistent à "reconstruire volume d'un objet à partir d'une série de mesures effectuées depuis l'extérieur de cet objet" (Wikipedia), utilisées en imagerie médicale ou sismique.
Exemple : La transformée de Radon, correspondant à la "tomographie axiale".
Exemple : La reconstruction de relief à partir des ombres (Shape from shading).
- *Modèle interprétable et explicatif.* Par analogie avec des modèles physiques, les EDPs permettent de définir un cadre interprétable et explicatif pour le traitement de données.
Exemple : Description multi-échelle d'une image, via un processus de diffusion.
Exemple : Les courbes elastica de Euler, qui correspondent à la position de repos d'une barre élastique, sont utilisées pour l'extraction de contours et de courbes dans des images.
- *Reconstruction d'information globale.* On peut considérer une EDP comme un système d'équations couplées, où chaque point du domaine porte (typiquement) une inconnue et une équation. La résolution d'un tel système est un objet global, qui synthétise les contraintes imposées localement.
Exemple : La résolution de l'équation eikionale permet de déterminer le plus court chemin dans un domaine contenant des obstacles et des zones plus ou moins rapides.
- *Structure mathématique, garanties.* Les EDPs ont une structure mathématique riche qui permet d'établir un certain nombre de garanties : terminaison de l'algorithme, robustesse au bruit, validité sur des cas simples, etc

1.2 Les EDPs considérées dans ce cours

Dans ce cours nous étudierons les schémas de discréttisation, la mise en oeuvre numérique, et les applications, de principalement deux équations (avec leurs variantes comme des modifications anisotropes ou non-linéaires, des cas limites, etc). Ce sont l'équation de la chaleur, et l'équation eikionale, qui sous leur forme la plus simple s'écrivent respectivement

$$\partial_t u = \Delta u, \quad \|\nabla u\| = 1, \quad (1)$$

avec des conditions au bord appropriées.

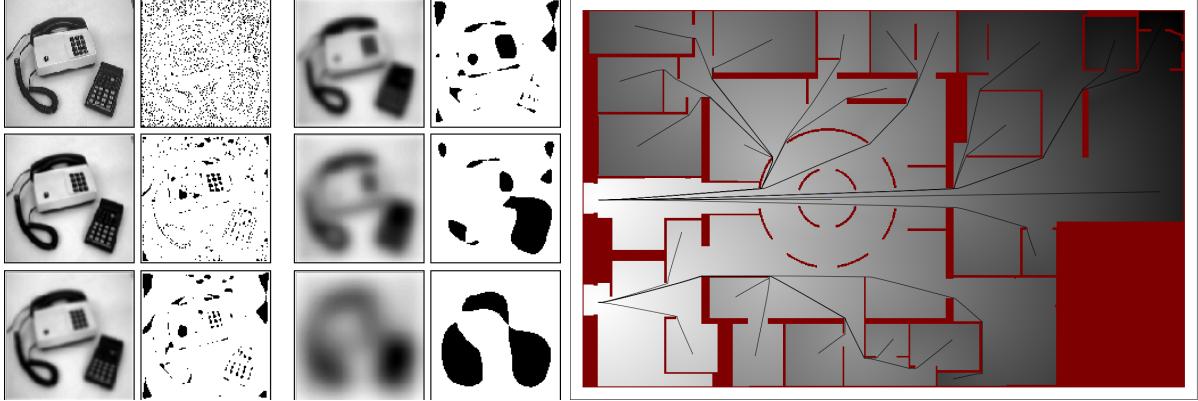


Figure 1: (Gauche) Equation de la chaleur et représentation multi-échelle d'une image. Crédit image: Tony Lindeberg. (Droite) Solution de l'équation eikonale (niveaux de grid) dans un domaine avec obstacles (rouge), et chemins minimaux.

- L'équation de la chaleur s'interprète comme le flot gradient (descente de gradient) de l'énergie de Dirichlet $\int_{\Omega} \|\nabla u\|^2$. Elle lisse les changements brusques dans la fonction u ; dans le cas d'une image elle élimine le bruit, mais rend flous les contours d'objets et les textures oscillantes. L'équation de la chaleur permet de séparer, au fil du temps, les différentes échelles d'une image.
- L'équation eikonale caractérise la fonction distance euclidienne u à un point ou à une région source. Le gradient de sa solution $\nabla u(x)$ donne la direction opposée à la source. L'équation eikonale est le modèle le plus simple de propagation de front, et remonter son gradient permet de calculer des plus courts chemins.

L'équation de la chaleur et l'équation eikonale standard (1) traitent toutes les directions d'espace de manière identique: on dit qu'elles sont *isotropes*. Plus précisément, elles sont définies en chaque point par des opérateurs différentiels Δu et $\|\nabla u\|$, qui sont invariants par rotation. Cette propriété de symétrie permet de discréétiser les opérateurs y intervenant de manière particulièrement simple: pour toute fonction u assez lisse, en notant $(e_i)_{1 \leq i \leq d}$ la base canonique de \mathbb{R}^d , on obtient

$$\Delta u(x) = \sum_{1 \leq i \leq d} \frac{u(x + he_i) - 2u(x) + u(x - he_i)}{h^2} + \mathcal{O}(h^2), \quad (2)$$

$$\|\nabla u(x)\|^2 = \sum_{1 \leq i \leq d} \frac{\max\{0, u(x) - u(x - he_i), u(x) - u(x + he_i)\}^2}{h^2} + \mathcal{O}(h). \quad (3)$$

Le choix des différences finies utilisées a son importance, et sera justifié plus loin dans le cours.

Anisotropie. On dit qu'un problème est anisotrope lorsque les directions d'espace dans un domaine ne sont pas interchangeables. C'est un phénomène générique, dont les causes sont variées:

- *Micro-structure.* Certains milieux physiques sont micro-structurés, ce qui affecte les ondes s'y propageant de manière anisotrope, notamment lorsque leur longueur d'onde excède l'échelle de la micro-structure, par un effet appelé homogénéisation. Ondes sismiques dans



Figure 2: Exemples d'anisotropies associées à une micro-structure. (Gauche) Empreinte digitale, (droite) minéral mica. Images wikipedia.

un milieu stratifié, ondes lumineuses dans un cristal, ... En traitement de l'image, il est aussi courant de détecter les directions préférentielles de textures oscillantes (comme dans une image d'empreinte digitale) pour les traiter de manière appropriée.

- *Rôle distinct des dimensions du domaine.* Certains domaines mathématiques ne correspondent pas à un milieu physique, mais à un espace d'états dont les dimensions jouent des rôles différents, ce qui crée des structures anisotropes. Par exemple l'espace $\mathbb{R}^2 \times \mathbb{S}^1$ des positions et orientations, qui correspond aux configurations d'un véhicule simple, possède une structure dite sous-riemannienne, permettant en (x_0, x_1, θ) seulement les déplacements engendrés par $(\cos \theta, \sin \theta, 0)$ et $(0, 0, 1)$.
- *Proximité des bords ou de fissures.* La modélisation des écoulements au bord d'un domaine, ou de l'élasticité au voisinage d'une fissure, se fait souvent par l'intermédiaire de modèles réduits traitant de manière spécifique la dimension tangente.
- *Paramétrisation d'un domaine complexe.* La résolution de d'EDPs dans des géométries complexes, surfaces ou volumes, peut se faire par l'intermédiaire de leur paramétrisation par des domaines de référence (rectangles). La paramétrisation induit alors en général une distortion anisotrope de l'EDP sur le domaine de référence, même si elle était isotrope sur la géométrie initiale.
- *Linéarisation d'un problème.* Certaines EDPs linéaires sont obtenues par perturbation d'une EDP non-linéaire au voisinage d'une solution. Selon les cas, l'équation linéarisée peut être anisotrope même si l'équation non-linéaire était isotrope. Par exemple, dans le cas de Monge-Ampère, on a formellement $\det(\nabla^2(u+h)) - \det(\nabla^2 u) = \text{Tr}(D(u)\nabla^2 h) + \mathcal{O}(h^2)$ où $D(u)$ est la comatrice de $\nabla^2 u$.

Formellement, l'introduction d'anisotropie dans une EDP se fait par l'introduction de champs qui décrivent la géométrie locale du problème. Par exemple si D est un champ de matrices symétriques définies positives, alors on peut considérer les variantes anisotropes de l'équation de la chaleur et de l'équation eikonal définies par

$$\partial_t u = \operatorname{div}(D \nabla u), \quad \langle \nabla u, D \nabla u \rangle = 1, \quad (4)$$

avec de nouveau des conditions au bord appropriées. Ces variantes favorisent la diffusion (chaleur) ou la propagation du front (eikonal) dans les directions associées aux grandes valeurs propres de D .

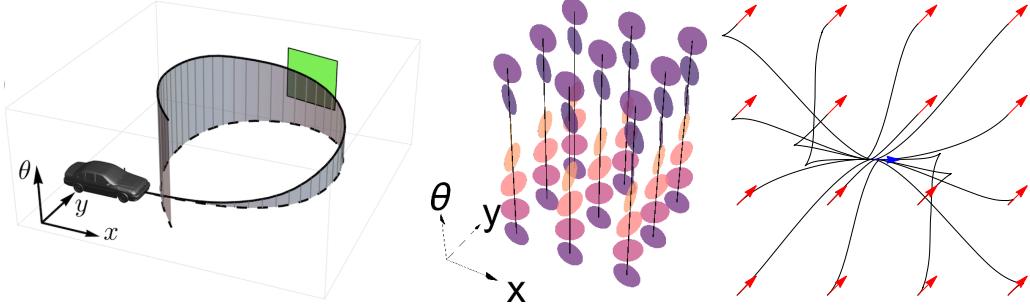


Figure 3: Exemple d'anisotropie liée au rôle distinct des dimensions du domaine. Ici le véhicule de Reeds-Shepp, dont l'espace d'états est $\mathbb{R}^2 \times \mathbb{S}^1$. (Gauche) Une trajectoire admissible. (Centre) Boule unité dans l'espace des vitesses admissibles. (Droite) Projection planaire de quelques trajectoires optimales.

2 Différentiation numérique

L'analyse numérique et l'optimisation requièrent de calculer numériquement les dérivées de fonctions. A cet effet, plusieurs méthodes existent: *différences finies*, *dé différentiation automatique*, ou simplement la différentiation symbolique suivie de l'implémentation des formules résultantes. La discussion présentée ci-dessous permet de faire un choix éclairé, et est suivie d'une description plus détaillée de la différentiation automatique.

Différences finies. On appelle différences finies les combinaisons (en général linéaires) de valeurs ponctuelles d'une fonction, destinées à approcher ses dérivées.

Par exemple, si $f : \mathbb{R} \rightarrow \mathbb{R}$ est lisse, si $x \in \mathbb{R}$ et h est petit, alors on peut considérer les différences finies *upwind*, *centrée*, et d'*ordre deux* définies par

$$\begin{aligned}\frac{f(x+h) - f(x)}{h} &= f'(x) + \mathcal{O}(h), \\ \frac{f(x+h) - f(x-h)}{2h} &= f'(x) + \mathcal{O}(h^2), \\ \frac{f(x+h) - 2f(x) + f(x-h)}{h^2} &= f''(x) + \mathcal{O}(h^2)\end{aligned}$$

Les différences finies se caractérisent par l'ordre de la dérivée qu'elles approchent (ici 1,1,2), et leur ordre de précision (ici 1,2,2). Une autre approche conceptuellement proche, pour approcher les dérivées d'une fonction définie ponctuellement, consiste à interpoler cette fonction (par éléments finis, splines, etc), puis à différentier cette interpolation.

Intérêts: Les différences finies permettent d'approcher les dérivées de fonctions qui ne sont définies que *ponctuellement*. Avec les éléments finis et autres méthodes conceptuellement proches, elles sont à la fondation des schémas numériques pour la discréétisation des équations aux dérivées partielles. Leur structure mathématique simple et (en général) linéaire permet d'étudier mathématiquement les quantités qui en sont dérivées.

Inconvénients: L'approximation des dérivées d'une fonction par différences finies, et autres méthodes proches, nécessite de faire des compromis entre la *précision* et la *stabilité*. (Les schémas d'ordre élevé sont généralement instables.)

Dé différentiation automatique. On appelle différentiation automatique les procédés informatiques permettant de calculer les dérivées de fonctions définies par des programmes, en utilisant

les dérivées exactes des fonctions usuelles (\exp , $\sqrt{\cdot}$, \sin , etc) et en appliquant les règles de composition des différentielles.

Intérêts: La différentiation automatique permet de calculer les dérivées de fonctions complexes de manière, en général, stable et précise. Grâce à des techniques comme la surcharge des opérateurs et des fonctions usuelles, son utilisation n'altère que peu la lisibilité des programmes. Son coût numérique est souvent raisonnable.

Inconvénients: Il existe au moins trois variétés de différentiation automatique: dense, creuse, et par rétro-propagation; et elles peuvent être composées entre elles. La mise en oeuvre doit être réfléchie en fonction des conditions d'utilisation, sous peine de coût de calcul excessif.

Differentiation formelle et implémentation. Lorsqu'une fonction apparaissant dans un programme informatique possède une expression simple, il peut sembler naturel de calculer ses dérivées en la dérivant formellement et en implémentant l'expression mathématique résultante.

Intérêt: Cette approche mène, parfois, à l'implémentation la plus efficace en termes de temps de calcul.

Inconvénient: Cette méthode est **à proscrire¹ en première approche**, en faveur de la différentiation automatique. En effet, elle rend le programme peu lisible, peu flexible, long à écrire, et introduit de nombreux bugs. Lorsqu'elle nécessaire, ses résultats doivent à minima être contrôlés sur des exemples par une approche alternative².

2.1 Rappels: différentielle et gradient

On rappelle pour fixer les conventions les définitions élémentaires des différentielle, gradient, hessienne d'une fonction. Dans cette sous-section, les lettres E et F désignent des espaces vectoriels normés (evn), et \mathbb{H} un espace de Hilbert.

Définition 2.1. Soient E, F evn. On note $\mathcal{L}(E, F)$ l'ensemble des applications linéaires continues de E dans F , qui est aussi un evn.

Définition 2.2 (Différentielle). Soient E, F evn. Une fonction $f : \Omega \rightarrow F$, où $\Omega \subseteq E$ est ouvert, est différentiable en $x \in \Omega$ s'il existe $L \in \mathcal{L}(E, F)$ telle que

$$f(x + h) = f(x) + L(h) + o(\|h\|)$$

On note $L = df|_x$, appelée différentielle de f en x .

Définition 2.3 (Gradient). Soit $f : \Omega \rightarrow \mathbb{R}$, où Ω est un ouvert d'un espace de Hilbert \mathbb{H} , différentiable en $x \in \Omega$. Le gradient de f en x , noté $\nabla f(x) \in \mathbb{H}$, est défini par l'identité

$$\langle \nabla f(x), v \rangle = df|_x(v)$$

¹Rappelons que les qualités à rechercher lors de la conception d'un programme sont, dans l'ordre:

1. La *lisibilité*, qui permet à d'autres programmeurs (voire vous-même) de continuer votre travail, et leur donne confiance quant à sa qualité.
2. La *robustesse*, c'est à dire les garanties que l'on peut apporter concernant l'exécution du programme, et sa gestion des erreurs.
3. La *généricité*, qui permet au programme d'être utilisé au sein d'applications non considérées initialement.
4. La *rapidité de conception*, par le choix des bons outils, car le temps humain a plus de valeur que le temps machine.
5. La *rapidité d'exécution*. La recherche de cette qualité ne doit pas se faire au détriment des précédentes.

²Citation appropriée : *If it's not tested, it's broken.* Bruce Eckel

pour tout $v \in H$. En d'autres termes $f(x + h) = f(x) + \langle \nabla f(x), h \rangle + o(h)$.

Exemple : soit H un Hilbert, et $f : H \rightarrow \mathbb{R}$ définie par $f(x) = \frac{1}{2}\|x\|^2$. On note que $f(x + h) = \frac{1}{2}\|x\|^2 + \langle x, h \rangle + \frac{1}{2}\|h\|^2$. On en déduit que $\nabla f(x) = x$, pour tout $x \in H$.

Définition 2.4 (Matrice jacobienne). Soit $f : \Omega \rightarrow \mathbb{R}^n$, où $\Omega \subseteq \mathbb{R}^m$, différentiable en $x \in \Omega$. On appelle matrice jacobienne de f en x , notée $Df|_x$, la matrice de $df|_x$ dans les bases canoniques de \mathbb{R}^m et \mathbb{R}^n . Ses composantes sont appelées dérivées partielles de f

$$Df|_x = \left(\frac{\partial f_i}{\partial x_j}(x) \right)_{\substack{1 \leq i \leq n \\ 1 \leq j \leq m}}$$

En l'absence d'ambiguité (choix de base), on ne se gènera pas pour identifier $df|_x$ et $Df|_x$.

Exemple: si $f : \mathbb{R}^m \rightarrow \mathbb{R}$, alors $\nabla f(x) = (Df|_x)^T$

Lemme 2.5 (Composition). Soient $f : E \rightarrow F$ et $g : F \rightarrow G$. Supposons f différentiable en $x \in E$, et g différentiable en $f(x) \in F$. Alors $g \circ f : E \rightarrow G$ est différentiable en x et

$$d(g \circ f)|_x = dg|_{f(x)} \circ df|_x.$$

La différentielle d'une composée est donc la composée des différentielles; de même pour les matrices jacobien, sous des hypothèses adéquates, $D(g \circ f)|_x = Dg|_{f(x)} Df|_x$.

2.2 Préliminaire: coût d'un produit matriciel

Étant données deux matrices, A de taille $I \times J$ et B de taille $J \times K$, leur produit AB de taille JK est défini par

$$(AB)_{ik} = \sum_{1 \leq j \leq J} A_{ij} B_{jk} \quad (5)$$

pour tous $1 \leq i \leq I$, $1 \leq k \leq K$. Le coût du calcul de AB est par la méthode "naive" est donc

$$\mathcal{O}(IJK).$$

Considérons A_1, \dots, A_n des matrices de taille $I_0 \times I_1, \dots, I_{n-1} \times I_n$. Leur produit est associatif, et peut donc être parenthésé de manière arbitraire

$$((A_1 A_2) A_3) \cdots A_n = A_1 \cdots A_n = A_1 (A_2 (\cdots A_n)). \quad (6)$$

Le coût de l'évaluation de l'expression parenthésée à gauche ou à droite est respectivement

$$\mathcal{O}(I_0(I_1 I_2 + I_2 I_3 + \cdots + I_{n-1} I_n)), \quad \mathcal{O}((I_0 I_1 + \cdots + I_{n-2} I_{n-1}) I_n). \quad (7)$$

Ces coûts ne sont en général pas égaux, bien qu'ils produisent le même résultat final. En particulier, si $I_0 = 1$, c'est à dire si la première matrice est un vecteur ligne, alors le parenthésage à gauche est optimal car son coût (7, gauche) correspond au coût de la lecture des données. De même manière, si $I_n = 1$, c'est à dire si la dernière matrice est un vecteur colonne, alors le parenthésage à droite est optimal.

Dans le cas général, trouver le parenthésage optimal pour minimiser le coût de calcul est un problème NP-complet.

Cas de matrices creuses Une matrice creuse possède peu de coefficients non-nuls, qui peuvent être stockés dans une structure de données adaptée, permettant une manipulation numérique efficace. Les matrices de schémas numériques pour la discréétisation des EDPs sont fréquemment de cette forme. La propriété d'être creux est partiellement compatible avec le produit matriciel, comme le montre le résultat suivant.

Lemme 2.6. *Soit A (resp. B) une matrice de taille $I \times J$ (resp. $J \times K$) dont chaque ligne au plus α (resp. β) coefficients non-nuls. Alors chaque ligne de AB contient au plus $\alpha\beta$ coefficients non-nuls.*

Proof. Soit \mathcal{A} (resp. \mathcal{B}) la matrice obtenue en remplaçant les coefficients non-nuls de A (resp. B) par 1. Alors $\mathcal{A}\mathcal{B}$ est une matrice dont les coefficients sont entiers, positifs, et non-nuls à chaque position où AB a un coefficient non-nul. Par ailleurs la somme des coefficients de $\mathcal{A}\mathcal{B}$ sur la ligne d'indice i , où $1 \leq i \leq I$, vaut

$$\sum_{1 \leq k \leq K} (\mathcal{A}\mathcal{B})_{ik} = \sum_{1 \leq j \leq J} \left(\mathcal{A}_{ij} \sum_{1 \leq k \leq K} \mathcal{B}_{jk} \right) \leq \sum_{1 \leq j \leq J} \mathcal{A}_{ij} \beta \leq \alpha \beta.$$

Le résultat annoncé s'ensuit. \square

Par transposition, on obtient un résultat analogue au Lemme 2.6 pour les colonnes. Par une récurrence immédiate, si A_1, \dots, A_n sont des matrices dont chaque ligne contient au plus $\alpha_1, \dots, \alpha_n$ coefficients non-nuls, alors leur produit $A_1 \cdots A_n$ est contient au plus $\alpha_1 \cdots \alpha_n$ coefficients non-nuls sur chaque ligne. Donc au total pas plus de

$$I_0 \alpha_1 \cdots \alpha_n \tag{8}$$

coefficients non-nuls, en notant I_0 le nombre de lignes de A_0 . Avec une implémentation raisonnable du calcul du produit de matrices de creuses, la quantité (8) borne aussi le coût du calcul de $A_1 \cdots A_n$, indépendamment de l'ordre d'associativité utilisé.

Cette estimation montre avant tout que l'utilisation des produits creux doit être réservée à un nombre très faible de facteurs très creux, car le taux de remplissage augmente de manière exponentielle au fil des produits.

Interprétation en tant que Jacobienne. Soient $f_1 : \mathbb{R}^{I_1} \rightarrow \mathbb{R}^{I_0}, \dots, f_n : \mathbb{R}^{I_n} \rightarrow \mathbb{R}^{I_{n-1}}$ des fonctions, et soit $x_n \in \mathbb{R}^{I_n}$. Supposons f_i différentiable au point x_i , où $x_{i-1} := f_i(x_i)$ pour tout $1 \leq i \leq n$. Alors

$$D(f_1 \circ \cdots \circ f_n)|_{x_n} = Df_1|_{x_1} \cdots Df_n|_{x_n}.$$

Les programmes informatiques décrivent des fonctions complexes comme composées de fonctions élémentaires. Leur différentiation automatique peut donc s'interpréter³ comme le produit matriciel des matrices Jacobiennes des étapes intermédiaires. Selon la structure de ces facteurs, on préférera utiliser un produit associatif à gauche, ou à droite, ou un produit creux; voire une combinaison de ces approches suivant les parties du facteur.

2.3 Trois approches de la différentiation automatique

On peut distinguer trois approches principales de la différentiation automatique: *dense*, *creuse*, et par *rétro-propagation*. Elles correspondent conceptuellement aux trois stratégies détaillées

³Même si elle n'est pas implémentée de cette manière en apparence, voir la sous-section suivante.

§2.2 pour le calcul d'un produit matriciel: associativité à droite, à gauche, ou produit creux, ce qui permet d'anticiper leurs forces et leurs faiblesses respectives. Cependant leur implémentation numérique s'éloigne de ce cadre, car pour plus de commodité et d'efficacité les matrices Jacobiennes des étapes intermédiaires ne sont en général pas construites explicitement.

Dans la suite, on suppose que l'on cherche à calculer la matrice jacobienne d'une fonction

$$f : \mathbb{R}^m \rightarrow \mathbb{R}^n, \quad (9)$$

définie via un programme informatique, donc comme composition de fonctions élémentaires qui ne seront pas explicitées ici. (Pour l'analogie avec le produit matriciel (7) des jacobienes, on note que (m, n) correspondent à (I_n, I_0) .)

Différentiation automatique dense (m petit).

Analogue matriciel. La différentiation automatique dense correspond au produit matriciel (6) par associativité à droite; on parle aussi de propagation *forward* puisque c'est le sens naturel de l'exécution du programme. Elle est particulièrement efficace lorsque le nombre d'entrées m de la fonction f est petit, par analogie avec (7, droite).

Utilisations. La différentiation automatique dense est idéale pour différentier des fonctions utilitaires en basse dimension (m, n petits). Elle convient aussi sur le principe pour analyser la dépendance d'une sortie en grande dimension ($n \gg 1$), par exemple la simulation numérique d'un problème physique, par rapport à un petit nombre m de paramètres.

Implémentation. La différentiation automatique dense s'implémente en replaçant les scalaires en entrée du programme définissant f par des paires $(x, v) \in \mathbb{R} \times \mathbb{R}^m$ représentant le développement limité

$$x + \langle v, h \rangle + o(h).$$

La variable $h \in \mathbb{R}^m$ représente ici une perturbation *symbolique*, qui n'a pas d'existence dans le programme informatique. La surcharge des opérateurs et des fonctions usuelles permet d'appliquer les règles de calcul des développements limités sans avoir à ré-écrire la fonction. Par exemple

$$\begin{aligned} \sin(x + \langle v, h \rangle + o(h)) &= \sin(x) + \cos(x)\langle v, h \rangle + o(\|h\|) \\ (x + \langle v, h \rangle + o(h))(x' + \langle v', h \rangle + o(h)) &= xx' + \langle xv' + x'v, h \rangle + o(h). \end{aligned}$$

Différentiation automatique creuse (f simple).

Analogue matriciel. La différentiation automatique creuse correspond au produit matriciel creux (et par associativité à droite, pour la simplicité). Elle est particulièrement efficace lorsque f a une structure très simple, faisant intervenir peu d'étapes internes, chacune ne dépendant que de quelques variables. Sous ces conditions, elle permet de traiter des entrées et sorties de grande taille.

Utilisations. La différentiation automatique creuse est idéale pour assembler les matrices Jacobiennes schémas numériques pour des équations aux dérivées partielles.

Implémentation. La différentiation automatique creuse s'implémente remplaçant les scalaires en entrée du programme définissant f par des paires $(x, \alpha, i) \in \mathbb{R} \times \mathbb{R}^K \times \{1, \dots, m\}^K$ représentant le développement limité

$$x + \sum_{1 \leq k \leq K} \alpha_k h_{i_k} + o(\|h\|).$$

De nouveau, la variable $h \in \mathbb{R}^m$ représente une perturbation *symbolique*, qui n'a pas d'existence dans le programme informatique. La surcharge des opérateurs et des fonctions usuelles permet

d'appliquer les règles de calcul des développements limités. Par exemple

$$\begin{aligned} \sin\left(x + \sum_{1 \leq k \leq K} \alpha_k h_{i_k} + o(\|h\|)\right) &= \sin(x) + \sum_{1 \leq k \leq K} \cos(x) \alpha_k h_{i_k} + o(\|h\|) \\ \left(x + \sum_{1 \leq k \leq K} \alpha_k h_{i_k} + o(\|h\|)\right) + \left(x' + \sum_{1 \leq k \leq K} \alpha'_k h_{i'_k} + o(\|h\|)\right) \\ &= x + x' + \left(\sum_{1 \leq k \leq K} \alpha_k h_{i_k} + \sum_{1 \leq k \leq K} \alpha'_k h_{i'_k} \right) + o(\|h\|). \end{aligned}$$

En particulier, la somme est représentée par $(x + x', \alpha \oplus \alpha', i \oplus i')$, où l'opérateur \oplus désigne la concaténation de vecteurs. (Cette représentation pourra éventuellement être simplifiée en regroupant les poids associés à des indices redondants.)

Par rétro-propagation (n petit).

Analogue matriciel. La différentiation automatique par rétro-propagation correspond au produit matriciel par associativité à gauche. Pour cette raison, elle est particulièrement efficace lorsque la sortie est de petite dimension, sans limite sur la taille de l'entrée.

Utilisations. La différentiation automatique par rétro-propagation est particulièrement utile pour les problèmes d'optimisation, car la sortie est alors de dimension 1. En particulier, elle est systématiquement utilisée pour l'entraînement des réseaux de neurones.

Implémentation. La différentiation automatique par rétro-propagation nécessite de rejouer les calculs dans l'ordre inverse de leur exécution initiale, et donc de les organiser dans un historique ou un graphe orienté. Pour cette raison elle est plus complexe à mettre en oeuvre que la différentiation dense ou creuse. Par ailleurs, la conservation de la totalité des états intermédiaires des variables a un coût mémoire potentiellement important, qui peut être réduit par des re-calculs partiels, ce qui mène à des compromis et astuces d'implémentation non-triviaux. Des libraries comme PyTorch implémentent ces techniques.

2.4 Dérivées d'ordre deux et supérieur

Rappels : définitions et propriétés élémentaires.

Définition 2.7 (Différentielle d'ordre supérieur). *Soient E, F des evn, et soit $f : \Omega \rightarrow F$ où $\Omega \subseteq E$ ouvert, différentiable en tout point de Ω . Si $df : \Omega \rightarrow \mathcal{L}(E, F)$ est différentiable au point $x \in \Omega$, alors on note $d^2 f|_x \in \mathcal{L}(E, \mathcal{L}(E, F))$ sa différentielle.*

On note que $\mathcal{L}(E_0, L(E_1, F))$ s'identifie à $\mathcal{L}^2(E_0 \times E_1, F)$, espace vectoriel des applications bilinéaires continues de $E_0 \times E_1$ dans F .

Théorème 2.8 (Schwartz). *Soit $f : \Omega \rightarrow F$, où $\Omega \subseteq E$ est ouvert, telle que $d^2 f : \Omega \rightarrow \mathcal{L}^2(E \times E, F)$ existe et est continue. Alors $d^2 f|_x \in \mathcal{L}^2(E \times E, F)$ est une forme bilinéaire symétrique, pour tout $x \in \Omega$.*

Sous les hypothèses du Théorème de Schwartz, on a le développement limité suivant, qui permet aussi de caractériser la différentielle seconde par identification

$$f(x + h) = f(x) + df|_x(h) + \frac{1}{2}d^2 f|_x(h, h) + o(\|h\|^2).$$

Définition 2.9 (Matrice hessienne). Soit $f : \mathbb{R}^n \rightarrow \mathbb{R}$ deux fois continument différentiable au voisinage de $x \in \mathbb{R}^n$. La matrice de la forme quadratique $d^2 f|_x$ dans la base canonique est appelée matrice hessienne, et noté $D^2 f|_x$.

Exemple: la Hessienne de $f : \mathbb{R}^n \rightarrow \mathbb{R}$ définie par $f(x) = \frac{1}{2}\|x\|^2$, est la matrice identité.

Différentiation automatique.

La différentiation automatique d'ordre deux, ou éventuellement supérieur, se justifie dans des cas particuliers dont voici quelques exemples:

- *Résolution de problèmes d'optimisation par la méthode de Newton.* Pour les problèmes d'optimisation ayant de bonnes propriétés mathématiques, dans l'idéal convexes, réguliers et non-contraints, l'efficacité de la méthode de Newton est souvent sans égal. Sa mise en oeuvre requiert d'évaluer les dérivées de la fonction à minimiser jusqu'à l'ordre deux.
- *Differentiation d'un programme faisant lui-même intervenir la différentiation automatique.* Exemple : trouver la géodésique joignant deux points donnés, par une méthode de tir. Les géodésiques sont des courbes obéissant aux équations de Hamilton, qui jouent un rôle fondamental en physique et en mathématiques

$$\partial_t q = \partial_p H, \quad \partial_t p = -\partial_q H.$$

Il est naturel d'utiliser la différentiation automatique pour dériver le Hamiltonien H , qui encode la géométrie du problème, et implémenter ces équations. Dans le cadre des méthodes de tir, on ajuste le moment initial p_0 (la position initiale q_0 étant fixée), pour atteindre une position finale $q(1)$ donnée; on peut utiliser pour cela méthode de Newton, ce qui requiert une seconde différentiation automatique.

- *Discrétisation de problèmes variationnels.* Certaines équations aux dérivées partielles sont présentées sous forme variationnelle: par exemple trouver $u \in H^1(\Omega)$ (espace de Sobolev sur un domaine Ω) tel que pour tout $v \in H^1(\Omega)$ on ait

$$\int_{\Omega} \nabla u \nabla v = \int_{\Omega} f v.$$

La construction automatique du schéma numérique, à partir d'une fonction implémentant une approximation numérique de ces intégrales (par différences finies, éléments finis, etc), requiert la différentiation d'ordre deux.

Les trois approches de la différentiation automatique, dense, creuse, et par rétro-propagation, s'étendent à l'ordre deux. Dans les deux premiers cas, il s'agit de remplacer les scalaires par des variables (x, v, m) ou $(x, \alpha, i, \beta, j, k)$ représentant des développements limités d'ordre deux

$$x + \langle v, h \rangle + \frac{1}{2} \langle h, mh \rangle + o(h^2), \quad x + \sum_{1 \leq r \leq R} \alpha_r h_{i_r} + \frac{1}{2} \sum_{1 \leq s \leq S} \beta_s h_{j_s} h_{k_s} + o(h^2),$$

et de surcharger les opérateurs et fonctions usuelles. La différentiation automatique par rétro-propagation à l'ordre deux, bien que possible, semble peu usitée.

3 Discrétisation de la diffusion anisotrope et non-linéaire

Dans cette leçon, on s'intéresse à l'équation de la chaleur anisotrope:

$$\partial_t u = \operatorname{div}(D \nabla u), \quad (10)$$

sur un domaine $\Omega \subseteq \mathbb{R}^d$, muni d'un champ de matrices symétriques définies positives $D : \Omega \rightarrow S_d^{++}$. On utilise des conditions au bord de Neumann sur $\partial\Omega$. On verra son interprétation en tant que flot gradient, sa discrétisation par différences finies *non-négatives*, ses variantes non-linéaires, et certaines de ses applications historiques⁴ en traitement d'image.

3.1 Flot gradient dans un espace abstrait

Les *flots gradients*, sont des analogues en temps continu de l'algorithme de descente de gradient, qui jouent un rôle important en analyse des EDPs [AGS08]. Contrairement à ce que leur nom suggère, les flots gradients gardent leur sens dans des espaces métriques, bien que le vecteur gradient n'y soit pas défini.

Définition 3.1 (Opérateur proximal). *Soit (X, d_X) un espace métrique, et soit $\mathcal{E} : X \rightarrow]-\infty, \infty]$ une fonction. On définit, sous réserve d'existence, pour tout $x \in X$ et tout $\varepsilon > 0$*

$$\operatorname{prox}_{\mathcal{E}}^\varepsilon(x) := \operatorname{argmin}_{y \in X} \frac{1}{2\varepsilon} d_X(x, y)^2 + \mathcal{E}(y). \quad (11)$$

L'opérateur proximal s'apparente à un pas de descente de gradient implicite. En effet, supposons que $X = \mathbb{H}$ est un espace de Hilbert, de sorte que $d_X(x, y) = \|x - y\|$, et que \mathcal{E} est différentiable en $y := \operatorname{prox}_{\mathcal{E}}^\varepsilon(x)$. Alors en différentiant (11, droite) à son optimum y on obtient

$$0 = \frac{y - x}{\varepsilon} + \nabla \mathcal{E}(y).$$

De manière équivalente, $y = x - \varepsilon \nabla \mathcal{E}(y)$, ce qui caractérise bien un pas de gradient implicite. On renvoie à [PB13] pour une étude détaillée des opérateurs proximaux, et en particulier la preuve [PB13, §4.1] que la suite $(x_n^\varepsilon)_{n \geq 0}$ définie dans la proposition suivante, qui correspond à une descente de gradient implicite, converge vers un minimiseur de \mathcal{E} sous des hypothèses adéquates.

Définition 3.2 (Flot gradient). *Sous les hypothèses de la Définition 3.1, soit $x_0 \in X$, et soit $\varepsilon > 0$. On définit, sous réserve d'existence, une suite $(x_n^\varepsilon)_{n \geq 0}$ et une application constante par morceaux $\mathbf{x}^\varepsilon : [0, \infty[\rightarrow X$ par*

$$x_0^\varepsilon = x_0 \quad x_{n+1}^\varepsilon = \operatorname{prox}_{\mathcal{E}}^\varepsilon(x_n^\varepsilon), \quad \forall n \geq 0, \quad \mathbf{x}^\varepsilon(t) = x_n^\varepsilon, \quad \forall t \in [n\varepsilon, (n+1)\varepsilon[.$$

Supposons qu'il existe une suite $\varepsilon_k \rightarrow 0$, telle que les fonctions $\mathbf{x}^{\varepsilon_k} : [0, \infty[\rightarrow X$ convergent localement uniformément vers une limite $\mathbf{x} : [0, \infty[\rightarrow X$. Alors on dit que \mathbf{x} est un flot gradient de \mathcal{E} pour la métrique d_X issu du point $x_0 \in X$.

En l'absence d'hypothèses sur l'espace X et la fonction \mathcal{E} , les Définitions 3.1 et 3.2 ne permettent d'établir ni l'existence ni l'unicité des objets considérés. L'existence d'un flot gradient et d'un minimiseur pour (11) s'établit sous des hypothèses de compacité. On dit qu'une partie d'un espace métrique est relativement compacte si son adhérence est compacte.

⁴L'approche EDP n'est plus l'état de l'art pour des tâches comme le débruitage d'image.

Proposition 3.3 (Existence). *Sous les hypothèses de la Définition 3.2. Supposons de plus que \mathcal{E} est semi-continue inférieurement, bornée inférieurement, telle que $\mathcal{E}(x_0) < \infty$, et que l'ensemble $\{x \in X; d_X(x_0, x) \leq C, \mathcal{E}(x) \leq C\}$ est relativement compact pour toute constante C .*

Alors le problème (11) admet toujours au moins une solution, et il existe au moins un flot gradient issu de x_0 au sens de la Définition 3.2.

Proof. Le problème (11) admet une solution car il s'agit de la minimisation d'une fonctionnelle s.c.i. sur un ensemble compact. On déduit de la Définition 3.1 que pour tout $k \geq 0$

$$d_X(x_k^\varepsilon, x_{k+1}^\varepsilon)^2 \leq \varepsilon(\mathcal{E}(x_k^\varepsilon) - \mathcal{E}(x_{k+1}^\varepsilon)),$$

et en particulier $(\mathcal{E}(x_n^\varepsilon))_{n \geq 0}$ est décroissante. Puis, en utilisant l'inégalité de Cauchy-Schwartz et une somme télescopique, on obtient pour tous $0 \leq m \leq n$

$$d_X(x_n^\varepsilon, x_m^\varepsilon)^2 \leq \left(\sum_{m \leq k < n} d_X(x_k^\varepsilon, x_{k+1}^\varepsilon) \right)^2 \leq (n-m) \sum_{m \leq k < n} d_X(x_k^\varepsilon, x_{k+1}^\varepsilon)^2 \leq (n-m)\varepsilon(\mathcal{E}(x_m^\varepsilon) - \mathcal{E}(x_n^\varepsilon)).$$

On en déduit, pour tous temps $0 \leq s \leq t$

$$d_X(\mathbf{x}^\varepsilon(t), \mathbf{x}^\varepsilon(s))^2 \leq (t-s+\varepsilon)(\mathcal{E}(\mathbf{x}^\varepsilon(s)) - \mathcal{E}(\mathbf{x}^\varepsilon(t))) \leq (t-s+\varepsilon)(\mathcal{E}(x_0) - \inf_X f).$$

Il s'agit d'une propriété d'équi-continuité des applications \mathbf{x}^ε qui, par le théorème d'Arzelà-Ascoli et grâce à l'hypothèse de compacité, assure l'existence d'une sous-famille $\mathbf{x}^{\varepsilon_n}$ convergeant uniformément sur tout segment de $[0, \infty[$. \square

L'unicité et la régularité du flot gradient et de l'opérateur proximal s'établissent sous des hypothèses de convexité⁵.

Proposition 3.4 (Unicité). *Sous les hypothèses de la Définition 3.1. Supposons de plus que X est un Hilbert et que \mathcal{E} est convexe. Alors, sous réserve d'existence, l'opérateur proximal est 1-Lipschitz, et les flots gradients \mathbf{x}, \mathbf{y} issus de points x_0, y_0 satisfont $\|\mathbf{x}(t) - \mathbf{y}(t)\| \leq \|x_0 - y_0\|$ pour tout $t \geq 0$.*

Proof. Soient $x = \text{prox}_{\mathcal{E}}^\varepsilon(x_0)$, $y = \text{prox}_{\mathcal{E}}^\varepsilon(y_0)$, et $v := y - x$. On a par (11) pour tout $t \in \mathbb{R}$

$$\frac{1}{2\varepsilon}\|x - x_0\|^2 + \mathcal{E}(x) \leq \frac{1}{2\varepsilon}\|x + tv - x_0\|^2 + \mathcal{E}(x + tv).$$

En sommant cette inégalité avec l'analogue obtenue en remplaçant (x, x_0) par (y, y_0) , on obtient

$$2\varepsilon(\mathcal{E}(x) + \mathcal{E}(y) - \mathcal{E}(x + tv) - \mathcal{E}(y - tv)) \leq \|x + tv - x_0\|^2 + \|y + tv - y_0\|^2 - \|x - x_0\|^2 - \|y - y_0\|^2.$$

Par convexité, le terme de gauche est positif pour tout $t \in [0, 1]$. Un développement limité du terme de droite lorsque $t \rightarrow 0$ donne $0 \leq 0 + t\langle v, x - x_0 - y + y_0 \rangle + o(t)$, donc en réarrangeant les termes

$$\|x - y\|^2 \leq \langle x - y, x_0 - y_0 \rangle.$$

Ceci implique $\|x - y\| \leq \|x_0 - y_0\|$ par Cauchy-Schwartz, et établit que $\text{prox}_{\mathcal{E}}^\varepsilon$ est 1-Lipschitz.

Par une récurrence immédiate, on obtient avec les notations de la Définition 3.2, $\|x_n^\varepsilon - y_n^\varepsilon\| \leq \|x_0 - y_0\|$ pour tout $n \geq 0$. Le résultat annoncé s'ensuit. \square

⁵On pourrait également montrer l'existence de (11) sous l'hypothèse que X est un Hilbert et que \mathcal{E} est s.c.i. convexe, comme pour l'existence du projeté orthogonal.

3.2 L'équation de la chaleur en tant que flot gradient

L'équation de la chaleur possède une double interprétation en tant que flot gradient: celui de l'*énergie de Dirichlet dans l'espace de Hilbert \mathbb{L}^2* , et celui de l'*entropie dans l'espace des mesures positives muni de la métrique de Wasserstein* (transport optimal) [JKO98]. Dans cette sous-section, on justifie l'interprétation formelle de l'équation de la chaleur par première approche (la plus ancienne). Ceci permet de justifier son existence, son unicité, et sa régularité par rapport aux paramètres, par application directe des Définitions 3.1 et 3.2 et Propositions 3.3 et 3.4. Soit $\Omega \subseteq \mathbb{R}^2$ un domaine borné de bord régulier, on note dans la suite

$$\mathbb{L}^2 := \mathbb{L}^2(\Omega), \quad \langle u, v \rangle_{\mathbb{L}^2} = \int_{\Omega} u(x)v(x) dx. \quad (12)$$

Soit $D \in C^0(\overline{\Omega}, S_d^{++})$ un champ de matrices symétriques définies positives. L'énergie de Dirichlet \mathcal{E} d'une fonction $u \in \mathbb{H}^1(\Omega)$ est définie par

$$\mathcal{E}(u) := \frac{1}{2} \mathcal{Q}(u, u), \quad \text{où } \mathcal{Q}(u, u) := \int_{\Omega} \|\nabla u(x)\|_{D(x)}^2 dx. \quad (13)$$

On a noté $\|v\|_M := \sqrt{\langle v, Mv \rangle}$ pour tous $v \in \mathbb{R}^d$, $M \in S_d^{++}$. La fonction \mathcal{Q} est une forme bilinéaire symétrique positive sur $\mathbb{H}^1(\Omega)$, dont la forme mixte $\mathcal{Q}(u, v)$ s'obtient par polarisation

$$\mathcal{Q}(u, v) = \frac{\mathcal{Q}(u + v, u + v) - \mathcal{Q}(u - v, u - v)}{4}. \quad \left(\text{Ici } \mathcal{Q}(u, v) = \int_{\Omega} \langle \nabla u(x), D(x) \nabla v(x) \rangle dx. \right) \quad (14)$$

On étend l'énergie de Dirichlet \mathcal{E} à $\mathbb{L}^2(\Omega) \setminus \mathbb{H}^1(\Omega)$, par la valeur $+\infty$.

On vérifie les hypothèses des Propositions 3.3 et 3.4, portant sur des propriétés de convexité et compacité, dans l'espace de Hilbert $X = \mathbb{L}^2(\Omega)$. Puis on calcule formellement le gradient de l'énergie de Dirichlet dans $\mathbb{L}^2(\Omega)$, et enfin on justifie la positivité de la solution.

Compacité. Les valeurs propres du champ de matrices D sont bornées supérieurement et inférieurement, par des constantes c_{\min} et c_{\max} . En effet, ces valeurs propres sont continues et positives sur l'ensemble compact $\overline{\Omega}$. Ainsi $c_{\min}\|v\|^2 \leq \|v\|_{D(x)}^2 \leq c_{\max}\|v\|^2$ pour tout $v \in \mathbb{R}^d$, $x \in \overline{\Omega}$. On en déduit pour tout $u \in \mathbb{H}^1(\Omega)$

$$c_{\min}\|\nabla u\|_{\mathbb{L}^2(\Omega)}^2 \leq \mathcal{E}(u) \leq c_{\max}\|\nabla u\|_{\mathbb{L}^2(\Omega)}^2. \quad (15)$$

Les *injections de Sobolev* établissent que l'ensemble suivant

$$\{u \in \mathbb{L}^2(\Omega); \|u\|_{\mathbb{L}^2(\Omega)} \leq C, \|\nabla u\|_{\mathbb{L}^2(\Omega)} \leq C\}$$

est une partie compacte de $\mathbb{L}^2(\Omega)$, pour toute constante C , ce qui avec (15) établit la propriété de compacité requise dans Proposition 3.3.

Convexité L'énergie de Dirichlet est convexe car c'est une forme quadratique positive sur $\mathbb{H}^1(\Omega)$, étendue par $+\infty$ hors de ce sous espace. Notons que \mathcal{E} satisfait pour tous $u, v \in \mathbb{H}^1(\Omega)$, et tout $t \in [0, 1]$, comme toute forme quadratique

$$\mathcal{E}((1-t)u + tv) = (1-t)\mathcal{E}(u) + t\mathcal{E}(v) - t(1-t)\mathcal{E}(u - v). \quad (16)$$

(Cette identité s'obtient, comme l'identité du parallélogramme, en développant chaque expression par bilinéarité.)

Gradient de l'énergie de Dirichlet. Pour u et h suffisament régulières, on calcule

$$\mathcal{Q}(u, h) = \int_{\Omega} \langle \nabla h, D\nabla u \rangle = \int_{\Omega} [\operatorname{div}(hD\nabla u) - h \operatorname{div}(D\nabla u)] = \int_{\partial\Omega} h \langle \mathbf{n}, D\nabla u \rangle - \int_{\Omega} h \operatorname{div}(D\nabla u),$$

Ainsi, compte tenu de la définition de l'énergie de Dirichlet \mathcal{E} via la forme quadratique \mathcal{Q} ,

$$d\mathcal{E}_u(h) = \mathcal{Q}(u, h) = \langle h, -\operatorname{div}(D\nabla u) \rangle_{\mathbb{L}^2(\Omega)} + \int_{\partial\Omega} h \langle \mathbf{n}, D\nabla u \rangle \quad (17)$$

Si la condition de Neumann $\langle \mathbf{n}, D\nabla u \rangle = 0$ sur Ω est satisfaite, et si $\operatorname{div}(D\nabla u) \in \mathbb{L}^2(\Omega)$, alors on obtient comme annoncé par la Définition 2.3 le gradient de \mathcal{E} en u vis à vis du produit scalaire $\mathbb{L}^2(\Omega)$

$$\nabla_u \mathcal{E}(u) = -\operatorname{div}(D\nabla_x u).$$

Cela justifie l'interprétation de (10) comme flot gradient de (13) dans $\mathbb{L}^2(\Omega)$. Dans le cas contraire, si la condition de Neumann n'est pas satisfaite ou si $\operatorname{div}(D\nabla u)$ n'est pas de carré intégrable, (17, droite) ne définit pas une forme linéaire continue sur $\mathbb{L}^2(\Omega)$, et \mathcal{E} n'est donc pas différentiable en u .

Semi-continuité inférieure. L'inégalité de Cauchy-Schwartz, valable pour toute forme quadratique positive \mathcal{Q} , s'écrit formellement

$$\sqrt{\mathcal{Q}(u, u)} = \sup_{\mathcal{Q}(v, v) \neq 0} \frac{\mathcal{Q}(u, v)}{\sqrt{\mathcal{Q}(v, v)}}$$

Précisons les espaces d'appartenance de u et de la fonction test v . Deux choix sont possibles:

- $u \in \mathbb{H}^1$, $v \in \mathbb{H}^1$. C'est l'espace naturel de définition de \mathcal{Q} .
- $u \in \mathbb{L}^2$, $v \in C^2(\bar{\Omega})$ telle que $\langle \mathbf{n}, D\nabla v \rangle = 0$ sur $\partial\Omega$. En effet, cet ensemble de fonctions test v est dense dans $\mathbb{H}^1(\Omega)$, et permet de donner sens à $\mathcal{Q}(u, v)$ lorsque $u \in \mathbb{L}^2$ via (17).

Le second choix permet d'écrire $\sqrt{2\mathcal{E}(u)} = \sqrt{\mathcal{Q}(u, u)}$ comme enveloppe de formes linéaires continues. Donc \mathcal{E} est s.c.i comme supremum d'une famille de fonctions s.c.i.

Positivité de la solution. Finalement on justifie de la positivité de la solution de l'équation de la chaleur, si la condition initiale est positive. Rappelons qu'elle l'expression de l'opérateur proximal dans ce contexte

$$\operatorname{prox}_{\mathcal{E}}^\varepsilon(u) := \operatorname{argmin}_{v \in \mathbb{L}^2(\Omega)} \frac{1}{2\varepsilon} \|u - v\|_{\mathbb{L}^2(\Omega)}^2 + \frac{1}{2} \int_{\Omega} \|\nabla v(x)\|_{D(x)}^2.$$

Compte tenu de la Définition 3.2 du flot, il suffit de prouver que $v := \operatorname{prox}_{\mathcal{E}}^\varepsilon(u)$ est une fonction positive dès que u est positive. Posons $v_+(x) := \max\{0, v(x)\}$, et notons que $|u(x) - v_+(x)| \leq |u(x) - v(x)|$, car u est positive, et que $\nabla v_+(x)$ est soit nul soit égal à $\nabla v(x)$, selon que v est positive ou non, pour presque tout $x \in \Omega$. Par unicité du minimiseur, voir Proposition 3.4, on a $v_+ = v$, donc v est positive comme annoncé.

3.3 Schéma aux différences finies

Nous souhaitons traiter numériquement l'équation de la chaleur en préservant les propriétés de l'équation continue (décroissance de l'énergie de Dirichlet, positivité). Dans cette optique, il est naturel de construire le schéma numérique comme descente de gradient d'une énergie de Dirichlet discrétisée. La preuve de convergence ne sera pas présentée ici, mais s'adapte de techniques isotropes [JS13].

Dans cette section, on fixe le domaine $\Omega \subseteq \mathbb{R}^d$ (borné, régulier), l'échelle de discrétilisation $h > 0$, et on introduit le domaine discrétisé et le produit scalaire

$$\Omega_h := \Omega \cap h\mathbb{Z}^d, \quad \langle u, v \rangle_h := h^d \sum_{x \in \Omega_h} u(x)v(x),$$

où $u, v : \Omega_h \rightarrow \mathbb{R}$. On introduit également une forme quadratique dédiée à l'approximation de l'énergie de Dirichlet (13). Pour tout $u : \Omega_h \rightarrow \mathbb{R}$, on pose $\mathcal{E}_h(u) := \frac{1}{2}\mathcal{Q}_h(u, u)$ où

$$Q_h(u, u) := h^d \sum_{x \in \Omega_h} Q_h^x(u, u), \quad \text{et où } Q_h^x(u, u) := \|\nabla u(x)\|_{D(x)}^2 + \mathcal{O}(h^2). \quad (18)$$

La forme locale de Q_h^x en $x \in \Omega_h$ dépend de celle de la matrice $D(x)$, et la consistance à l'ordre deux n'est attendue que pour x intérieur au domaine. On se contente de donner $Q_h^x(u, u)$, pour la lisibilité, car la forme bilinéaire se retrouve par polarisation, voir (14).

Cas isotrope $D(x) = d(x) \text{Id}$. Etant donnée $u : \Omega_h \rightarrow \mathbb{R}$, et $x \in \Omega_h$, on pose

$$Q_h^x(u, u) = \frac{d(x)}{2h^2} \sum_{1 \leq i \leq d} \left[(u(x + he_i) - u(x))^2 + (u(x - he_i) - u(x))^2 \right], \quad (19)$$

où $(e_i)_{1 \leq i \leq d}$ désigne la base canonique de \mathbb{R}^d . Si x est suffisamment loin de $\partial\Omega$, alors un développement limité donne facilement la consistance (18, droite). Les différences finies faisant intervenir des points hors du domaine sont ignorées, ce qui approche des conditions au bord de Neumann.

Cas anisotrope. On présente une généralisation du schéma isotrope, issue de [FM14], et fondée sur la décomposition matricielle de Selling discutée §3.4. Celle-ci, limitée à la dimension $d \in \{2, 3\}$, prend la forme suivante,

$$D(x) = \sum_{1 \leq i \leq I} \rho_i(x) e_i e_i^T \quad \text{où } \rho_i(x) \geq 0, \quad e_i \in \mathbb{Z}^d, \quad \forall 1 \leq i \leq I. \quad (20)$$

On montre par ailleurs que les poids ρ_i dépendent continument de x , qu'au plus $d(d+1)/2$ sont non nuls pour chaque x , et que les offsets e_i sont bornés en fonction du conditionnement de $D(x)$.

Pour la commodité de notation, notons $\rho_{-i}(x) := \rho_i(x)$ et $e_{-i} := e_i$ pour tout $1 \leq i \leq I$, et posons

$$Q_h^x(u, u) = \frac{1}{2h^2} \sum_{1 \leq |i| \leq I} \rho_i(x) (u(x + he_i) - u(x))^2, \quad (21)$$

Nous obtenons, pour u suffisamment lisse et x suffisamment loin de $\partial\Omega$, la propriété de consistance

$$\begin{aligned}
Q_h^x(u, u) &= \sum_{1 \leq i \leq I} \rho_i(x) (\langle \nabla u(x), e_i \rangle^2 + \mathcal{O}(h^2)), \\
&= \sum_{1 \leq i \leq I} \rho_i(x) \operatorname{Tr}(\nabla u(x) \nabla u(x)^T e_i e_i^T) + \mathcal{O}(h^2), \\
&= \operatorname{Tr}\left(\nabla u(x) \nabla u(x)^T \sum_{1 \leq i \leq I} \rho_i(x) e_i e_i^T\right) + \mathcal{O}(h^2), \\
&= \operatorname{Tr}(\nabla u(x) \nabla u(x)^T D(x)) + \mathcal{O}(h^2), \\
&= \|\nabla u(x)\|_{D(x)}^2 + \mathcal{O}(h^2).
\end{aligned}$$

Condition Courant-Freidrichs-Levy (CFL). La discréétisation de l'équation de la chaleur par un schéma explicite s'écrit

$$I_h \frac{u_{n+1} - u_n}{\delta t} = Q_h u_n, \quad \text{soit } u_{n+1} = (\operatorname{Id} - \delta_t I_h^{-1} Q_h) u_n,$$

où l'on a noté Id la matrice identité, $I_h := h^d \operatorname{Id}$ la matrice du produit scalaire $\langle \cdot, \cdot \rangle_h$, et Q_h celle de la forme quadratique du même nom. Le schéma est stable en norme $\mathbb{L}^2(\Omega_h)$ pourvu⁶ que $\operatorname{Id} - \delta_t I_h^{-1} Q_h$ ait toutes ses valeurs propres dans $[-1, 1]$. De manière équivalente

$$\delta t Q_h \preceq 2I_h,$$

au sens de l'ordre sur les matrices symétriques, puisque par construction on a déjà $Q_h \succeq 0$. Or

$$Q_h^x(u, u) \leq \frac{1}{h^2} \sum_{1 \leq |i| \leq I} \rho_i(x) (u(x + he_i)^2 + u(x)^2),$$

en utilisant l'inégalité $(a + b)^2 \leq 2(a^2 + b^2)$, $a, b \in \mathbb{R}$. Ici et dans la suite, les indices tels que $x + he_i \notin \Omega_h$ sont exclus de la somme. Ainsi, en regroupant les coefficients de $u(x)^2$,

$$Q_h(u, u) \leq h^{d-2} \sum_{x \in \Omega_h} u(x)^2 \sum_{1 \leq |i| \leq I} (\rho_i(x) + \rho_i(x + he_i)) \leq Ch^{-2} \langle u, u \rangle_h$$

où

$$C := \max_{x \in \Omega_h} \sum_{1 \leq |i| \leq I} (\rho_i(x) + \rho_i(x + he_i)), \quad \left(\text{CFL : } C\delta t \leq 2h^2 \right). \quad (22)$$

Pour majorer la constante (22, gauche) on note que

$$\sum_{1 \leq i \leq I} \rho_i(x) \leq \sum_{1 \leq i \leq I} \rho_i(x) \|e_i\|^2 = \operatorname{Tr}(D(x)),$$

car les offsets $(e_i)_{1 \leq i \leq I}$ sont à coordonnées entières et non nuls. Par ailleurs les poids $\rho_i(x)$ issus de la décomposition de Selling dépendent de manière Lipschitz de la matrice $D(x)$. Ainsi, pour un champ D régulier et aux échelles petites, on a $C \lesssim 4 \max\{\operatorname{Tr}(D(x)); x \in \bar{\Omega}\}$.

⁶Sous la même condition, l'énergie définie par Q_h (ou toute puissance de celle-ci) décroît. Le schéma implicite est pour sa part inconditionnellement stable.

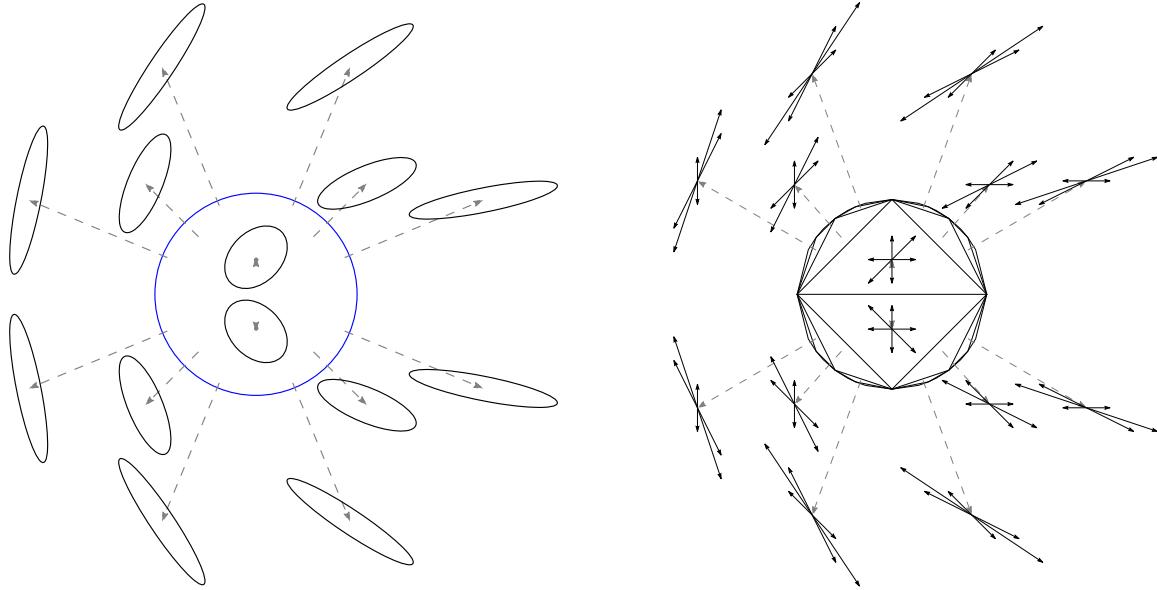


Figure 4: (Gauche) Ellipse définie par $\{v \in \mathbb{R}^2; \langle v, D(x, y)v \rangle = 1\}$, où $D(x, y)$ est définie par la paramétrisation de Pauli (23). (Droite) Superbase $D(x, y)$ -obtuse, et son opposée.

Positivité de la solution. Par construction, la matrice Q_h a toutes ses entrées négatives hors de la diagonale, et toutes ses entrées diagonales bornées par Ch^{-2} , où C est défini par (22). Sous la condition CFL (22), la matrice $I - \delta t I_h^{-1} Q_h$ a donc toutes ses entrées positives, et la positivité de la solution est préservée au fil des itérations du schéma.

3.4 Décomposition matricielle de Selling

La décomposition de Selling fait partie du domaine mathématique nommé *géométrie des réseaux Euclidiens*, ayant de nombreuses applications en arithmétique, cryptographie, théorie des groupes, étude des empilements de sphères, etc [CS13, Sch09]. Elle est introduite en 1874 [Sel74], voir également [CS92]. Dans cette section, on suppose toujours $i, j \in \{0, \dots, d\}$.

Définition 3.5 (Superbase). *On appelle superbase de \mathbb{Z}^d une famille $(v_0, \dots, v_d) \in (\mathbb{Z}^d)^{d+1}$ telle que $v_0 + \dots + v_d = 0$ et $|\det(v_1, \dots, v_d)| = 1$. Une superbase est dite D -obtuse, où $D \in S_d^{++}$, si $\langle v_i, Dv_j \rangle \leq 0$ pour tous $i \neq j$.*

Voir Fig. 4 les superbases obtuses associées à des matrices symétriques de taille 2×2 , paramétrées par

$$D(x, y) := \begin{pmatrix} 1+x & y \\ y & 1-x \end{pmatrix}, \quad \text{où } x^2 + y^2 < 1. \quad (23)$$

Définition 3.6. *A toute superbase (v_0, \dots, v_d) de \mathbb{Z}^d , on associe la famille $(e_{ij})_{i \neq j}$ définie par les relations $\langle e_{ij}, v_k \rangle = \delta_{ik} - \delta_{jk}$ pour tout $0 \leq k \leq d$*

Proposition 3.7 (Decomposition de Selling). *Pour toute matrice $D \in S_d$, et toute superbase (v_0, \dots, v_d) on a*

$$D = - \sum_{i < j} \langle v_i, Dv_j \rangle e_{ij} e_{ij}^T. \quad (24)$$

Proof. En notant D' le membre de droite, on note que $\langle v_i, Dv_j \rangle = \langle v_i, D'v_j \rangle$ pour tous $i \neq j$. Comme $v_i = -\sum_{k \neq i} v_k$, l'identité vaut aussi lorsque $i = j$. Comme (v_1, \dots, v_d) est une base, on a $D = D'$. \square

La formule (24) fait par construction intervenir des offsets e_{ij} à coefficients entiers. De plus si la superbase (v_0, \dots, v_d) est D -obtuse, alors les poids $\rho_{ij} = -\langle v_i, Dv_j \rangle$ sont positifs, comme désiré (20). L'algorithme de Selling permet de garantir cette propriété géométrique.

Proposition 3.8 (Algorithme de Selling). *Etant donnée une superbase (v_0, \dots, v_d) et une matrice $D \in S_d^{++}$, l'algorithme de Selling en dimension $d = 2$ (resp. $d = 3$) répète l'opération suivante: s'il existe $0 \leq i < j \leq d$ tels que $\langle v_i, Dv_j \rangle > 0$ alors (resp. en notant $\{i, j, k, l\} = \{0, 1, 2, 3\}$)*

$$(v_0, v_1, v_2) \leftarrow (-v_i, v_j, v_j - v_i) \quad (\text{resp. } (v_0, v_1, v_2, v_3) \leftarrow (-v_i, v_j, v_k + v_i, v_k + v_l)).$$

Cet algorithme termine, et la superbase finalement obtenue est D -obtuse.

Proof. Un calcul direct montre que la quantité $\mathcal{E}(v) := \sum_{0 \leq i \leq d} \|v_i\|_D^2$ décroît strictement à chaque itération, de $4\langle v_i, v_j \rangle$ en dimension $d = 2$ (resp. $2\langle v_i, v_j \rangle$ en dimension $d = 3$). En particulier la superbase reste bornée et prend des valeurs deux à deux distinctes à chaque itération de l'algorithme; donc celui-ci termine. Par construction, la superbase finale est D -obtuse. \square

Interprétation et généralisation par la programmation linéaire. Etant donnée $D \in S_d^{++}$, Voronoi définit sa première réduction par le programme linéaire

$$\min_{M \in S_d} \text{Tr}(DM), \quad \text{sous contrainte que } \langle e, Me \rangle \geq 1, \forall e \in \mathbb{Z}^d \setminus \{0\}.$$

A toute superbase $b = (v_0, \dots, v_d)$, associons $M_b := \frac{1}{2} \sum_{0 \leq i \leq d} v_i v_i^T$. On peut montrer que la matrice M_b satisfait (3.4, droite), et que si b est D -obtuse, alors elle résout (3.4). De plus, en dimension $d \in \{2, 3\}$, l'algorithme de Selling s'apparente à la méthode du simplexe pour la résolution du programme linéaire (3.4).

Par ailleurs, le programme linéaire dual à (3.4) prend la forme d'un problème de décomposition matricielle

$$\max_{\rho \geq 0} \sum_{e \in \mathbb{Z}^d \setminus \{0\}} \rho(e), \quad \text{sous contrainte que } \sum_{e \in \mathbb{Z}^d} \rho(e)ee^T = D.$$

Ainsi, la réduction de Voronoi permet d'étendre la décomposition de Selling en dimension $d > 3$, au prix de complications certaines à mesure que la dimension augmente.

3.5 Non-linéarité

La diffusion anisotrope et non-linéaire a fait l'objet d'une série de travaux dédiés à son utilisation pour débruiter des images, détecter et réhausser leurs contours, ou leur appliquer des effets artistiques [PM90, CLMC92, Wei98]. Dans ce cadre, les tenseurs de diffusion D sont en général construits à partir du gradient des données à traiter, sous la forme générale suivante:

$$\partial_t u = \text{div}(D[\nabla u] \nabla u), \tag{25}$$

où $D[\nabla u](t, x)$ dépend de manière non-linéaire de $\nabla u(t, \cdot)$, et est une matrice symétrique semi-définie positive. Le traitement mathématique, numérique, et les applications de l'équation non-linéaire dépendent de deux propriétés principales des tenseurs de diffusion $D[\nabla u]$: leur expression est-elle *locale* ou non, et leur forme *isotrope* ou non.

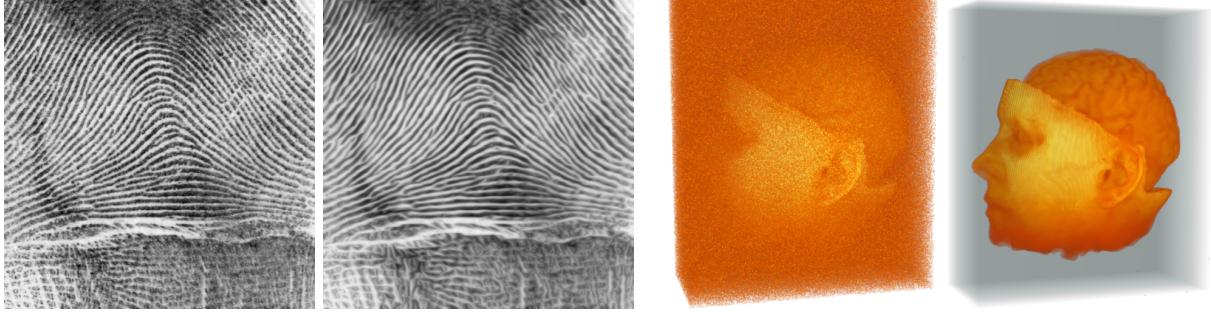


Figure 5: Avant/après l'effet de $\partial_t u = \operatorname{div}(D_u \nabla u)$, avec coefficients de diffusion non-linéaires et anisotropes $D[\nabla u]$ (Coherence enhancing diffusion [Wei98]).

- L'expression de $D[\nabla u](x)$ est *locale* si elle dépend seulement de $\nabla u(x)$. L'expression est au contraire *non-locale* si $D[\nabla u](x)$ dépend de $\nabla u(\cdot)$ au voisinage de x , en général via une convolution. Paradoxalement, l'analyse de l'EDP (existence, unicité, stabilité, discréétisation, ...) est en général plus simple dans le cas non-local grâce à l'effet régularisant de la convolution, et les contraintes induites due la forme de $D[\nabla u]$ (anisotropie, propriétés de monotonie, ...) sont moins sévères également dans ce cas.
- Diffusion *isotrope*⁷ si les matrices $D[\nabla u]$ sont en tout point proportionnelles à l'identité, et *anisotrope* sinon. L'anisotropie donne des degrés de liberté appréciables dans les applications, en permettant de diffuser tangentiellement à certaines directions liées aux données (contours d'objets, textures oscillantes).

Deux classes de modèles ont été particulièrement étudiées: faisant intervenir des tenseurs *locaux* et *isotropes*, ou non-locaux et anisotropes.

Dépendance locale et isotrope. Dans ces travaux [PM90], le coefficient de diffusion est proportionnel à l'identité, selon un coefficient dépendant de la norme du gradient:

$$D[\nabla u](x) = g(|\nabla u(x)|) \operatorname{Id}. \quad (26)$$

La fonction g est en général décroissante, de manière à diffuser faiblement dans les zones de fort gradient, et ainsi préserver les contours et discontinuités présents dans l'image. Les modèles suivants sont standard, et tirent leur nom de la fonctionnelle de Variation Totale (TV) ou des auteurs Perona et Malik (PM) [PM90]:

$$g_{\text{TV}}(s) = \frac{1}{s} \quad g_{\text{PM}}(s) = \frac{1}{1 + \lambda s^2}$$

Formellement, la diffusion non-linéaire (25) s'interprète dans le cas (26) comme le flot gradient de la fonctionnelle

$$\mathcal{E}_f(u) := \int_{\Omega} f(|\nabla u(x)|) dx, \quad \text{où } f(r) := \int_0^r s g(s) ds.$$

⁷Une certaine confusion de terminologie règne dans la littérature du fait que l'équation de diffusion peut être reformulée sous forme non-divergence : $\partial_t u = \langle \operatorname{div}(D[\nabla u]), \nabla u \rangle + \operatorname{Tr}(D[\nabla u] \nabla^2 u)$. Dans ce document l'isotropie/anistropie fait référence à la forme de tenseur de diffusion sous forme divergence.

Pour les modèles considérés ici, on note que $f_{\text{TV}}(s) = s$ et $f_{\text{PM}}(s) = (2\lambda)^{-1} \ln(1 + \lambda s^2)$. Si f est convexe et croissante, de manière équivalente si $s \in [0, \infty[\mapsto sg(s)$ est croissante et positive, alors l'énergie \mathcal{E}_f est convexe. Cette propriété est satisfaite dans le cas de la variation totale, mais pas du modèle de Perona et Malik, pour lequel l'unicité de la solution est donc perdue. Ce phénomène s'illustre visuellement par des instabilités menant à des artefacts *en marche d'escalier*.

Dépendance non-locale et anisotrope. Dans ces travaux, le coefficient de diffusion dépend d'une régularisation par convolution $G_\sigma \star u$ de la solution, où σ désigne l'échelle du bruit dans l'image, et G_σ est typiquement un noyau Gaussien. L'effet régularisant de la convolution permet d'établir l'existence, l'unicité, et la stabilité de la solution par rapport aux conditions initiales, à l'aide du lemme de Gronwall et de techniques relativement standard en analyse parabolique, présentées⁸ dans [CLMC92], qui ne seront pas reprises en détail ici.

Des constructions particulières de tenseurs de diffusion sont proposées dans [Wei98]. Dans une première étape, les directions locales de l'image à traiter sont identifiées grâce au *tenseur de structure*, qui écrit

$$S[\nabla u](x) := G_\rho \star (\nabla u_\sigma \nabla u_\sigma^T), \quad \text{où } u_\sigma := G_\sigma \star u.$$

Noter que $\nabla u_\sigma = (\nabla G_\sigma) \star u = G_\sigma \star (\nabla u)$ ne dépend que du gradient de u . Le paramètre σ correspond à l'échelle de cohérence du bruit (noise scale), et ρ à l'échelle des détails (feature scale). Les directions principales identifiées par le tenseur de structure, a.k.a. ses vecteurs propres $(e_i)_{1 \leq i \leq d}$, sont conservées dans le tenseur de diffusion. En revanche les valeurs propres sont ajustées en fonction de l'effet désiré

$$D[\nabla u](x) = \sum_{1 \leq i \leq d} \mu_i e_i e_i^T, \quad \text{où } S[\nabla u](x) = \sum_{1 \leq i \leq d} \lambda_i e_i e_i^T, \quad \text{et } (\mu_1, \dots, \mu_d) = f(\lambda_1, \dots, \lambda_d).$$

La fonction de transfert $f : \mathbb{R}^d \rightarrow \mathbb{R}^d$ des valeurs propres doit être invariante par permutation et continue, pour que $D[\nabla u]$ dépende continument de $S[\nabla u]$. Plusieurs choix sont considérés dans [Wei98], dédiés au renforcement des arêtes (“edge enhancing”, diffusion tangentielle aux arêtes exclusivement), ou à l'uniformisation des régions délimitées par celles-ci (“coherence enhancing”, diffusion partout excepté transversalement aux arêtes), ...

4 Chaleur et distances géodésiques

Dans cette section, on présente le lien entre les équations de Poisson et de la chaleur d'une part, et la distance géodésique riemannienne sur un domaine d'autre part, introduite §4.1. Sur le plan formel, ce lien découle d'une transformation logarithmique de la solution, présentée §4.2. Le cadre de des solutions de viscosité, présenté §4.3 permet d'établir la convergence sous des hypothèses adéquates. Les applications au traitement de l'image et de la géométrie sont abordées §4.4. Elles sont soit directes comme méthode de calcul de distances [CWW13], soit indirectes via le transport optimal [SdGP⁺15].

Les équations de Poisson (resp. de la chaleur), s'écrivent respectivement

$$u_\varepsilon - \varepsilon^2 \operatorname{div}(D\nabla u_\varepsilon) = u_0, \quad \left(\text{resp. } \partial_t u = \operatorname{div}(D\nabla u), \quad u|_{t=0} = u_0, \right) \quad (27)$$

⁸Cette référence suppose une diffusion isotrope, mais s'adapte immédiatement au cas anisotrope

sur un domaine $\Omega \subseteq \mathbb{R}^d$ muni d'un champ de matrices symétriques définies positives D , avec une trace de Neumann nulle au bord⁹. Compte tenu de leur caractère linéaire, les solutions de ces équations admettent une représentation intégrale

$$u_\varepsilon(x) = \int_{\Omega} \mathcal{P}_\varepsilon(x, y) u_0(y) dy, \quad (\text{resp. } u(t, x) = \int_{\Omega} \mathcal{Q}_t(x, y) u_0(y) dy) \quad (28)$$

Le noyau de Poisson (resp. de la chaleur) est lié à la distance géodésique d_M sur $\overline{\Omega}$ associée aux champ de tenseurs inverse $M = D^{-1}$. En effet, sous des hypothèses adéquates [Var67], on a pour $x, y \in \Omega$

$$\mathcal{P}_\varepsilon(x, y) = \exp\left(-\frac{d_M(x, y) + o(1)}{\varepsilon}\right). \quad (\text{resp. } \mathcal{Q}_t(x, y) = \exp\left(-\frac{d_M(x, y)^2 + o(1)}{4t}\right)). \quad (29)$$

Hypothèse 4.1. *Dans toute cette section, on suppose que $\Omega \subseteq \mathbb{R}^d$ est ouvert, borné, connexe, et a un bord de classe C^∞ . On suppose également $D \in C^\infty(\overline{\Omega}, S_d^{++})$, et on note $M := D^{-1}$.*

Remarque 4.2 (Intérêt de l'approche). *Il existe plusieurs méthodes pour calculer la distance géodésiques entre des points, comme la résolution directe de l'équation eikonale que l'on verra dans un chapitre ultérieur, alternatives à l'utilisation de l'équation de Poisson ou de la chaleur présentée dans cette section. Les principaux points forts de cette dernière sont:*

- La facilité de mise en oeuvre *sur des objets géométriques complexes (triangulations, patches, etc)*. En effet, une discrétisation de l'opérateur laplacien est généralement disponible ou connue, ce qui est plus rarement le cas pour l'opérateur non-linéaire eikonal.
- La rapidité, surtout en dimension 2 et lorsque les distances entre de nombreuses paires de points doivent être calculées. En effet, on peut tirer parti dans ce cas de la factorisation de Cholevski creuse de la discrétisation de l'EDP.
- La régularisation naturelle et implicite de la fonction distance, induite par cette approche, qui est appréciée dans certaines applications. En effet, l'approximation produite satisfait une relaxation de l'équation eikonale par un terme d'ordre deux (38).
- Une implémentation du transport optimal [SdGP⁺15], rendue possible par le calcul efficace des convolutions (28).

La résolution de l'équation eikonale est en revanche à préférer en dimension $d \geq 3$, pour des métriques non-riemannniennes, ou pour l'obtention d'ordres de convergences élevés.

4.1 Distance riemannienne

On rappelle la définition de la distance géodésique, sur un ouvert de \mathbb{R}^d muni d'une métrique riemannienne, et quelques propriétés de régularité. On discute également des méthodes de calcul de géodésiques riemannniennes. La distance riemannienne séparant $x, y \in \overline{\Omega}$ est définie par

$$d_M(x, y) := \min_{\gamma \in \Gamma_x^y} \int_0^1 \|\gamma'(t)\|_{M(\gamma(t))} dt, \quad (30)$$

où Γ_x^y désigne l'ensemble des chemins $\gamma \in \text{Lip}([0, 1], \Omega)$ tels que $\gamma(0) = x$ et $\gamma(1) = y$. On considère également la distance au bord du domaine, définie pour tout $x \in \overline{\Omega}$ par

$$v_0(x) := \min_{y \in \partial\Omega} d_M(x, y). \quad (31)$$

⁹L'étude présentée dans les sections suivantes utilise un cadre un peu différent, voir §4.2.

Dans la suite, on utilise souvent la notation \dot{x} pour les vecteurs, et \hat{x} pour les co-vecteurs. On se place toujours sous l'Hypothèse 4.1 de régularité du domaine et de la métrique.

Proposition 4.3. *La fonction d_M est une distance sur $\overline{\Omega}$, équivalente à la distance euclidienne, et v_0 est 1-Lipschitzienne vis-à-vis de d_M . Le minimum est atteint dans (30) et (31).*

Proof. Par compacité de $\overline{\Omega}$ et continuité de M , il existe des constantes $0 < c_{\min} \leq c_{\max} < \infty$ telles que $c_{\min}\|\dot{x}\| \leq \|\dot{x}\|_{M(x)} \leq c_{\max}\|\dot{x}\|$ pour tout $x \in \overline{\Omega}$ et $\dot{x} \in \mathbb{R}^d$. Ainsi, pour tout $\gamma \in \Gamma_x^y$

$$\int_0^1 \|\gamma'(t)\|_{M(\gamma(t))} dt \geq c_{\min} \int_0^1 \|\gamma'(t)\| dt \geq c_{\min} \|\gamma(1) - \gamma(0)\| = c_{\min} \|y - x\|.$$

Il s'ensuit que $d_M(x, y) \geq c_{\min}\|x - y\|$. Par ailleurs, compte tenu de la régularité de Ω , il existe toujours un chemin $\gamma \in \Gamma_x^y$ de longueur euclidienne bornée par $C_0\|x - y\|$, où $C_0 = C_0(\Omega)$. On en déduit $d_M(x, y) \leq c_{\max}C_0\|x - y\|$.

Par cette estimation, et l'invariance de la longueur par reparamétrisation, on peut se restreindre dans (30) aux chemins $\gamma \in \Gamma_x^y$ qui sont C_0c_{\max}/c_{\min} -Lipschitz vis-à-vis de la distance euclidienne. Comme ils forment un sous ensemble compact de $C^0([0, 1], \overline{\Omega})$, et que la longueur riemannienne est semi-continue inférieurement pour cette topologie, l'infimum (30) est atteint.

La fonction v_0 est 1-Lipschitzienne, vis-à-vis de d_M , car c'est l'infimum de la famille de fonctions $x \mapsto d_M(x, y)$, $y \in \partial\Omega$, qui ont toutes cette propriété. L'infimum (31) est atteint par continuité de $y \mapsto d_M(x, y)$ et compacité de $\partial\Omega$. \square

Proposition 4.4 (Backtracking). *Soit $x \in \Omega$, et soit $\gamma : [0, T] \rightarrow \overline{\Omega}$ un chemin minimisant pour (31), paramétré à vitesse riemannienne unité avec $T := v_0(x)$, $\gamma(0) = x$, $\gamma(T) \in \partial\Omega$. Alors*

$$d_M(x, \gamma(t)) = t, \quad v_0(\gamma(t)) = v_0(x) - t \tag{32}$$

pour tout $t \in]0, T]$, et on a

$$\gamma(t) \in \operatorname{argmin}_{y \in \overline{\Omega}} \frac{1}{2t} d_M(x, y)^2 + v_0(y) \tag{33}$$

Preuve de (32). Soit $y_* \in \partial\Omega$ le minimiseur de (31), de sorte que $d_M(x, y_*) = v_0(x) = T$. Comme γ est 1-Lipschitz pour d_M , on a $d_M(x, \gamma(t)) \leq t$ et $d_M(y_*, \gamma(t)) \leq T - t$. Par inégalité triangulaire ce sont des égalités. Par ailleurs comme v_0 est 1-Lipschitz pour d_M , $|v_0(\gamma(t)) - T| = |v_0(\gamma(t)) - v_0(x)| \leq d_M(x, \gamma(t)) = t$ et $0 \leq v_0(\gamma(t)) \leq d_M(y_*, \gamma(t)) = T - t$, donc $v_0(\gamma(t)) = T - t$.

Preuve de (33). Soit $y \in \Omega$, et soit $\delta := d_M(x, y)$. On a

$$\frac{1}{2t} d_M(x, y)^2 + v_0(y) \geq \frac{1}{2t} d_M(x, y)^2 + v_0(x) - d_M(x, y) = \frac{\delta^2}{2t} - \delta + v_0(x) \geq v_0(x) - \frac{t}{2},$$

en utilisant successivement que v_0 est 1-Lipschitz pour d_M , puis l'optimisation d'une fonction quadratique en δ . L'égalité a lieu lorsque $v_0(y) = v_0(x) - d_M(x, y)$ et $d_M(x, y) = t$, ce qui est justement satisfait pour $y = \gamma(t)$. \square

La Proposition 4.4 exprime, sous réserve d'unicité du minimiseur de (33), que $\gamma(\varepsilon) = \operatorname{prox}_{v_0}^\varepsilon(x)$. En d'autres termes, le chemin optimal γ correspond à une descente du gradient de v_0 vis à vis de la métrique d_X . Sous la forme d'une EDO, on a $\gamma(0) = x$ et sous réserve de différentiabilité

$$\gamma'(t) = -V(\gamma(t)), \quad \text{où } V(x) := D(x)\nabla v_0(x). \tag{34}$$

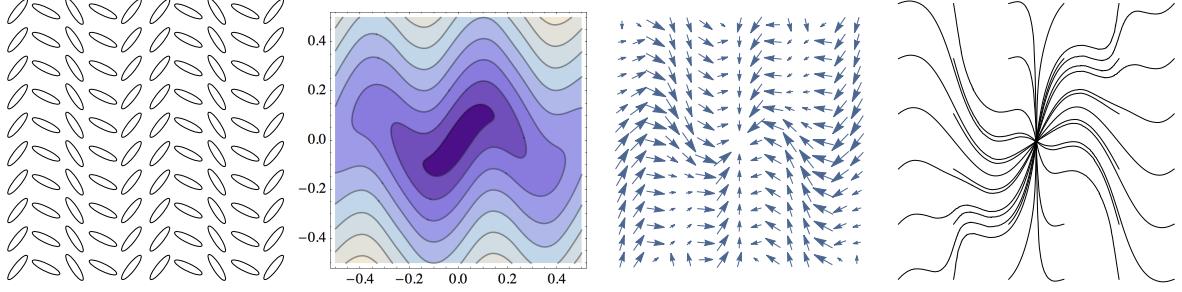


Figure 6: (i) Représentation d'une métrique riemannienne via l'indicatrice de Tissot. (ii) Distance riemannienne au point central (31). (iii) Opposée du gradient riemannien de cette distance (34, droite), et (iv) géodésiques minimisantes vers le point central.

On appelle *backtracking* (retour en arrière) la méthode de calcul de géodésiques minimisantes fondée sur le calcul préliminaire de v_0 , puis la résolution numérique de cette EDO du premier ordre.

Les chemins optimaux pour (30) et (31) satisfont également une EDO du second ordre, dite d'Euler-Lagrange, qui ne fait pas intervenir v_0 . On appelle *shooting* (méthode de tir) l'approche du calcul de géodesiques minimisantes fondée sur sa résolution numérique, en ajustant la vitesse initiale dans le but de trouver des chemins minimaux.

Proposition 4.5 (Euler-Lagrange). *Les minima (30) et (31) sont atteints et les chemins optimaux correspondants satisfont, après reparamétrisation à vitesse riemannienne constante,*

$$\frac{d}{dt} \nabla_{\dot{x}} L(\gamma, \gamma') = \nabla_x L(\gamma, \gamma'), \quad \text{où } L(x, \dot{x}) = \frac{1}{2} \|\dot{x}\|_{M(x)}^2. \quad (35)$$

Dans le cas (31), on a de plus $\gamma'(1) \propto D(y)\mathbf{n}(y)$, où \mathbf{n} désigne la normale extérieure à Ω .

Proof. (On admet la régularité du chemin optimal γ .) Par l'inégalité de Cauchy-Schwartz, on a

$$d_M(x, y)^2 = \min_{\gamma \in \Gamma_x^y} \int_0^1 \|\gamma'(t)\|_{M(\gamma(t))}^2 dt,$$

le minimum étant atteint lorsque γ minimise (30) et est paramétré à vitesse constante, ce qui motive l'introduction du lagrangien (35, droite). Considérons un chemin γ optimal pour (30) et une petite perturbation η , supposés réguliers. Un développement limité d'ordre un donne

$$\int_0^1 L(\gamma + \eta, \gamma' + \eta') dt = \int_0^1 L(\gamma, \gamma') dt + \int_0^1 \left(\langle \eta, \nabla_x L \rangle + \langle \eta', \nabla_{\dot{x}} L \rangle \right) dt + o(\|\eta\|).$$

où pour alléger l'écriture on omet les arguments de $\gamma(t), \eta(t), L(\gamma(t), \gamma'(t))$ sous l'intégrale. Par une intégration par parties, le terme d'ordre un en η s'écrit

$$\int_0^1 \left(\langle \eta, \nabla_x L \rangle - \langle \eta, \frac{d}{dt} \nabla_{\dot{x}} L \rangle \right) dt + \langle \eta(1), \nabla_{\dot{x}} L(\gamma(1), \gamma'(1)) \rangle - \langle \eta(0), \nabla_{\dot{x}} L(\gamma(0), \gamma'(0)) \rangle. \quad (36)$$

Dans le cas (30) de la distance riemannienne entre points fixés, la perturbation est sujette à $\eta(0) = \eta(1) = 0$ ce qui élimine les termes de bord dans (36), et l'annulation du terme intégral donne (35, gauche). Dans le cas (31), la perturbation satisfait $\eta(0) = 0$ mais est seulement sujette à $y + \eta(1) \in \partial\Omega$, donc $\nabla_{\dot{x}} L(\gamma(1), \gamma'(1)) \propto \mathbf{n}(y)$, ce qui équivaut à $\gamma'(1) \propto D(y)\mathbf{n}(y)$. \square

Proposition 4.6. *La distance au bord v_0 est C -concave, où $C = C(D, \Omega)$.*

Proof. On démontre ce résultat en établissant d'une part que v_0 est lisse au voisinage de $\partial\Omega$, et d'autre part qu'elle est localement C -concave pour une constante C dépendant de la distance au bord.

(Régularité au bord.) Pour tout $y \in \partial\Omega$, et tout $t \geq 0$ assez petit, soit $\gamma_y(t)$ la solution de (35) avec pour conditions initiales $\gamma_y(0) = y$ et $\gamma'_y(0) = -D(y)\mathbf{n}(y)$. Alors $(y, t) : \partial\Omega \times [0, \varepsilon] \mapsto \gamma_y(t)$ est un difféomorphisme sur son image. De plus $d(\gamma_y(t), \partial\Omega) = t$ par la Proposition 4.5, donc par le théorème des fonctions implicites v_0 est C^∞ au voisinage de $\partial\Omega$.

(C -convexité) Soit $x \in \Omega$ tel que $v_0(x) > \varepsilon$. Soit $\gamma : [0, T] \rightarrow \overline{\Omega}$ un chemin minimisant pour (31), paramétré à vitesse riemannienne constante, où $T = v_0(x) > \varepsilon$, $\gamma(0) = x$, $\gamma(T) \in \partial\Omega$. Pour h assez petit, le chemin $t \mapsto \gamma(t) + th/T$ prend ses valeurs dans $\overline{\Omega}$, et l'on a donc

$$v_0(x+h) \leq F(h) := \int_0^T \|\gamma'(t) + h/T\|_{M(\gamma(t)+th/T)} dt \leq v_0(x) + \langle \nabla F(0), h \rangle + C\|h\|^2$$

La fonction F , définie ci-dessus, est C^2 au voisinage de 0. Sa Hessienne peut être bornée indépendamment de x car $\gamma'(t)$ est de norme riemannienne unité pour tout $t \in [0, T]$, et car T est borné inférieurement et supérieurement. \square

4.2 Transformation logarithmique

L'image de la solution d'une EDP par une transformation lisse, est solution d'une nouvelle EDP. Dans le cas de l'équation de Poisson, le logarithme de sa solution satisfait une équation non-linéaire, donc plus complexe d'un certain point de vue, mais dans laquelle le petit coefficient ε joue un rôle plus transparent. Cette technique est introduite dans [Var67].

Pour simplifier l'étude, on considère un cadre un peu différent de celui qui motive l'introduction.

Définition 4.7. *Sous l'Hypothèse 4.1. Pour tout $\varepsilon > 0$, on note $u_\varepsilon : \overline{\Omega} \rightarrow \mathbb{R}$ la solution de*

$$u_\varepsilon - \varepsilon^2 \operatorname{div}(D\nabla u_\varepsilon) = 0 \quad \text{dans } \Omega, \quad u_\varepsilon = 1 \quad \text{sur } \partial\Omega. \quad (37)$$

Sous les hypothèses de la Définition 4.7, on montre [Eva10] (theorem 6 section 6.3) que $u_\varepsilon \in C^2(\overline{\Omega},]0, 1[)$ est une solution classique et strictement positive de (37), pour tout $\varepsilon > 0$. Notons l'interprétation énergétique

$$u_\varepsilon = \operatorname{argmin}_u \int_{\Omega} (u^2 + \|\nabla u\|_D^2) \quad \text{sous contrainte } u = 1 \text{ sur } \partial\Omega.$$

Présentons maintenant la transformation logarithmique et l'EDP non-linéaire associée, une "relaxation visqueuse" de l'équation eikionale riemannienne, qui correspond au cas limite $\varepsilon > 0$.

Proposition 4.8. *Sous les hypothèses de la Définition 4.7. Définissons $v_\varepsilon \in C^2(\overline{\Omega},]0, 1[)$ par*

$$v_\varepsilon := -\varepsilon \ln u_\varepsilon. \quad \left(\text{equiv: } u_\varepsilon = \exp\left(\frac{-v_\varepsilon}{\varepsilon}\right). \right)$$

Cette fonction est solution classique de l'EDP

$$\|\nabla v_\varepsilon\|_D^2 - 1 - \varepsilon \operatorname{div}(D\nabla v_\varepsilon) = 0 \quad \text{dans } \Omega, \quad v_\varepsilon = 0 \quad \text{sur } \partial\Omega. \quad (38)$$

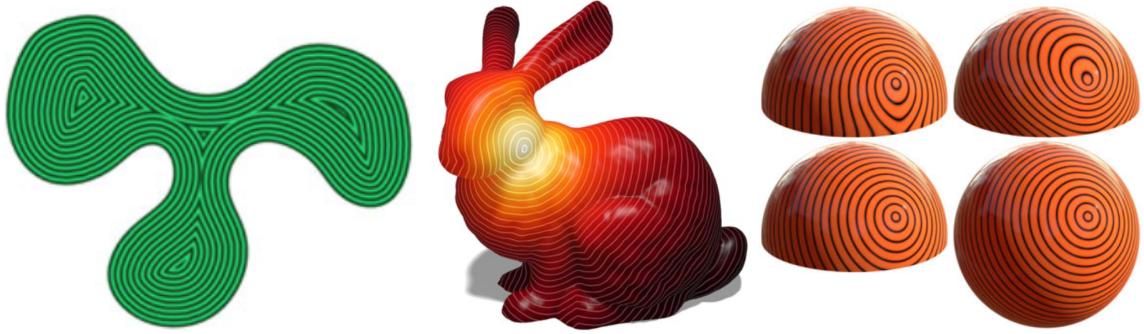


Figure 7: Approximation de la distance géodésique, via le noyau de Poisson ou l'équation de la chaleur. (i) Distance au bord d'un domaine du plan. (ii) Distance à un point d'une variété fermée. (iii) Distance à un point d'une variété à bord, avec différentes conditions au bord (Neumann, Dirichlet, Robin, sphère complète). Credits image : [CWW13].

Proof. La fonction v_ε est C^2 comme composée de fonctions C^2 , et on a $v_\varepsilon = -\varepsilon \ln(1) = 0$ sur $\partial\Omega$ comme annoncé. La formule de dérivation composée donne $\varepsilon \nabla u_\varepsilon = -u_\varepsilon \nabla v_\varepsilon$, d'où l'on déduit

$$\varepsilon^2 \operatorname{div}(D\nabla u_\varepsilon) = -\varepsilon \operatorname{div}(u_\varepsilon D\nabla v_\varepsilon) = \langle u_\varepsilon \nabla v_\varepsilon, D\nabla v_\varepsilon \rangle - \varepsilon u_\varepsilon \operatorname{div}(D\nabla v_\varepsilon).$$

On a utilisé $\operatorname{div}(a\omega) = \langle \nabla a, \omega \rangle + a \operatorname{div}(\omega)$, valable pour tout un champ a de scalaires et ω de vecteurs assez réguliers. Ceci conclut la preuve. \square

Le formalisme des solutions de viscosité, présenté §4.3, permet de passer à la limite $\varepsilon \rightarrow 0$ dans (38), et de donner un sens à l'équation eikonale $\|\nabla v_0\|_D^2 - 1 = 0$, dont la solution se trouve être la distance riemannienne au bord (31).

Remarque 4.9 (Variantes et améliorations). *Plusieurs heuristiques sont proposées dans [CWW13] pour améliorer la précision l'approximation v_ε de la distance géodésique, dont une étape de post-traitement, ou encore l'utilisation de conditions au bord de Robin $u + \langle \nabla u, D\mathbf{n} \rangle = 0$.*

4.3 Solution de viscosité

Les solutions de viscosité sont un formalisme mathématique pour l'étude des EDPs [CIL92], qui est tout à fait distinct des flots gradient considérés §3.1. Ses objets principaux sont les fonctions continues ou semi-continues qui sont étudiées via des principes de comparaison, au lieu des espaces métriques ou de Hilbert et des considérations énergétiques vus précédemment. Le formalisme discret correspondant sera présenté dans un chapitre ultérieur.

Définition 4.10 (Opérateur elliptique). *Un opérateur F sur un domaine Ω est dit dégénéré elliptique si pour tout $u \in C^2(\overline{\Omega})$ et tout $x \in \overline{\Omega}$ on a*

$$Fu(x) := \mathcal{F}(x, u(x), \nabla u(x), \nabla^2 u(x)), \quad (39)$$

où $\mathcal{F} : \overline{\Omega} \times \mathbb{R} \times \mathbb{R}^d \times S_d$ est croissante par rapport à sa seconde variable (a.k.a. $u(x)$), et décroissante par rapport à sa dernière variable (a.k.a. $\nabla^2 u(x)$) pour l'ordre usuel¹⁰ sur les matrices symétriques.

On dit que F est elliptique s'il existe $\delta > 0$ tel que $u \mapsto Fu - \delta u$ est dégénéré elliptique.

¹⁰ $A \preceq B$ ssi $B - A$ est semi-définie positive.

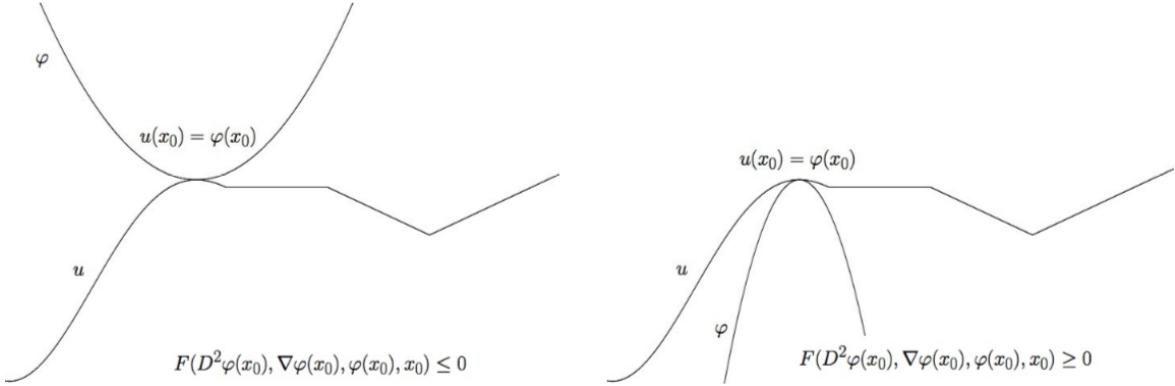


Figure 8: (i) Fonction test tangente *supérieurement* et *sous-solution*, (ii) fonction test tangente *inférieurement* et *sur-solution*, voir la Définition 4.12. Crédits image : [FGN13].

Les opérateurs apparaissant dans l'équation de Poisson (27) et l'équation eikonale relaxée (38) peuvent s'écrire sous la forme (39):

$$u - \varepsilon^2(\langle \operatorname{div} D, \nabla u \rangle + \operatorname{Tr}(D\nabla^2 u)), \quad \|\nabla v\|_D^2 - 1 - \varepsilon(\langle \operatorname{div} D, \nabla v \rangle + \operatorname{Tr}(D\nabla^2 v)). \quad (40)$$

On a utilisé l'identité $\operatorname{div}(D\nabla u) = \langle \operatorname{div}(D), \nabla u \rangle + \operatorname{Tr}(D\nabla^2 u)$, où la divergence de D est définie ligne par ligne. La monotonie par rapport à $\nabla^2 u(x)$ est satisfaite dans les deux cas car les matrices $D(x)$ sont symmétiques définies positives, pour tout $x \in \Omega$. On note que (40, gauche) est elliptique, tandis que (40, droite) est seulement dégénéré elliptique.

La théorie des solutions de viscosité permet de donner sens à l'équation (39) lorsque u est continue, mais pas forcément différentiable, par l'intermédiaire de fonctions test φ qui lui sont tangentes supérieurement ou inférieurement.

Définition 4.11. Soient $u \in C^0(\bar{\Omega})$, $\varphi \in C^2(\bar{\Omega})$, et $x \in \Omega$. On dit que φ est tangente supérieurement (resp. inférieurement) à u en x si

$$u(x) = \varphi(x), \quad \text{et} \quad u \leq \varphi \text{ sur } \bar{\Omega} \quad (\text{resp. } u \geq \varphi \text{ sur } \bar{\Omega}).$$

Définition 4.12. Soit F un opérateur dégénéré elliptique, et soit $u \in C^0(\bar{\Omega})$. On définit, au sens des solutions de viscosité:

- $Fu(x) \leq 0$: si $F\varphi(x) \leq 0$ pour toute $\varphi \in C^2(\bar{\Omega})$ tangente supérieurement à u en x .
- $Fu(x) \geq 0$: si $F\varphi(x) \geq 0$ pour toute $\varphi \in C^2(\bar{\Omega})$ tangente inférieurement à u en x .

Si $Fu \leq 0$ (resp. $Fu \geq 0$, resp. $Fu \leq 0$ et $Fu \geq 0$) sur $\bar{\Omega}$ on dit que u est sous-solution (resp. sur-solution, resp. solution) de viscosité de l'EDP définie par F .

Si la fonction u n'admet pas de tangente supérieure (resp. inférieure) en un point x , alors la condition apparaissant dans la Définition 4.12 est vide, et l'on a donc $Fu(x) \leq 0$ (resp. $Fu(x) \geq 0$) au sens des solutions de viscosité.

On dit qu'une EDP admet un principe de comparaison lorsque la propriété suivante est vérifiée : si \underline{u} et \bar{u} en sont une sous-solution et une sur-solution sur Ω , et si $\underline{u} \leq \bar{u}$ sur $\partial\Omega$, alors $\underline{u} \leq \bar{u}$ sur Ω .

Les EDPs elliptiques admettent souvent un tel principe, qui se démontre simplement *sous l'hypothèse simplificatrice* que \underline{u} et \bar{u} sont C^2 . En effet, soit $x \in \bar{\Omega}$ le maximiseur de $\underline{u} - \bar{u}$, et supposons pour la contradiction que $\underline{u}(x) - \bar{u}(x) > 0$. Alors $x \in \Omega$, car $\underline{u} \leq \bar{u}$ sur $\partial\Omega$, et

$$\nabla \underline{u}(x) = \nabla \bar{u}(x), \quad \nabla^2 \underline{u}(x) \preceq \nabla^2 \bar{u}(x), \quad \text{donc } F\underline{u}(x) - F\bar{u}(x) \geq \delta(\underline{u}(x) - \bar{u}(x)) > 0.$$

Cette dernière propriété contredit l'hypothèse que $F\underline{u}(x) \leq 0$ et $F\bar{u}(x) \geq 0$, ce qui conclut.

Pour établir le principe de comparaison lorsque \underline{u} et \bar{u} sont seulement continues, des hypothèses supplémentaires sont nécessaires, ainsi que des techniques comme le doublement de variables : maximiser $(x, y) \in \bar{\Omega}^2 \mapsto \underline{u}(x) - \bar{u}(y) - \frac{1}{\rho} \|x - y\|^2$, où $\rho > 0$ est petit.

On renvoie à [BR98, Théorème 2.1] pour la preuve du principe de comparaison pour l'équation de Poisson (40, gauche). Comme l'équation eikionale relaxée (40, droite) s'en déduit par un changement de variables monotone, la transformation logarithmique, elle le satisfait également.

Proposition 4.13. *Au sens des solutions de viscosité, on a $\|\nabla v_0\|_D^2 = 1$ et $-\text{Tr}(D\nabla^2 v_0) \geq -C$ dans Ω , où $C = C(D, \Omega)$ est une constante, et où v_0 est définie par (31).*

Proof. (Preuve de $\|\nabla v_0\|_D \leq 1$.) Soit $x \in \Omega$, et soit $\varphi \in C^2(\Omega)$ tangente supérieurement à v_0 en x . Alors pour tout $h \in \mathbb{R}^d$ assez petit

$$\varphi(x) - \varphi(x + h) \leq v_0(x) - v_0(x + h) \leq d_M(x, x + h) = \|h\|_{M(x)} + o(\|h\|),$$

en utilisant successivement $\varphi(x) = v_0(x)$, le caractère 1-Lipschitz de v_0 pour la distance d_M , et l'approximation locale de la distance riemannienne au voisinage de x par la norme sur l'espace tangent. On en déduit $\|\nabla \varphi(x)\|_{D(x)} \leq 1$, donc $\|\nabla v_0\|_D \leq 1$ sur Ω au sens des solutions de viscosité.

(Preuve de $-\text{Tr}(D\nabla^2 v_0) \geq -C$.) Supposons φ tangente inférieurement à v_0 en x , et rappelons que v_0 est C_0 -concave, voir la Proposition 4.6. Donc

$$\varphi(x + h) - 2\varphi(x) + \varphi(x - h) \leq v_0(x + h) - 2v_0(x) + v_0(x - h) \leq \frac{1}{2} C_0 \|h\|^2.$$

Ainsi $\nabla^2 \varphi \lesssim C_0 \text{Id}$, et $-\text{Tr}(D(x)\nabla^2 \varphi(x)) \geq -C_0 \text{Tr}(D(x)) \geq -C_1$ car D est borné sur Ω . On a montré que $-\text{Tr}(D\nabla^2 v_0) \geq -C$ sur Ω , au sens des solutions de viscosité.

(Preuve de $\|\nabla v_0\|_D \geq 1$.) Soit $\varphi \in C^2(\Omega)$ tangente inférieurement à v_0 en x . Soit $\gamma : [0, T] \rightarrow \Omega$ un chemin optimal pour (31) paramétré à vitesse riemannienne unité, où $T := v_0(x)$. On a donc $\gamma(t) = x + t\dot{x} + o(t)$ où $\|\dot{x}\|_{M(x)} = 1$. Alors pour tout $t \in [0, T]$,

$$t = v_0(x) - v_0(\gamma(t)) \leq \varphi(x) - \varphi(\gamma(t)) = -t \langle \nabla \varphi(x), \dot{x} \rangle + o(t).$$

Or $|\langle \nabla \varphi(x), \dot{x} \rangle| \leq \|\nabla \varphi(x)\|_{D(x)} \|\nabla \dot{x}\|_{M(x)}$, car $M(x) = D(x)^{-1}$. On en déduit $\|\nabla \varphi(x)\|_{D(x)} \geq 1$, donc $\|\nabla v_0\|_D \geq 1$ sur Ω au sens des solutions de viscosité. \square

Dans le résultat suivant, on construit des sous- et sur-solutions de viscosité de l'équation eikionale relaxée (38), donc une borne supérieure et inférieure pour sa solution. A cet effet, on introduit un noyau de convolution $G_\delta(x) := \frac{1}{\delta^d} G_1(x/\delta)$, où $\delta > 0$ et où $G_1 \in C^\infty(\mathbb{R}^d, \mathbb{R}_+)$ est positive, d'intégrale unité, et de support compact.

Corollaire 4.14. *Il existe des constantes $\varepsilon_0 > 0$ et C telles que pour tout $0 < \varepsilon \leq \varepsilon_0$, les fonctions suivantes sont des sur- et sous- solution de viscosité de (38), où l'on note $\delta := \sqrt{\varepsilon}$*

$$\bar{v}_\varepsilon := (1 + \varepsilon C)v_0 \quad \underline{v}_\varepsilon := (1 - C\delta)G_\delta * v_0 - C\delta.$$

Par convention, on étend v_0 à \mathbb{R}^d par 0 avant convolution avec le noyau G_δ de largeur δ .

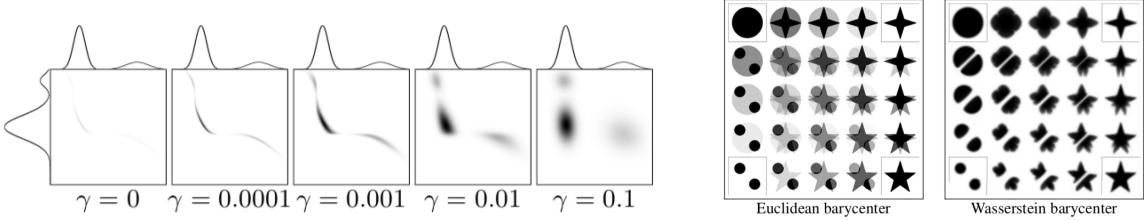


Figure 9: Gauche: effet de la relaxation entropique sur le plan de transport. Droite: interpolation de mesures dans les espaces Euclidiens et de Wasserstein. Credits image : [SdGP⁺15]

Proof. (Sur-solution.) Soit $\lambda = 1 + \varepsilon C \geq 1$. Alors λv_0 est nulle sur $\partial\Omega$, et on a dans Ω , en utilisant la Proposition 4.13

$$\|\lambda \nabla v_0\|_D^2 - 1 - \varepsilon(\text{Tr}(\lambda D \nabla^2 v_0) + \langle \text{div}(D), \nabla v_0 \rangle) \geq \lambda^2 - 1 - \lambda \varepsilon C_0 = \varepsilon(2C - C_0) + \mathcal{O}(\varepsilon^2).$$

Cette quantité est positive pourvu que $2C > C_0$ et que ε soit assez petit.

(Sous-solution) La fonction v_0 est globalement K -Lipschitz, donc pour tout $\delta > 0$

$$G_\delta * v_0 \leq C_1 \delta \text{ sur } \partial\Omega, \quad \|\nabla^2(G_\delta * v_0)\| \leq C_2 \delta^{-1} \text{ dans } \Omega.$$

On déduit de la première inégalité que $\underline{v}_\varepsilon \leq 0$ sur $\partial\Omega$ pourvu que $C \geq C_1$. Pour la seconde inégalité, on utilise que $\|\nabla G_\delta\| \leq C_3 \delta^{-1}$ et que v_0 est Lipschitz donc de gradient borné (défini presque partout). On en déduit, en rappelant que $\varepsilon = \delta^2$ et en posant $\lambda = 1 - C\delta$

$$\begin{aligned} \|\lambda \nabla \underline{v}_\varepsilon\|_D^2 - 1 - \varepsilon(\text{Tr}(\lambda D \nabla^2 \underline{v}_\varepsilon) + \langle \text{div}(D), \nabla \underline{v}_\varepsilon \rangle) &\leq \lambda^2(1 + C_3 \delta) - 1 + \varepsilon \lambda(C_2 \delta^{-1} + C_4) \\ &= \delta(-2C + C_2 + C_3) + \mathcal{O}(\delta^2). \end{aligned}$$

Cette quantité est négative pourvu que $2C > C_2 + C_3$ et que δ soit assez petit. Ce qui conclut. \square

Par le principe de comparaison, on obtient $\underline{v}_\varepsilon \leq v_\varepsilon \leq \overline{v}_\varepsilon$, pour tout $0 < \varepsilon \leq \varepsilon_0$ où $\underline{v}_\varepsilon$ et \overline{v}_ε sont définies au Corollaire 4.14. On vérifie par ailleurs que $\overline{v}_\varepsilon = v_0 + \mathcal{O}(\varepsilon)$ et $\underline{v}_\varepsilon = v_0 + \mathcal{O}(\sqrt{\varepsilon})$. Finalement, il existe une constante C telle que pour tout $0 < \varepsilon \leq \varepsilon_0$

$$v_\varepsilon - C\varepsilon \leq v_0 \leq v_\varepsilon + C\sqrt{\varepsilon}.$$

Ceci achève la preuve de la convergence.

4.4 Transport optimal entropique

On présente ici le problème de transport optimal avec relaxation entropique, discréétisé. Lorsque la fonction de coût est une distance Riemannienne, ou le carré d'une distance Riemannienne, on peut le résoudre de manière numériquement efficace, en exploitant l'asymptotique du noyau de poisson ou de la chaleur (29) et la formule de convolution (28). Ces techniques sont présentées en détail dans [SdGP⁺15], ainsi que dans [PCo19] avec d'autres approches.

Soient (X, μ) et (Y, ν) des ensembles finis munis de mesures de probabilité¹¹, soit $c : X \times Y \rightarrow \mathbb{R}$ un coût, et soit $\varepsilon > 0$. On souhaite calculer la relaxation entropique de la distance de Wasserstein, définie par

$$W_\varepsilon(\mu, \nu) := \min_{\pi \in \Pi(\mu, \nu)} \sum_{x \in X, y \in Y} c(x, y) \pi_{xy} + \varepsilon \text{KL}(\pi; \mu \otimes \nu), \quad (41)$$

¹¹Une mesure de probabilité sur un ensemble fini X s'identifie à une fonction $\mu : X \rightarrow \mathbb{R}_+$ positive et dont la somme des entrées est unitaire: $\sum_{x \in X} \mu_x = 1$.

On a noté $\Pi(\mu, \nu)$ l'ensemble des couplages entre les mesures de probabilité μ et ν , défini comme l'ensemble des $\pi : X \times Y \rightarrow \mathbb{R}_+$ satisfaisant aux contraintes suivantes

$$\forall x \in X, \sum_{y \in Y} \pi_{xy} = \mu_x, \quad \forall y \in Y, \sum_{x \in X} \pi_{xy} = \nu_y. \quad (42)$$

On a noté $\mu \otimes \nu$ la mesure de probabilité $(x, y) \in X \times Y \mapsto \mu_x \times \nu_y$, et KL la divergence de Kullback-Lieber, aussi appelée entropie relative, définie pour deux mesures de probabilités π, π^* sur un ensemble Z par

$$\text{KL}(\pi; \pi^*) := \sum_{z \in Z} \pi_z \ln \left(\frac{\pi_z}{\pi_z^*} \right) \quad \left(= 1 + \sum_{z \in Z} \pi_z \ln \left(\frac{\pi_z}{e\pi_z^*} \right) \right) \quad (43)$$

avec dans notre cas $Z = X \times Y$ et $z = (x, y)$. La divergence de Kullback-Lieber $\text{KL}(\pi; \pi_*)$ est une fonction convexe de son premier argument π car $s \in \mathbb{R}_+ \mapsto s \ln s$ est convexe. L'expression (43, droite) est équivalente si π est une mesure de probabilité. On la préfèrera car sa dérivée par rapport à π a une expression un peu plus simple.

Le problème (41) fait intervenir une fonction objectif convexe et C^∞ sur son domaine de définition $]0, \infty[^{X \times Y}$. Ce problème est soumis à $N_X + N_Y$ contraintes linéaires (42), la contrainte de positivité de π étant gérée implicitement par la pénalisation entropique. (En effet, $s \ln s$ a un taux d'accroissement $-\infty$ en 0^+ .) La principale obstruction à une résolution numérique directe est la dimension $N_X \times N_Y$ de l'inconnue π , qui excède typiquement la capacité mémoire d'un ordinateur si X et Y sont des images. Pour cette raison on considère le problème dual.

Dualité de Kantorovitch. Pour exprimer le problème dual, on introduit de deux fonctions inconnues $\phi : X \rightarrow \mathbb{R}$ et $\psi : Y \rightarrow \mathbb{R}$ appelées potentiels de Kantorovitch, qui permettent de reformuler les contraintes marginales (42) au sein de la fonction objectif. Ainsi (41) s'écrit

$$\begin{aligned} & \min_{\pi \geq 0} \left(\sum_{\substack{x \in X \\ y \in Y}} c(x, y) \pi_{xy} + \varepsilon \text{KL}(\pi; \mu \otimes \nu) + \sup_{\phi} \sum_{x \in X} \phi(x) \left(\mu_x - \sum_{y \in Y} \pi_{xy} \right) + \sup_{\psi} \sum_{y \in Y} \psi(y) \left(\nu_y - \sum_{x \in X} \pi_{xy} \right) \right) \\ &= \min_{\pi \geq 0} \sup_{\phi, \psi} \left(\sum_{x \in X} \phi(x) \mu_x + \sum_{y \in Y} \psi(y) \nu_y + \sum_{\substack{x \in X \\ y \in Y}} \left(c(x, y) - \phi(x) - \psi(y) + \varepsilon \ln \left(\frac{\pi_{xy}}{e\mu_x\nu_y} \right) \right) \pi_{xy} \right) + \varepsilon. \end{aligned} \quad (44)$$

On a utilisé l'expression (43, droite) de la divergence KL , car dans les calculs qui suivent π n'est pas forcément de somme unité. Le théorème du *minimax de Sion*¹² permet d'échanger l'ordre des optimisations dans la dernière expression, et donc d'obtenir

$$\begin{aligned} W_\varepsilon(\mu, \nu) - \varepsilon &= \sup_{\phi, \psi} \left(\sum_{x \in X} \phi(x) \mu_x + \sum_{y \in Y} \psi(y) \nu_y + \min_{\pi \geq 0} \sum_{\substack{x \in X \\ y \in Y}} \left(c(x, y) - \phi(x) - \psi(y) + \varepsilon \ln \left(\frac{\pi_{xy}}{e\mu_x\nu_y} \right) \right) \pi_{xy} \right) \\ &= \sup_{\phi, \psi} \left(\sum_{x \in X} \phi(x) \mu_x + \sum_{y \in Y} \psi(y) \nu_y - \varepsilon \sum_{\substack{x \in X \\ y \in Y}} \mu_x \nu_y \exp \left(\frac{\phi(x) + \psi(y) - c(x, y)}{\varepsilon} \right) \right). \end{aligned} \quad (45)$$

¹²Les hypothèses [Kom88] sont la convexité des ensembles d'optimisation, la compacité de l'un deux (ici π est bornée), la quasi-convexité s.c.i de la fonction objectif par rapport à la première variable, et sa quasi-concavité s.c.s par rapport à la seconde (ici elle est continue et respectivement convexe et linéaire).

L'expression (45) porte le nom de formulation duale de Kantorovitch du problème du transport optimal, ici discret et avec relaxation entropique. On a utilisé l'identité $\min\{(a+b \ln(s/e))s; s > 0\} = b \exp(-a/b)$, atteint lorsque $\ln(s) = -a/b$, pour tous $a, b > 0$, ce qui se vérifie aisément par différentiation. En particulier, étant connus les potentiels de Kantorovitch ϕ et ψ optimaux, le plan de transport π est caractérisé par

$$\pi_{xy} = \mu_x \nu_y \exp\left(\frac{\phi(x) + \psi(y) - c(x, y)}{\varepsilon}\right) = \mu_x \nu_y \Phi(x) \Psi(y) K(x, y).$$

Pour simplifier l'écriture, on a introduit $\Phi = \exp(\phi/\varepsilon)$, $\Psi = \exp(\psi/\varepsilon)$, et $K(x, y) = \exp(-c(x, y)/\varepsilon)$.

Algorithme de Sinkhorn. Notons $F(\phi, \psi)$ la quantité maximisée dans (45). L'algorithme de maximisation alternée appliquée à cette fonction porte le nom d'algorithme de Sinkhorn. Pour certains coûts c , il peut se calculer grâce aux identités données en introduction de §4, de manière approchée mais numériquement efficace [SdGP⁺15].

Différentions $F(\phi, \psi)$ par rapport à $\phi(x)$, où $x \in X$ est un point quelconque mais fixé:

$$\frac{\partial}{\partial \phi(x)} F(\phi, \psi) = \mu_x - \sum_{y \in Y} \mu_x \nu_y \exp\left(\frac{\phi(x) + \psi(y) - c(x, y)}{\varepsilon}\right)$$

En résolvant $\frac{\partial}{\partial \phi(x)} F(\phi, \psi) = 0$, c'est à dire en maximisant la fonction concave F par rapport à $\phi(x)$, on trouve

$$\exp(-\phi(x)/\varepsilon) = \sum_{y \in Y} \nu_y \exp\left(\frac{\psi(y) - c(x, y)}{\varepsilon}\right), \quad \text{equiv à: } \Phi(x)^{-1} = \sum_{y \in Y} \nu_y \Psi(y) K(x, y). \quad (46)$$

Dans les applications, on choisit souvent $c(x, y) = d_M(x, y)$, ou $c(x, y) = d_M(x, y)^2$, et on utilise un paramètre $\varepsilon > 0$ de relaxation entropique assez petit. Dans ces conditions (46, droite), est proche de la convolution (28), qui se calcule par la résolution d'un système linéaire.

5 Schémas monotones

Dans cette section, on présente les schémas numériques discrets dégénérés elliptiques (DDE), qui sont la contrepartie discrète des opérateurs différentiels dégénérés elliptiques (DE) considérés §4.3. Ils sont particulièrement efficaces pour la discréétisation directe de l'équation eikonale, et d'une manière plus générale des équations issues de la théorie du contrôle optimal, sans le terme régularisant rencontré dans la méthode de Varadhan (38).

On présente §5.1 les principes de comparaison, preuve d'existence et d'unicité de solutions, associés de manière abstraite à ces schémas. Des exemples de discréétisations d'EDPs sont donnés §5.2. Les sections §5.3 et 5.4 sont dédiées aux méthodes de résolution numérique.

Définition 5.1 (Schéma discret dégénéré elliptique). *Un schéma numérique F sur un ensemble fini X est dit DDE s'il est de la forme suivante: pour tout $u : X \rightarrow \mathbb{R}$ et tout $x \in X$*

$$Fu(x) := \mathcal{F}(x, u(x), (u(x) - u(y))_{y \in X \setminus \{x\}})$$

où \mathcal{F} est croissante en sa seconde variable, et en sa troisième variable coordonnée par coordonnée.

Un schéma est dit δ -elliptique, où $\delta > 0$, si $u \mapsto Fu - \delta u$ est DDE.

Si F est δ -elliptique, alors pour tout $u : X \rightarrow \mathbb{R}$ et tout $\lambda \geq 0$ on a

$$F(u + \lambda) \geq Fu + \delta\lambda, \quad F(u - \lambda) \leq Fu - \delta\lambda. \quad (47)$$

5.1 Principe de comparaison, existence d'une solution

A l'instar des opérateurs dégénérés elliptiques, les schémas DDE bénéficient de principes de comparaison, qui sont utilisés dans les preuves d'existence, d'unicité, et de stabilité des solutions. Les preuves sont plus simples dans le cadre discret, et ne nécessitent pas de passer par des fonctions test comme Définition 4.12 et autres technicités. Voir aussi [Obe06].

On commence par énoncer, Proposition 5.2 et Corollaire 5.4, les principes de comparaison associés aux schémas DDE.

Proposition 5.2 (Principe de comparaison strict). *Soit F un schéma DDE sur un ensemble fini X , et soient $\underline{u}, \bar{u} : X \rightarrow \mathbb{R}$ tels que $F\underline{u} < F\bar{u}$ sur X . Alors $\underline{u} < \bar{u}$ sur X .*

Proof. Soit $x \in X$ le maximiseur de $\underline{u} - \bar{u}$. Par définition on a pour tout $y \in X$

$$\underline{u}(y) - \bar{u}(y) \leq \underline{u}(x) - \bar{u}(x) \quad \text{donc} \quad \bar{u}(x) - \bar{u}(y) \leq \underline{u}(x) - \underline{u}(y).$$

En supposant par l'absurde que $\underline{u}(x) - \bar{u}(x) \geq 0$ on obtient par les propriétés de croissance de \mathcal{F}

$$\mathcal{F}(x, \bar{u}(x), (\bar{u}(x) - \bar{u}(y))_{y \in X \setminus \{x\}}) \leq \mathcal{F}(x, \underline{u}(x), (\underline{u}(x) - \underline{u}(y))_{y \in X \setminus \{x\}}),$$

ce qui contredit l'hypothèse $F\bar{u}(x) < F\underline{u}(x)$, et conclut la preuve. \square

Définition 5.3. *Soit F un schéma DDE, et soit $u : X \rightarrow \mathbb{R}$. On dit que u est une sous-solution stricte (resp. sous-solution, resp. solution, resp. sur-solution, resp. sur-solution stricte) si on a $Fu < 0$ sur X (resp. $Fu \leq 0$, resp. $Fu = 0$, resp. $Fu \geq 0$, resp. $Fu > 0$).*

On dit que F admet un principe de comparaison si on a $\underline{u} \leq \bar{u}$ pour toute sous-solution \underline{u} et toute sur-solution \bar{u} .

Corollaire 5.4 (Principe de comparaison). *Soit F un schéma DDE, telle que toute sur-solution est limite de sur-solutions strictes (resp. toute sous-solution est limite de sous-solutions strictes). Alors F admet un principe de comparaison. C'est en particulier le cas si F est elliptique.*

Proof. Soient \underline{u} une sous-solution, soit \bar{u} une sur-solution, et soit $(\bar{u}_n)_{n \geq 0}$ une suite de sur-solutions convergeant vers \bar{u} lorsque $n \rightarrow \infty$. Alors $\underline{u} < \bar{u}_n$ sur X pour tout $n \geq 0$ par la Proposition 5.2, et donc $\underline{u} \leq \bar{u}$ sur X par passage à la limite, comme annoncé.

Si F est elliptique, et \bar{u} est une sur-solution, alors $\bar{u} + \varepsilon$ est une sur-solution stricte pour tout $\varepsilon > 0$, par (47), ce qui conclut. \square

Le principe de comparaison implique l'unicité de la solution à un schéma DDE. Nous présentons maintenant deux résultats d'existence.

Proposition 5.5 (Solution de Perron). *Soit F un schéma dégénéré elliptique continu et admettant une sous-solution \underline{u} et une sur-solution \bar{u} . Alors F admet une solution telle que $\underline{u} \leq u \leq \bar{u}$, a savoir*

$$u(x) := \max\{v(x); v : X \rightarrow \mathbb{R} \text{ sous-solution et } \underline{u} \leq v \leq \bar{u} \text{ sur } X\}. \quad (48)$$

Proof. Par construction, on a $\underline{u} \leq u \leq \bar{u}$ sur X , et le max est atteint dans (48) par continuité de F . Soit $x \in X$, et soit $v : X \rightarrow \mathbb{R}$ tel que $v(x) = u(x)$; par définition on a $v(y) \leq u(y)$ pour tout $y \in X \setminus \{x\}$, et donc

$$Fu(x) = \mathcal{F}(x, u(x), (u(x) - u(y))_{y \in X \setminus \{x\}}) \leq \mathcal{F}(x, v(x), (v(x) - v(y))_{y \in X \setminus \{x\}}) \leq 0$$

Par ailleurs supposons par l'absurde que $Fu(x) < 0$. Alors $u(x) < \bar{u}(x)$ car on aurait sinon $Fu(x) \geq F\bar{u}(x) \geq 0$ par monotonie. Définissons $v_\varepsilon : X \rightarrow \mathbb{R}$ par $v_\varepsilon(y) = u(y)$ pour tout

$y \in X \setminus \{x\}$, et $v_\varepsilon(x) = u(x) + \varepsilon$. Alors pour tout $\varepsilon > 0$ on a $Fv_\varepsilon \leq Fu \leq 0$ sur $X \setminus \{x\}$ par monotonie, et pour ε assez petit on a $Fv_\varepsilon(x) \leq 0$ par continuité. De plus $\underline{u} \leq v_\varepsilon \leq \bar{u}$, pour ε assez petit, donc v_ε convient pour (48). Ceci contredit la définition de u , donc on a $Fu(x) \geq 0$, ce qui conclut. \square

Corollaire 5.6. *Soit F un schéma δ -elliptique et continu sur un ensemble fini X . Alors F admet une unique solution $u_* : X \rightarrow \mathbb{R}$, et on a $\|u - u_*\|_\infty \leq \delta^{-1}\|Fu\|_\infty$ pour tout $u : X \rightarrow \mathbb{R}$.*

Proof. Soit $u : X \rightarrow \mathbb{R}$, et soit $\lambda := \delta^{-1}\|Fu\|_\infty$. Alors $F(u + \lambda) \geq 0$ et $F(u - \lambda) \leq 0$ par (47). Ainsi F admet une solution u_* telle que $u - \lambda \leq u_* \leq u + \lambda$ par la Proposition 5.5. De plus u_* est unique par le principe de comparaison, Corollaire 5.4. \square

5.2 Exemples

Les schémas DDE permettent de discréteriser des EDPs dégénérées elliptiques de manière très générale [KT92, BS91, Obe08]. Elles sont cependant limitées en terme de précision, car les schémas DDE ne peuvent être précis qu'à l'ordre un pour des EDPs d'ordre un, et l'ordre deux pour des EDPs d'ordre deux. Par ailleurs, un certain nombre d'approches font intervenir des stencils de discréterisation très grands, ce qui dégrade la précision numérique des solutions.

On présente ici quelques discréterisations fondées sur la décomposition de Selling, voir la Proposition 3.7, ce qui permet d'obtenir des schémas DDE relativement compacts et précis. Observons d'abord que les conditions de croissance qui définissent les schémas DDE, voir la Définition 5.1, sont stables par bon nombre d'opérations.

Proposition 5.7 (Combinaison de schémas). *Si F et G des schémas DDE sur un ensemble fini X , alors c'est aussi le cas des schémas suivants*

$$\alpha F + \beta G, \quad \max\{F, G\}, \quad \min\{F, G\}, \quad \eta \circ F \quad v \mapsto F(\eta \circ v),$$

où $\alpha, \beta \geq 0$ et η est croissante. De plus FG est DDE si F et G sont toujours positifs.

Dans la suite des exemples, on suppose que le schéma est posé sur une grille cartésienne X d'échelle $h > 0$, et on se donne une fonction test u

$$X \subseteq h\mathbb{Z}^d, \quad u : X \rightarrow \mathbb{R}.$$

On se donne la décomposition d'une matrice $D \in S_d^+$ sous la forme suivante

$$D = \sum_{1 \leq i \leq I} \rho_i e_i e_i^T, \quad \text{où } \rho_i \geq 0, e_i \in \mathbb{Z}^d, \forall 1 \leq i \leq I. \quad (49)$$

Une telle décomposition est triviale si D est diagonale, et s'obtient typiquement par l'algorithme de Selling ou ses généralisations dans le cas contraire, voir la Proposition 3.7.

Différences finies upwind. Les différences finies décentrées d'ordre un, souvent dites upwind, permettent des discréterisations DDE des opérateurs d'ordre un. Elles sont définies par

$$\delta_h^{+e} u := \frac{u(x + he) - u(x)}{h}$$

où $x \in X$ et $e \in \mathbb{Z}^d$. Cette expression n'a de sens que si $x + he \in X$.

Lemme 5.8. *Là où il est défini sur X , le schéma $-\delta_h^{+e}u$ est dégénéré elliptique, et satisfait la propriété de consistance $\delta_h^{+e}u(x) = \langle \nabla u(x), e \rangle + \mathcal{O}(h)$ pour u lisse.*

Par combinaison linéaire *non-négative*, on obtient une discrétisation des opérateurs linéaires

$$\sum_{1 \leq i \leq d} |\alpha_i| \frac{u(x) - u(x - h \operatorname{sign}(\alpha_i) b_i)}{h} = \langle \nabla u(x), \dot{x} \rangle + \mathcal{O}(h), \quad \dot{x} = \sum_{1 \leq i \leq d} \alpha_i b_i.$$

A l'aide de la Proposition 5.7, on obtient des discrétisations d'opérateurs non-linéaires du premier ordre, tels que les opérateurs de Bellman qui s'écrivent $\max_{\dot{x} \in A} \langle \nabla u(x), \dot{x} \rangle$. Alternativement, on a une discrétisation particulièrement efficace de l'opérateur eikonal, introduite dans [RT92] dans le cadre isotrope, et qui s'étend au cadre anisotrope [Mir19] via la décomposition de Selling (49).

Corollaire 5.9. *Sous l'hypothèse (49), le schéma suivant est DDE*

$$F_h u(x) = h^{-2} \sum_{1 \leq i \leq I} \rho_i \max\{0, u(x) - u(x + he_i), u(x) - u(x - he_i)\}^2,$$

et consistant à l'ordre un avec l'opérateur de l'équation eikonaire Riemannienne, là où il est défini

$$F_h u(x) = \|\nabla u(x)\|_D^2 + \mathcal{O}(h).$$

Proof. Le caractère DDE découle de la Proposition 5.7; noter l'importance du 0, qui assure que le terme mis au carré est non-négatif. Par ailleurs, on note que

$$h^{-1} \max\{0, u(x) - u(x + he), u(x) - u(x - he)\} = |\langle \nabla u(x), e \rangle| + \mathcal{O}(h).$$

Ainsi, par linéarité de la trace

$$Fu(x) + \mathcal{O}(h) = \sum_{1 \leq i \leq I} \rho_i \langle \nabla u(x), e_i \rangle^2 = \operatorname{Tr} \left(\nabla u(x) \nabla u(x)^T \sum_{1 \leq i \leq I} \rho_i e_i e_i^T \right) = \|\nabla u(x)\|_D^2. \quad \square$$

Différences finies d'ordre deux. Les différences finies centrées d'ordre deux permettent des discrétisations DDE des opérateurs d'ordre deux. Elles sont définies par

$$\Delta_h^e u(x) := \frac{u(x + he) - 2u(x) + u(x - he)}{h^2}, \quad (50)$$

où $x \in X$ et $e \in \mathbb{Z}^d$. Cette expression n'a de sens que si $x + he \in X$ et $x - he \in X$.

Lemme 5.10. *Là où il est défini sur X , le schéma $-\Delta_h^e$ est dégénéré elliptique, et satisfait la propriété de consistance $\Delta_h^e u(x) = \langle e, \nabla^2 u(x) e \rangle + \mathcal{O}(h^2) = \operatorname{Tr}(\nabla^2 u(x) ee^T) + \mathcal{O}(h^2)$ pour u lisse.*

Par combinaison linéaire non-négative, on obtient une discrétisation DDE de l'opérateur laplacien sous forme non-divergence.

Corollaire 5.11. *Sous l'hypothèse (49), le schéma $-\Delta_h^D$ est DDE, où l'on définit*

$$\Delta_h^D u(x) := \sum_{1 \leq i \leq I} \rho_i \Delta_h^{e_i} u(x).$$

Il est de plus consistant à l'ordre deux avec l'opérateur laplacien anisotrope, sous forme non-divergence

$$\Delta_h^D u(x) = \operatorname{Tr}(D \nabla^2 u(x)) + \mathcal{O}(h^2).$$

Proof. Le caractère DDE découle de $-\Delta_h^D$ découle de celui de $-\Delta_h^e$ et de la stabilité par combinaisons linéaires positives. La consistance s'obtient par linéarité de la trace

$$\Delta_h^D u(x) + \mathcal{O}(h^2) = \text{Tr} \left(\nabla^2 u(x) \sum_{1 \leq i \leq I} \rho_i e_i e_i^T \right) = \text{Tr}(\nabla^2 u(x) D). \quad \square$$

La Proposition 5.7, combinée au Corollaire 5.11, permet d'obtenir discrétisations d'opérateurs dégénérés elliptiques non-linéaires d'ordre deux, voir par exemple [KT92, Kry05, BCM16]. On les exprime pour cela comme extrema de familles d'opérateurs linéaires. Par exemple, pour l'opérateur de Monge-Ampère, on utilise la formule suivante valable pour toute $M \in S_d^+$

$$d \det(M)^{\frac{1}{d}} = \inf\{\text{Tr}(DM); D \in S_d^+, \det(D) = 1\}. \quad (51)$$

Différences finies centrées. Les différences finies centrées d'ordre un ne permettent pas *seules* de construire des schémas DDE. Cependant, elles peuvent intervenir dans de tels schémas si elles sont combinées avec des différences d'ordre deux de même support. Elles sont définies par

$$\delta_h^e u(x) := \frac{u(x + he) - u(x - he)}{2h},$$

où $x \in X$, $e \in \mathbb{Z}^d$. Cette expression n'a de sens que si $x + he \in X$ et $x - he \in X$.

Lemme 5.12. *Là où il est défini sur X , le schéma $-\Delta_h^e + \lambda \delta_h^e$ est dégénéré elliptique pourvu que $2|\lambda| h \leq 1$. De plus on a la propriété de consistance $\delta_h^e u(x) = \langle \nabla u(x), e \rangle + \mathcal{O}(h^2)$.*

Corollaire 5.13. *Sous l'hypothèse (49), définissons la discrétisation suivante du gradient*

$$\nabla_h^D u(x) = \sum_{1 \leq i \leq I} \rho_i \delta_h^{e_i} u(x) D^{-1} e_i, \quad \nabla_h^D u(x) = \nabla u(x) + \mathcal{O}(h^2).$$

Si G est K -Lipschitz, et $\varepsilon > 0$, alors l'opérateur suivant est DDE dès que $2Kh \leq \varepsilon$.

$$-\varepsilon \Delta_h^D u + G(\nabla_h^D u).$$

5.3 Itérations d'Euler et de Newton

Nous avons prouvé l'existence de solutions aux schémas DDE §5.1, via la méthode de Perron qui n'est pas constructive. On présente ici deux approches alternatives, les méthodes d'Euler et de Newton, dont l'implémentation informatique est pertinente.

La première méthode fait intervenir une variable de temps supplémentaire, et l'EDO suivante

$$\partial_t u + F(u) = 0$$

dont les états stationnaires satisfont $F(u) = 0$. Un schéma d'Euler explicite pour cette EDO permet d'approcher ces états, sous des hypothèses de régularité et d'ellipticité.

Proposition 5.14 (Méthode d'Euler). *Soit F un schéma δ -elliptique sur un ensemble fini X , et soit $u_0 : X \rightarrow \mathbb{R}$. On définit une suite $(u_n)_{n \geq 0}$ par la relation de récurrence*

$$u_{n+1} := u_n - \varepsilon F(u_n).$$

Si F est K -Lipschitz pour la norme $\|\cdot\|_\infty$ sur l'ensemble $\{\|Fu\|_\infty \leq \|F_{u_0}\|_\infty\}$, alors

$$\|F(u_n)\|_\infty \leq \rho^n \|F(u_0)\|_\infty, \quad \text{où } \rho = \frac{1 + K^2 \varepsilon^2}{1 + \delta \varepsilon}.$$

En particulier, $\|u_n - u_\infty\|_\infty \leq \delta^{-1} \rho^n \|F(u_0)\|_\infty$, où u_∞ est l'unique solution du schéma, par le Corollaire 5.6. En choisissant $0 < \varepsilon < \delta/K^2$ on obtient $\rho < 1$, et donc convergence géométrique.

Proof. Soit $u : X \rightarrow \mathbb{R}$, soit $v := u - \varepsilon Fu$. On suppose qu'il existe $x \in X$ tel que $Fv(x) = \|Fv\|_\infty$ (sans nuire à la généralité, on traiterait de même le cas $Fv(x) = -\|Fv\|_\infty$). Alors par ellipticité de F

$$F(v + \varepsilon Fu)(x) \geq Fv(x) + \varepsilon \delta Fv(x).$$

On en déduit, en utilisant le caractère Lipschitz de F

$$Fu(x) = F(v + \varepsilon Fu)(x) \geq F(v + \varepsilon Fv)(x) - K\varepsilon\|Fu - Fv\|_\infty \geq (1 + \varepsilon\delta)Fv(x) - K^2\varepsilon^2\|Fu\|_\infty.$$

On en déduit $(1 + K^2\varepsilon^2)\|Fu\|_\infty \geq (1 + \delta\varepsilon)\|Fv\|_\infty$, ce qui conclut. \square

La méthode d'Euler souffre de sa convergence relativement lente. En effet, dans les cas d'intérêt, la constante de Lipschitz K est souvent grande et difficile à évaluer, tandis que la constante d'ellipticité δ est souvent nulle ou très petite; donc le pas de temps $\varepsilon = \delta/K^2$ est très faible.

La méthode de Newton est une alternative redoutablement efficace, qui converge généralement en une dizaine d'itérations. L'ellipticité n'est pas directement une hypothèse de cette méthode, voir la Proposition 5.15, mais elle peut servir indirectement à établir l'inversibilité de la matrice jacobienne du schéma. Une limitation théorique - mais dont les effets numériques ne sont pas systématiquement visibles - est la nécessité que la jacobienne de F soit Lipschitz, ce qui n'est pas vérifié par tous les schémas DDE, notamment ceux qui sont construits comme des max ou min de schémas élémentaires, voir les Proposition 5.7, Corollaire 5.9, et (51).

Algorithm 1 Algorithme de Newton avec damping

Entrées: F , u_0 .

Pour tout $n \geq 0$:

Calculer $v_n := [\mathrm{d}F(u_n)]^{-1}F(u_n)$.

Trouver le plus petit $k \geq 0$ tel que $\delta := 2^{-k}$ satisfait $\|F(u_n - \delta v_n)\|_\infty \leq (1 - \delta/2)\|Fu_n\|_\infty$.

Poser $\delta_n := \delta$, $u_{n+1} := u_n - \delta_n v_n$.

Proposition 5.15 (Méthode de Newton). *Soit F un schéma sur un ensemble fini X , et $u_0 : X \rightarrow \mathbb{R}$. Supposons que la matrice Jacobienne de F est inversible et Lipschitz sur le domaine $\{\|Fu\|_\infty \leq \|Fu_0\|_\infty\}$. Alors les itérées $(u_n)_{n \geq 0}$ de l'Algorithme 1 satisfont: (convergence globale linéaire) $\|Fu_n\|_\infty \leq \rho^n$, où $\rho < 1$ dépend seulement des bornes sur les propriétés de F , et (convergence locale quadratique) $\|u_n - u_*\|_\infty \leq C\eta^{2^n}$ pour certaines constantes $\eta < 1$ et C .*

Proof. Un développement de Taylor, avec reste intégral, donne pour tout $\delta > 0$

$$F(u_n - \delta v_n) = Fu_n - \delta[\mathrm{d}F(u_n)]v_n - \delta \int_0^1 [\mathrm{d}F(u_n + t\delta v_n) - \mathrm{d}F(u_n)]v_n dt$$

On en déduit, compte tenu du choix de v_n ,

$$\|F(u_n - \delta v_n)\|_\infty \leq (1 - \delta)\|Fu_n\|_\infty + \frac{1}{2}\delta^2 KL\|Fu_n\|_\infty^2,$$

où K est une borne pour la constante de Lipschitz de $\mathrm{d}F$, et L pour la norme de $[\mathrm{d}F]^{-1}$, sur l'ensemble considéré. En particulier, $\|F(u_n - \delta v_n)\|_\infty \leq (1 - \delta/2)\|Fu_n\|_\infty$ dès que $\delta KL\|Fu_n\|_\infty \leq 1$. On en déduit que l'entier k recherché dans l'Algorithme 1 existe, et que soit $\delta_n = 1$, soit $2KL\|Fu_n\|_\infty \geq 1$. En particulier $\|Fu_n\|_\infty \leq \rho^n\|Fu_0\|_\infty$ où $\rho := 1 - (4KL\|u_0\|_\infty)^{-1}$, ce qui établit la convergence globale linéaire.

Par ailleurs, si $2KL\|Fu_n\|_\infty \leq 1$ pour un certain $n \geq 0$, alors $\delta_n = 1$. Dans ce cas on obtient $KL\|Fu_{n+1}\|_\infty \leq (KL\|Fu_n\|_\infty)^2$, puis $KL\|Fu_{n+k}\|_\infty \leq (KL\|Fu_n\|_\infty)^{2^k}$ pour tout $k \geq 0$ ce qui établit la convergence quadratique locale de $\|Fu_n\|_\infty$ vers 0, donc celle de u_n car $[\mathrm{d}F]^{-1}$ est borné. \square

5.4 Itérations de Jacobi

On présente une approche de la résolution des schémas DDE fondée sur la résolution de problèmes locaux - où l'on recherche la solution du schéma en un point, ayant fixé les valeurs de ses voisins. L'implémentation de cette approche est donc un peu plus complexe que les méthodes de Euler et Newton, mais en contrepartie on se passe de certaines de leurs hypothèses restrictives: l'ellipticité pour Euler, et différentiabilité et nécessité d'inverser des systèmes linéaires de grande taille pour Newton.

A cet effet, on introduit la notion d'*opérateur monotone* sur un ensemble fini X . On note $\mathbb{U} := \mathbb{R}^X$ l'ensemble des fonctions $u : X \rightarrow \mathbb{R}$.

Définition 5.16. *Un opérateur sur un ensemble X est une fonction $\Lambda : \mathbb{U} \rightarrow \mathbb{U}$. Il est dit monotone si pour tous $u, v \in \mathbb{U}$, on a: $u \leq v \Rightarrow \Lambda u \leq \Lambda v$.*

Proposition 5.17. *Soit Λ un opérateur monotone et continu sur X , et soient $\underline{u}, \bar{u} : X \rightarrow \mathbb{R}$ tels que $\Lambda \underline{u} \geq \underline{u}$, $\Lambda \bar{u} \leq \bar{u}$, et $\underline{u} \leq \bar{u}$. Alors les limites suivantes existent et sont point fixes de Λ*

$$\underline{u}_\infty := \lim_{n \rightarrow \infty} \Lambda^n \underline{u}, \quad \bar{u}_\infty := \lim_{n \rightarrow \infty} \Lambda^n \bar{u}, \quad \underline{u} \leq \underline{u}_\infty \leq \bar{u}_\infty \leq \bar{u}$$

Proof. On déduit des hypothèses, par une récurrence immédiate, que pour tout $n \geq 0$

$$\underline{u} \leq \Lambda \underline{u} \leq \cdots \leq \Lambda^n \underline{u} \leq \Lambda^n \bar{u} \leq \cdots \leq \Lambda \bar{u} \leq \bar{u}.$$

Ainsi la limite définissant \underline{u}_∞ (resp. \bar{u}_∞) est croissante (resp. décroissante) donc convergente. Par continuité de Λ on a $\Lambda \underline{u}_\infty = \underline{u}_\infty$ et $\Lambda \bar{u}_\infty = \bar{u}_\infty$ sur X , ce qui conclut. \square

Etant donné un schéma numérique F sur un ensemble fini X , l'opérateur Λ associé est obtenu en résolvant (si c'est possible) $Fu(x) = 0$ en un point donné $x \in X$ par rapport à la variable $u(x)$, les autres valeurs $u(y)$, $y \in X \setminus \{x\}$ étant fixées.

Définition 5.18 (Opérateur de mise à jour). *Soit F un schéma sur un ensemble X . On fait l'hypothèse que pour tout $u : X \rightarrow \mathbb{R}$ et tout $x \in \mathbb{R}$, la quantité suivante*

$$F(x, \lambda, (\lambda - u(y))_{y \in X \setminus \{x\}}). \tag{52}$$

admet une unique racine $\lambda \in \mathbb{R}$, notée $\lambda = \Lambda u(x)$ et appelée mise à jour de Jacobi.

Etant donné un schéma F , et étant connu l'opérateur Λ associé, les équations $Fu = 0$ et $\Lambda u = u$ sont clairement équivalentes. On peut donc reformuler la recherche d'une racine de F en celle d'un point fixe de Λ . La proposition suivante montrer de plus que, si F est DDE, alors Λ est monotone, ce qui permet d'en trouver un point fixe par itération, voir la Proposition 5.17.

Proposition 5.19. *Soit F un schéma et Λ l'opérateur associé par la Définition 5.18, que l'on suppose exister. Si F est dégénéré elliptique, alors Λ est monotone.*

Proof. Soient $u, v : X \rightarrow \mathbb{R}$ telles que $u \leq v$. Si F est DDE alors pour tout $\lambda \in \mathbb{R}$

$$F(x, \lambda, (\lambda - u(y))_{y \in X \setminus \{x\}}) \geq F(x, \lambda, (\lambda - v(y))_{y \in X \setminus \{x\}}).$$

Par ailleurs ces deux quantités sont des fonctions croissantes de λ , donc leurs racines (supposées exister) satisfont $\Lambda u(x) \geq \Lambda v(x)$, comme annoncé. \square

Note: la mise à jour locale est la même pour les itérations de Gauss-Siedel et de Jacobi. La distinction dépend de l'utilisation ou non d'une variable tampon : mises à jour simultanées pour Jacobi (pas de variable tampon), ou successives pour Gauss-Siedel.

Pour conclure, on calcule les opérateurs monotones associés à quelques schémas DDE.

Cas d'un opérateur linéaire. Considérons un schéma F linéaire et δ -elliptique sur l'ensemble $X = \{1, \dots, N\}$. Ainsi $Fu = Au + b$, où A est une matrice de taille $N \times N$, et où $b \in \mathbb{R}^N$. On peut écrire $A = D - M$ où D est une matrice diagonale, M est une matrice ayant des zéros sur la diagonale. Par monotonie de F , les entrées de D et M sont positives, satisfont $D_{ii} \geq \delta + \sum_{1 \leq j \leq N} M_{ij}$ pour tout $1 \leq i \leq N$. Par la Définition 5.18 on a

$$v = \Lambda u \Leftrightarrow Dv = Mu + b \Leftrightarrow v = D^{-1}Mu + D^{-1}b.$$

On reconnaît la méthode de Jacobi de résolution de systèmes linéaires. Sous les hypothèses faites ici, la matrice $D^{-1}M$ a toutes entrées positives, et les sommes de ses lignes bornées strictement par 1. Elle est donc contractante pour $\|\cdot\|_\infty$, de sorte que les itérations convergent géométriquement.

Cas de l'équation eikonale. Considérons le schéma F de discréétisation de l'équation eikonale du Corollaire 5.9. Le problème local résolu pour la mise à jour de Jacobi (52) prend la forme

$$\sum_{1 \leq i \leq I} \rho_i (\lambda - v_i)_+^2 - h^2 = 0, \quad v_i := \min\{u(x + he_i), u(x - he_i)\}$$

Sans nuire à la généralité, on peut supposer $v_1 \leq \dots \leq v_I$. Le membre de gauche est égal à -1 sur $] -\infty, v_1]$, puis strictement croissant et quadratique sur chacun des intervalles $[v_1, v_2], \dots, [v_{I-1}, v_I], [v_I, \infty[$. Le calcul de l'opérateur de mise à jour se ramène donc au tri de I valeurs, puis à la résolution de I équations du second degré.

References

- [AGS08] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science and Business Media, 2008.
- [BCM16] Jean-David Benamou, Francis Collino, and Jean-Marie Mirebeau. Monotone and consistent discretization of the Monge-Ampere operator. *Mathematics of computation*, 85(302):2743–2775, 2016.
- [BR98] Guy Barles and Elisabeth Rouy. A strong comparison result for the Bellman equation arising in stochastic exit time control problems and its applications. *Communications in Partial Differential Equations*, 23(11-12):1995–2033, 1998.
- [BS91] Guy Barles and Panagiotis E Souganidis. Convergence of approximation schemes for fully nonlinear second order equations. *Asymptotic analysis*, 4(3):271–283, 1991.
- [CIL92] Michael G Crandall, Hitoshi Ishii, and Pierre-Louis Lions. User's guide to viscosity solutions of second order partial differential equations. *SIAM Journal on Numerical Analysis*, 27(1):1–67, 1992.
- [CLMC92] Francine Catté, Pierre-Louis Lions, Jean-Michel Morel, and Tomeu Coll. Image selective smoothing and edge detection by nonlinear diffusion. *SIAM Journal on Numerical Analysis*, 29(1):182–193, 1992.
- [CS92] J H Conway and N J A Sloane. Low-Dimensional Lattices. VI. Voronoi Reduction of Three-Dimensional Lattices. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 436(1896):55–68, January 1992.

- [CS13] John Horton Conway and Neil James Alexander Sloane. *Sphere packings, lattices and groups*, volume 290. Springer Science and Business Media, 2013.
- [CWW13] Keenan Crane, Clémence Weischedel, and Max Wardetzky. Geodesics in heat: A new approach to computing distance based on heat flow. *ACM Transactions on Graphics (TOG)*, 32(5):152, 2013.
- [Eva10] Lawrence C Evans. *Partial Differential Equations*. American Mathematical Soc., 2010.
- [FGN13] Xiaobing Feng, Roland Glowinski, and Michael Neilan. Recent Developments in Numerical Methods for Fully Nonlinear Second Order Partial Differential Equations. *SIAM Review*, 55(2):205–267, January 2013.
- [FM14] Jérôme Fehrenbach and Jean-Marie Mirebeau. Sparse non-negative stencils for anisotropic diffusion. *Journal of Mathematical Imaging and Vision*, 49(1):123–147, 2014.
- [JKO98] Richard Jordan, David Kinderlehrer, and Felix Otto. The Variational Formulation of the Fokker–Planck Equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, January 1998.
- [JS13] Boško S Jovanović and Endre Süli. *Analysis of finite difference schemes: for linear partial differential equations with generalized solutions*, volume 46. Springer Science and Business Media, 2013.
- [Kom88] Hidetoshi Komiya. Elementary proof for Sion’s minimax theorem. *Kodai mathematical journal*, 11(1):5–7, 1988.
- [Kry05] Nicolai V Krylov. The rate of convergence of finite-difference approximations for Bellman equations with Lipschitz coefficients. *Applied Mathematics and Optimization*, 52(3):365–399, 2005.
- [KT92] Hung-Ju Kuo and Neil S Trudinger. Discrete Methods for Fully Nonlinear Elliptic Equations. *SIAM Journal on Numerical Analysis*, 29(1):123–135, February 1992.
- [Mir19] Jean-Marie Mirebeau. Riemannian Fast-Marching on Cartesian Grids, Using Voronoi’s First Reduction of Quadratic Forms. *SIAM Journal on Numerical Analysis*, 57(6):2608–2655, 2019.
- [Obe06] A M Oberman. Convergent Difference Schemes for Degenerate Elliptic and Parabolic Equations: Hamilton-Jacobi Equations and Free Boundary Problems. *SIAM Journal on Numerical Analysis*, 44(2):879–895, January 2006.
- [Obe08] A M Oberman. Wide stencil finite difference schemes for the elliptic Monge-Ampere equation and functions of the eigenvalues of the Hessian. *Discrete Contin Dyn Syst Ser B*, 2008.
- [PB13] N Parikh and S Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 2013.
- [PCo19] Gabriel Peyré, Marco Cuturi, and others. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

- [PM90] P Perona and J Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7):629–639, July 1990.
- [RT92] Elisabeth Rouy and Agnès Tourin. A Viscosity Solutions Approach to Shape-From-Shading. *SIAM Journal on Numerical Analysis*, 29(3):867–884, July 1992.
- [Sch09] Achill Schürmann. Computational geometry of positive definite quadratic forms. *University Lecture Series*, 49, 2009.
- [SdGP⁺15] Justin Solomon, Fernando de Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (TOG)*, 34(4):66, 2015.
- [Sel74] Eduard Selling. Ueber die binären und ternären quadratischen Formen. *Journal fur die Reine und Angewandte Mathematik*, 77:143–229, 1874.
- [Var67] S R S Varadhan. On the behavior of the fundamental solution of the heat equation with variable coefficients. *Communications on Pure and Applied Mathematics*, 20(2):431–455, May 1967.
- [Wei98] Joachim Weickert. *Anisotropic diffusion in image processing*, volume 1. Teubner Stuttgart, 1998.