

# EDPs numériques pour l'analyse d'images

Jean-Marie Mirebeau\*

March 14, 2022

## Abstract

L'objectif du cours est de présenter des outils et méthodes permettant la résolution numérique efficace de certaines équations aux dérivées partielles - liées à la diffusion anisotrope, au transport optimal, et au contrôle optimal - en vue d'applications au traitement de l'image.

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
1.1	Pertinence des EDPs en traitement d'image . . . . .	2
1.2	Les EDPs considérées dans ce cours . . . . .	3
<b>2</b>	<b>Définition et propriétés fondamentales</b>	<b>6</b>
2.1	Rappels: différentielle et gradient . . . . .	7
2.2	Préliminaire: coût d'un produit matriciel . . . . .	8
2.3	Trois approches de la différentiation automatique . . . . .	10
2.4	Dérivées d'ordre deux et supérieur . . . . .	11
<b>3</b>	<b>Discrétisation de la diffusion anisotrope et non-linéaire</b>	<b>13</b>
3.1	Flot gradient dans un espace abstrait . . . . .	13
3.2	L'équation de la chaleur en tant que flot gradient . . . . .	15
3.3	Schéma aux différences finies . . . . .	17
3.4	Décomposition matricielle de Selling . . . . .	20
3.5	Non-linéarité . . . . .	22
<b>4</b>	<b>Chaleur et distances géodésiques</b>	<b>24</b>
4.1	Distance riemannienne . . . . .	25
4.2	Transformation logarithmique . . . . .	28
4.3	Solution de viscosité . . . . .	29
4.4	Transport optimal entropique . . . . .	33
<b>5</b>	<b>Schémas monotones</b>	<b>35</b>
5.1	Principe de comparaison, existence d'une solution . . . . .	36
5.2	Exemples . . . . .	37
5.3	Itérations d'Euler et de Newton . . . . .	41
5.4	Itérations de Jacobi/Gauss-Siedel . . . . .	43

---

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

<b>6 Résolution numérique et applications de l'équation eikonale</b>	<b>45</b>
6.1 Convergence vers la solution continue . . . . .	45
6.2 Méthodes itératives . . . . .	47
6.3 Méthode du fast marching . . . . .	50
6.4 Applications de l'équation eikonale . . . . .	51
<b>7 Contrôle non-holonomie</b>	<b>53</b>
7.1 Modèle sous-Riemannien . . . . .	56
7.2 Autres modèles non-holonomes . . . . .	57
7.3 Discrétisation de l'équation eikonale généralisée . . . . .	61
7.4 Applications . . . . .	63

## 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 discutons dans cette section de leur pertinence dans le cadre du traitement de l'image, et donnons 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 le 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) [RT92].
- *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 [Lin13].  
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 [CMC17].
- *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 solution d'un tel système est un objet global, qui synthétise les contraintes imposées localement.  
Exemple : La résolution de l'équation eikonale permet de déterminer le plus court chemin dans un domaine contenant des obstacles et des zones plus ou moins rapides, voir Fig. 1 et §6.

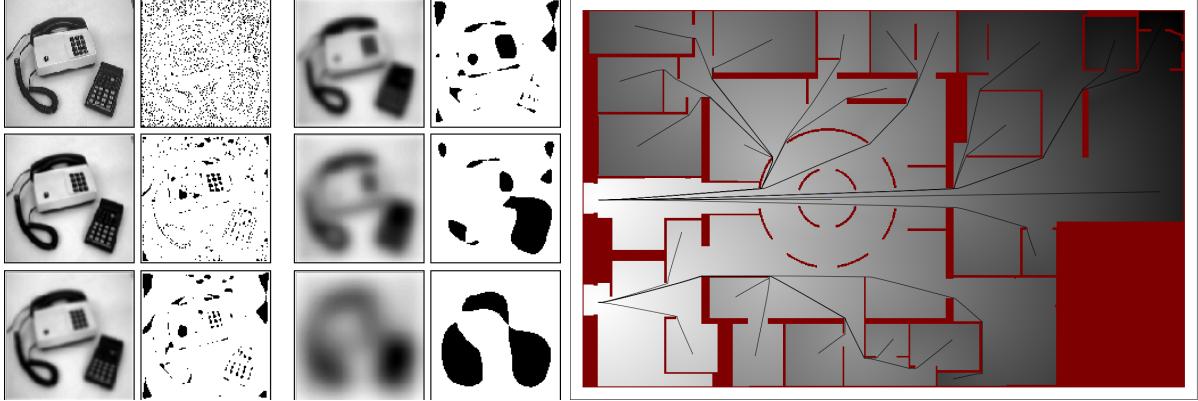


Figure 1: (Gauche) Equation de la chaleur et représentation multi-échelle d'une image [Lin13]. Crédit image: Tony Lindeberg. (Droite) Solution de l'équation eikonale (niveaux de grid) dans un domaine avec obstacles (rouge), et chemins minimaux.

- *Structure mathématique, garanties.* Les EDPs ont une structure mathématique riche qui permet d'établir un certain nombre de garanties : terminaison de l'algorithme de résolution, 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 discréétisation, 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 eikonale, 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. Leur application au traitement de l'image est décrite dans les références classiques [Lin13, PPKC10]. Des exemples de solutions sont donnés Fig. 1.

- L'équation de la chaleur s'interprète comme le flot gradient (descente de gradient) de l'énergie de Dirichlet  $\int_{\Omega} \|\nabla u\|^2$ , voir §3.2. Elle lissee 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, voir Fig. 1 (gauche).
- 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. Voir §6.

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  $\Delta u$  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 montre par un développement de Taylor les propriétés de consistance:

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

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

Ces deux schémas aux différences finies mènent aux discrétisations classiques sur grille cartésienne de l'équation de la chaleur et de l'équation eikonale [RT92]. Leur structure les rend monotones, plus précisément : *dégénérés elliptiques*, ce qui permet de les résoudre de manière stable et efficace, voir §5.

**Anisotropie.** On dit qu'un problème est anisotrope lorsque les directions d'espace dans le domaine géométrique associé 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 [All92]. Exemple : ondes sismiques dans un milieu stratifié, ondes lumineuses dans un crystal, ... L'anisotropie issue d'une micro-structure peut aussi être détectée et traitée artificiellement. En traitement de l'image, il est ainsi courant de détecter les directions préférentielles de textures oscillantes (comme dans une image d'empreinte digitale) pour les traiter de manière appropriée, par exemple à l'aide du tenseur de structure [Wei98, §2.2] et §3.5. Illustration Fig. 2.
- *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  $(x_0, x_1, \theta)$  seulement les déplacements engendrés par  $(\cos \theta, \sin \theta, 0)$  et  $(0, 0, 1)$ . Voir Fig. 3 et §7.
- *Proximité des bords ou de fissures.* La modélisation des écoulements au bord d'un domaine, ou de l'élasticité au voisinage d'une fissure [CMA<sup>+</sup>20], 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 l'équation de Monge-Ampère  $\det \nabla^2 u = f$  liée entre autres au problème du transport optimal [Fig17], 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  $\nabla^2 u$ .



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

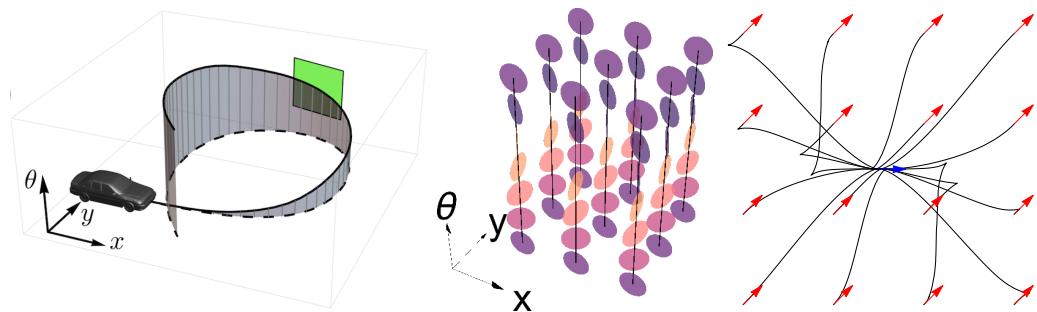


Figure 3: Exemple d'anisotropie liée au rôle distinct des dimensions du domaine. Ici le véhicule de Reeds-Shepp, dont l'espace d'états est  $\mathbb{R}^2 \times \mathbb{S}^1$ . (Gauche) Une trajectoire admissible. (Centre) Boule unité dans l'espace des vitesses admissibles. (Droite) Projection planaire de quelques trajectoires optimales. Voir §7.

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 eikonale 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, gauche) ou la propagation du front (eikonal, droite) dans les directions associées aux grandes valeurs propres de  $D$ , et sont naturelles dans le cadre de la géométrie Riemannienne. Nous verrons comment discréteriser les EDPs anisotropes (4) aux § 3.3 et 5.2, en généralisant les approches classiques isotropes (2) et (3), et en identifiant et conservant leurs bonnes propriétés.

## 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*, *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 introduction à la différentiation automatique [GW08].

**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 façon alternative mais liée d'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 discréterisation des équations aux dérivées partielles, tels que (2) et (3). Leur structure mathématique simple et (en général) linéaire permet d'étudier mathématiquement les quantités qui en sont dérivées, voir par exemple §5.

*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.)

**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 [GW08].

*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. Un grand nombre de bibliothèques logicielles proposent des implémentations de qualité (PyTorch®, TensorFlow®, etc).

*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, voir §2.3.

**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 proscrire<sup>1</sup> 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 est nécessaire, ses résultats doivent être contrôlés sur des exemples par une approche alternative<sup>2</sup>.

## 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 lorsque  $h \rightarrow 0$

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

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

---

<sup>1</sup>Rappelons 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.

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

**Définition 2.3** (Gradient). Soit  $f : \Omega \rightarrow \mathbb{R}$ , où  $\Omega$  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

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

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  $x \in \mathbb{H}$ .

**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  $Df|_x$  de taille  $n \times m$ , 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'ambiguïtés (qui pourraient par exemple être liées aux choix de bases) on ne se gênera 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 on a  $D(g \circ f)|_x = Dg|_{f(x)} Df|_x$  sous les hypothèses adéquates de différentiabilité.

## 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

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

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

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

Des algorithmes plus efficaces existent [DS13], mais ils ne sont pas forcément adaptés à la différentiation automatique, ni pertinents dans cette analyse heuristique de sa complexité.

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

$$((A_1 A_2) A_3) \cdots A_n = A_1 \cdots A_n = A_1 (A_2 (\cdots A_n)). \tag{6}$$

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

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

Ces coûts ne sont en général pas égaux, bien qu'ils correspondent au calcul du même résultat, par associativité du produit matriciel. 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.

Il existe de nombreux autres parenthésages possibles du produit associatif (6), et leur ensemble est en bijection avec les arbres binaires à  $n$  feuilles. 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 contient au plus  $\alpha$  (resp.  $\beta$ ) 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 la valeur 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 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. Sous ces conditions, il est plus favorable que le produit dense (7).

**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

$$D(f_1 \circ \cdots \circ f_n)|_{x_n} = Df_1|_{x_1} \cdots Df_n|_{x_n}. \tag{9}$$

Les programmes informatiques décrivent des fonctions complexes comme composées de fonctions élémentaires. Leur différentiation automatique peut donc s'interpréter comme le produit

matriciel des matrices Jacobiennes des étapes intermédiaires. Noter que l’implémentation informatique de la différentiation automatique, décrite section suivante §2.3, peut paraître éloignée de ce formalisme mathématique, toutefois nécessaire pour son étude [GW08, §3.2]. Selon la structure des facteurs du produit (9), on préfèrera utiliser un produit associatif à gauche, ou à droite, ou un produit creux; voire une combinaison de ces approches suivant les sous-facteurs.

### 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 §2.2 pour le calcul d’un produit matriciel: associativité à droite, à gauche, ou produit creux respectivement, 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. Le lecteur souhaitant aller plus loin sur ce thème est renvoyé vers [GW08].

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

$$f : \mathbb{R}^m \rightarrow \mathbb{R}^n, \quad (10)$$

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 jacobien, on note que  $(n, m)$  correspond à  $(I_0, I_n)$ .)

#### 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 d’entrée.

*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

$$x + \langle v, h \rangle + o(h). \quad (11)$$

La variable  $h \in \mathbb{R}^m$  représente ici une perturbation *symbolique*, qui n’a pas d’existence dans le programme informatique. Les règles usuelles de calcul des développements limités s’appliquent, par exemple:

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

La surcharge des opérateurs et des fonctions usuelles permet d’appliquer ces règles de calcul sans avoir à ré-écrire la fonction. Ainsi la fonction sinus et le produit s’étendent aux paires de variables  $(x, v)$  représentant un développement limité du premier ordre (11) par  $\sin(x, v) = (\sin x, v \cos x)$  et  $(x, v)(x', v') = (xx' + xv' + vx')$ .

### Différentiation automatique creuse ( $f$ simple).

*Analogue matriciel.* La différentiation automatique creuse correspond au produit matriciel creux, qui est évalué 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 dimension.

*Utilisation.* La différentiation automatique creuse est idéale pour assembler les matrices jacobienes de 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 triplets  $(x, \alpha, i) \in \mathbb{R} \times \mathbb{R}^K \times \{1, \dots, m\}^K$  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. Un développement limité donne, par exemple:

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

De nouveau, la surcharge des opérateurs et des fonctions usuelles permet d'appliquer ces règles de calcul des développements limités sans avoir à ré-écrire la fonction. Ainsi  $\sin(x, \alpha, i) = (\sin x, \alpha \cos x, i)$ . De manière plus intéressante, 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 et en sommant 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 [GW08]. Des libraries comme PyTorch® et Tensorflow® implémentent ces techniques.

## 2.4 Dérivées d'ordre deux et supérieur

Les définitions et techniques de différentiation présentées ci-dessus s'étendent aux dérivées d'ordre deux, voire supérieur. On notera cependant que leur coût peut augmenter rapidement si la dimension de l'espace de départ est grande.

## Rappels : définitions et propriétés élémentaires.

**Définition 2.7** (Différentielle d'ordre supérieur). Soient  $E, F$  des evn, et soit  $f : \Omega \rightarrow F$  où  $\Omega \subseteq E$  est ouvert, différentiable en tout point de  $\Omega$ . 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  $\mathcal{L}(E_0, \mathcal{L}(E_1, F))$  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 notée  $D^2 f|_x$ .

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<sup>3</sup>.
- *Définition 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 [AF11], ce qui requiert une seconde différentiation automatique.

---

<sup>3</sup>On rappelle que la méthode de Newton nécessite la dérivée d'ordre 1 (différentielle ou matrice jacobienne) pour résoudre une équation du type  $f(x) = y$ , et d'ordre 2 (matrice hessienne) pour minimiser une fonction  $g$ , car cela revient à résoudre l'équation  $\nabla g(x) = 0$ .

- *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  $\Omega$ ) tel que pour tout  $v \in H^1(\Omega)$  on ait

$$\int_{\Omega} \nabla u \nabla v = \int_{\Omega} f v. \quad (12)$$

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. En effet l'intégrale (12, gauche) est une quantité d'ordre deux en la solution inconnue  $u$ , et en la fonction test  $v$ , qui sont toutes deux des variables symboliques.

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 les 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 ce chapitre, on s'intéresse à l'équation de la chaleur anisotrope:

$$\partial_t u = \operatorname{div}(D \nabla u), \quad (13)$$

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 historiques<sup>4</sup> en traitement d'image [Wei98].

#### 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  $\varepsilon > 0$*

$$\operatorname{prox}_{\mathcal{E}}^\varepsilon(x) := \operatorname{argmin}_{y \in X} \frac{1}{2\varepsilon} d_X(x, y)^2 + \mathcal{E}(y). \quad (14)$$

---

<sup>4</sup>L'approche EDP n'est plus l'état de l'art pour des tâches comme le débruitage d'image [XXC12].

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  $y := \text{prox}_{\mathcal{E}}^\varepsilon(x)$ . Alors en différentiant (14, 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. Un second cas particulier à garder en tête, opposé à l'hypothèse de différentiabilité précédente, est celui où  $\mathcal{E}$  est la fonction caractéristique d'un fermé  $E$  (valeur nulle sur  $E$ , infinie sur  $X \setminus E$ ), et donc où  $\text{prox}_{\mathcal{E}}^\varepsilon$  s'apparente à une projection sur l'ensemble  $E$ .

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*

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

*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. Dans le résultat suivant, on établit l'existence d'un flot gradient et d'un minimiseur pour (14) sous des hypothèses de compacité. On dit qu'une partie d'un espace métrique est relativement compacte si son adhérence est compacte.

**Proposition 3.3** (Existence). *Sous les hypothèses de la Définition 3.2. Supposons de plus que  $\mathcal{E}$  est semi-continue inférieurement (s.c.i.), 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 (14) 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 (14) 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  $k \geq 0$

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

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  $0 \leq m \leq n$

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

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

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

Il s'agit d'une propriété d'équi-continuité des applications  $\mathbf{x}^\varepsilon$  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 convexité, comme dans le résultat suivant, généralement considérées comme plus fortes que les hypothèses de compacité précédemment utilisées dans la preuve d'existence. (On pourrait également montrer l'existence de l'opérateur proximal (14) sous ce type d'hypothèses: si  $X$  est un Hilbert et que  $\mathcal{E}$  est s.c.i convexe.)

**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 (14) 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'analogie 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.  $\square$

### 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, qui est la plus classique, voir par exemple [Bre11]. 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

$$\mathbb{L}^2 := \mathbb{L}^2(\Omega), \quad \langle u, v \rangle_{\mathbb{L}^2} = \int_{\Omega} u(x)v(x) dx. \quad (15)$$

On rappelle que  $\mathbb{L}^2(\Omega)$  désigne l'espace de Hilbert des fonctions mesurables de carré intégrable, modulo la relation d'égalité presque partout, et que  $\mathbb{H}^1(\Omega)$  désigne celles dont le gradient (défini au sens des distributions) a ces mêmes propriétés.

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

$$\mathcal{E}(u) := \frac{1}{2} \mathcal{Q}(u, u), \quad \text{où } \mathcal{Q}(u, u) := \int_{\Omega} \|\nabla u(x)\|_{D(x)}^2 dx. \quad (16)$$

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

$$\mathcal{Q}(u, v) = \frac{\mathcal{Q}(u + v, u + v) - \mathcal{Q}(u - v, u - v)}{4}. \quad \left( \text{Ici } \mathcal{Q}(u, v) = \int_{\Omega} \langle \nabla u(x), D(x) \nabla v(x) \rangle dx. \right) \quad (17)$$

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

Dans les paragraphes suivants, 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  $u \in \mathbb{H}^1(\Omega)$

$$c_{\min}\|\nabla u\|_{\mathbb{L}^2(\Omega)}^2 \leq \mathcal{E}(u) \leq c_{\max}\|\nabla u\|_{\mathbb{L}^2(\Omega)}^2. \quad (18)$$

Les *injections de Sobolev* [Bre11] 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 (18) 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. En effet, notons que  $\mathcal{E}$  satisfait pour tous  $u, v \in \mathbb{H}^1(\Omega)$ , et tout  $t \in [0, 1]$ , comme toute forme quadratique

$$\mathcal{E}((1-t)u + tv) = (1-t)\mathcal{E}(u) + t\mathcal{E}(v) - t(1-t)\mathcal{E}(u - v). \quad (19)$$

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

**Gradient de l'énergie de Dirichlet.** Pour  $u$  et  $h$  suffisamment régulières, on calcule

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

Ainsi, compte tenu de la définition (16) de l'énergie de Dirichlet  $\mathcal{E}$  via la forme quadratique  $\mathcal{Q}$ ,

$$\mathrm{d}\mathcal{E}|_u(h) = \mathcal{Q}(u, h) = \langle h, -\operatorname{div}(D \nabla u) \rangle_{\mathbb{L}^2(\Omega)} + \int_{\partial\Omega} h \langle \mathbf{n}, D \nabla u \rangle, \quad (20)$$

où  $\mathbf{n}$  désigne la normale extérieure au domaine  $\Omega$ .

Si la condition de Neumann  $\langle \mathbf{n}, D \nabla u \rangle = 0$  sur  $\Omega$  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  $\mathbb{L}^2(\Omega)$

$$\nabla_u \mathcal{E}(u) = -\operatorname{div}(D \nabla_x u). \quad (21)$$

Cela justifie l'interprétation de (13) comme flot gradient de (16) 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, alors (20, 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  $\mathcal{Q}$ .
- $u \in \mathbb{L}^2$ ,  $v \in C^2(\overline{\Omega})$  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 (20).

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 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 discréttisation  $h > 0$ , et on introduit le domaine discrétisé et le produit scalaire

$$\Omega_h := \Omega \cap h\mathbb{Z}^d, \quad \langle u, v \rangle_h := h^d \sum_{x \in \Omega_h} u(x)v(x), \quad (22)$$

où  $u, v : \Omega_h \rightarrow \mathbb{R}$ . Posons également  $\|\eta\|_h := \sqrt{\langle \eta, \eta \rangle_h}$ . On introduit également une forme quadratique dédiée à l'approximation de l'énergie de Dirichlet (16). Pour tout  $u : \Omega_h \rightarrow \mathbb{R}$ , on pose  $\mathcal{E}_h(u) := \frac{1}{2}\mathcal{Q}_h(u, u)$  où

$$Q_h(u, u) := h^d \sum_{x \in \Omega_h} Q_h^x(u, u), \quad \text{et où } Q_h^x(u, u) := \|\nabla u(x)\|_{D(x)}^2 + \mathcal{O}(h^2). \quad (23)$$

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 (17).

On utilisera en certaines occasions les matrices  $I_h$  et  $Q_h$  du produit scalaire  $\langle \cdot, \cdot \rangle_h$  et de la forme quadratique  $Q_h$ , qui satisfont par définition aux relations suivantes: pour tous  $u, v : \Omega_h \rightarrow \mathbb{R}$

$$u^\top I_h v = \langle u, v \rangle_h, \quad u^\top Q_h v = Q_h(u, v). \quad (24)$$

Noter que  $I_h := h^d \text{Id}$  compte tenu de la discréétisation utilisée (22) du produit scalaire.

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

$$Q_h^x(u, u) = \frac{d(x)}{2h^2} \sum_{1 \leq i \leq d} \left[ (u(x + he_i) - u(x))^2 + (u(x - he_i) - u(x))^2 \right], \quad (25)$$

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 (23, 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,

$$D(x) = \sum_{1 \leq i \leq I} \rho_i(x) e_i e_i^\top \quad \text{où } \rho_i(x) \geq 0, \quad e_i \in \mathbb{Z}^d, \quad \forall 1 \leq i \leq I. \quad (26)$$

On montre par ailleurs que les poids  $\rho_i$  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

$$Q_h^x(u, u) = \frac{1}{2h^2} \sum_{1 \leq |i| \leq I} \rho_i(x) (u(x + he_i) - u(x))^2, \quad (27)$$

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

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

**Gradient de l'énergie de Dirichlet discréétisée.** Dans le cadre continu, nous avons vu que l'équation de la chaleur (13) s'identifie au flot gradient (descente de gradient en temps continu) de l'énergie de Dirichlet (16), dont le gradient est explicitement calculé en (21). Nous mimons ici ce calcul dans le cadre discréétisé.

Rappelons qu'un gradient est toujours défini en termes d'une fonctionnelle à différentier, *et d'un produit scalaire*, voir la Définition 2.3. Dans le cadre de la discrétisation d'EDPs, le produit scalaire à utiliser sur l'espace des solutions  $\mathbb{L}^2(\Omega_h)$  est (22), qui approche le cadre continu (15), et non le produit scalaire euclidien usuel. Cette distinction est assez bénigne (bien que nécessaire) dans le cas des différences finies car  $\langle \cdot, \cdot \rangle_h$  est proportionnel au produit scalaire usuel, mais elle peut poser de véritables difficultés numériques<sup>5</sup> dans le cas des éléments finis (non-étudié ici).

Concrètement, le gradient  $g = \nabla \mathcal{E}_h(u) \in L^2(\Omega_h)$  de l'énergie  $\mathcal{E}_h$  au point  $u \in L^2(\Omega_h)$ , *par rapport au produit scalaire*  $\langle \cdot, \cdot \rangle_h$ , est caractérisé par le développement de Taylor:

$$\mathcal{E}_h(u + \eta) = \mathcal{E}(u) + \langle g, \eta \rangle_h + \mathcal{O}(\|\eta\|_h^2)$$

où  $\eta$  est une petite perturbation. En termes matriciels

$$\frac{1}{2}(u + \eta)^\top Q_h(u + \eta) = \frac{1}{2}u^\top Q_h u + g^\top I_h \eta + \mathcal{O}(\|\eta\|_h^2),$$

Rappelons que  $I_h$  est la matrice du produit scalaire  $\langle \cdot, \cdot \rangle_h$ , et  $Q_h$  celle de la forme quadratique  $Q_h$ . Ainsi par identification des termes du premier ordre, sans oublier la bilinéarité et la symétrie de  $Q_h$  et  $I_h$ , nous obtenons

$$\nabla \mathcal{E}_h(u) = g = I_h^{-1} Q_h u. \quad (28)$$

**Opérateur proximal discrétisé.** L'opérateur proximal est défini en termes de l'énergie  $\mathcal{E}_h$  d'intérêt, *et de la distance sur l'espace de fonctions*, ici  $L^2(\Omega_h)$  muni de  $\|\cdot\|_h$ . Calculons le

$$\begin{aligned} \text{prox}_{\mathcal{E}_h}^\varepsilon(u) &:= \underset{v \in L^2(\Omega_h)}{\operatorname{argmin}} \frac{1}{2\varepsilon} \|u - v\|_h^2 + \mathcal{E}_h(v), \\ &= \underset{v \in L^2(\Omega_h)}{\operatorname{argmin}} \frac{1}{2\varepsilon} (u - v)^\top I_h (u - v) + \frac{1}{2} v^\top Q_h v, \\ &= \underset{v \in L^2(\Omega_h)}{\operatorname{argmin}} \frac{1}{2} v^\top (I_h + \varepsilon Q_h) v - v^\top I_h u, \\ &= (I_h + \varepsilon Q_h)^{-1} I_h u. \end{aligned} \quad (29)$$

**Condition de Courant-Freidrichs-Levy (CFL).** La discrétisation de l'équation de la chaleur par un schéma explicite, avec un pas de temps  $\tau > 0$ , s'écrit compte tenu de (28)

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

où l'on a noté  $\text{Id}$  la matrice identité. Ce schéma est stable en norme  $\mathbb{L}^2(\Omega_h)$  pourvu<sup>6</sup> que  $\text{Id} - \tau I_h^{-1} Q_h$  ait toute ses valeurs propres dans  $[-1, 1]$ . (Noter que le schéma implicite, défini par  $u_{n+1} = \text{prox}_{\mathcal{E}_h}^\tau u_n = (I_h + \tau Q_h)^{-1} I_h u$ , est pour sa part inconditionnellement stable.)

La stabilité du schéma explicite s'écrit de manière équivalente

$$\tau Q_h \preceq 2I_h,$$

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

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

<sup>5</sup>L'inversion de la matrice du produit scalaire, aussi appelée matrice de masse, est relativement coûteuse car celle-ci est non-diagonale, et requise en de nombreuses occasions.

<sup>6</sup>Sous la même condition, l'énergie définie par  $Q_h$  (ou toute puissance de celle-ci) décroît.

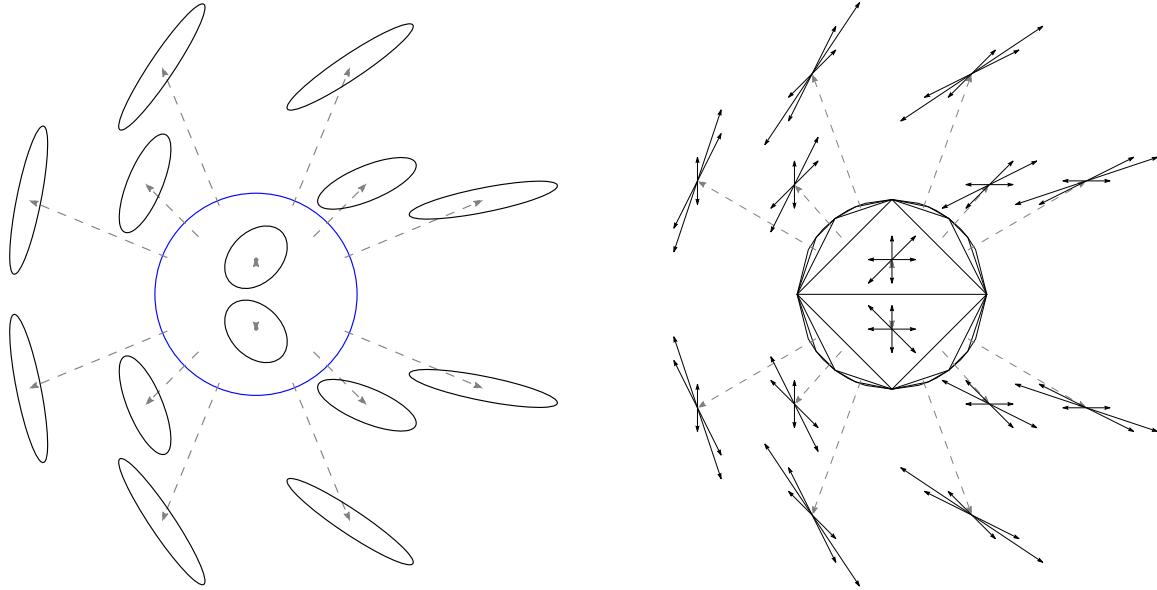


Figure 4: (Gauche) Ellipse définie par  $\{v \in \mathbb{R}^2; \langle v, D(x, y)v \rangle = 1\}$ , où  $D(x, y)$  est définie par la paramétrisation de Pauli (31). (Droite) Superbase  $D(x, y)$ -obtuse, et son opposée.

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  $u(x)^2$  :

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

donc en termes matriciels  $\mathbf{Q}_h \preceq Ch^{-2} \mathbf{I}_h$  puisque  $u$  est arbitraire, où

$$C := \max_{x \in \Omega_h} \sum_{1 \leq |i| \leq I} (\rho_i(x) + \rho_i(x + he_i)), \quad (\text{CFL : } C\tau \leq 2h^2). \quad (30)$$

Pour majorer la constante (30, gauche) on note que

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

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  $C \lesssim 4 \max\{\text{Tr}(D(x)); x \in \bar{\Omega}\}$ .

**Positivité de la solution.** Par construction, la matrice  $\mathbf{Q}_h$  a toutes ses entrées négatives hors de la diagonale, et toutes ses entrées diagonales bornées par  $Ch^{-2}$ , où  $C$  est défini par (30). Sous la condition CFL (30), la matrice  $I - \tau I_h^{-1} \mathbf{Q}_h$  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  $D \in S_d^{++}$ , ssi  $\langle v_i, Dv_j \rangle \leq 0$  pour tous  $i \neq j$ .*

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

$$D(x, y) := \begin{pmatrix} 1+x & y \\ y & 1-x \end{pmatrix}, \quad \text{où } x^2 + y^2 < 1. \quad (31)$$

**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  $0 \leq k \leq d$ .*

Les vecteurs  $(e_{ij})$ , définis ci-dessus par dualité, admettent des expressions plus simples et explicites en dimension 2 et 3. En dimension  $d = 2$ , si  $\{i, j, k\} = \{0, 1, 2\}$ , on vérifie que  $e_{ij} = \pm v_k^\perp$ . En dimension  $d = 3$ , si  $\{i, j, k, l\} = \{0, 1, 2, 3\}$ , on vérifie que  $e_{ij} = \pm v_k \wedge v_l$ .

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

$$D = - \sum_{i < j} \langle v_i, Dv_j \rangle e_{ij} e_{ij}^\top. \quad (32)$$

*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 (32) 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 (26). Selling montre l'existence d'une superbase satisfaisant cette propriété géométrique de manière constructive, via l'algorithme présenté ci-dessous, dont l'efficacité pratique est suffisante pour les applications considérées en EDP.

**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  $\{i, j, k, l\} = \{0, 1, 2, 3\}$ )*

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

Cet algorithme termine, et la superbase finale 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.  $\square$

On appelle *décomposition de Selling* la formule de Selling (32) appliquée à une matrice  $D \in S_d^{++}$  et à une superbase  $D$ -obtuse  $(v_0, \dots, v_d)$ , obtenue typiquement via l'algorithme de Selling. On montre que la décomposition de Selling ne dépend pas du choix de la superbase  $D$ -obtuse.

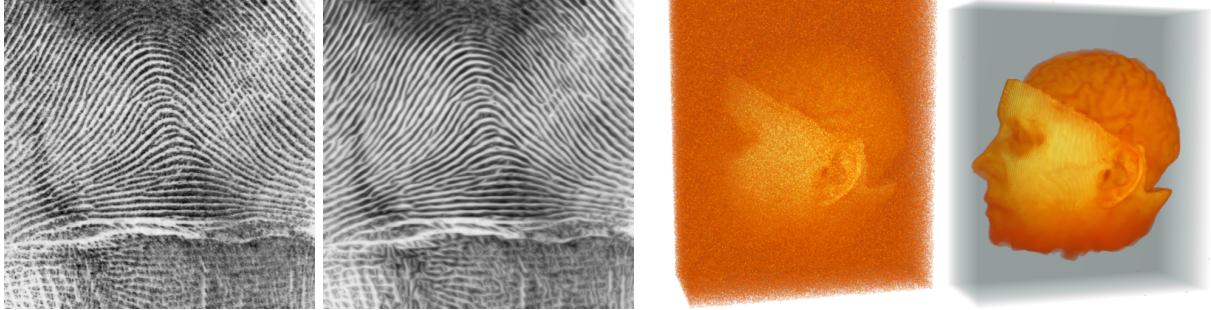


Figure 5: Avant/après l'effet de  $\partial_t u = \operatorname{div}(D_u \nabla u)$ , avec coefficients de diffusion non-linéaires et anisotropes  $D[\nabla u]$  (Coherence enhancing diffusion [Wei98]).

**Interprétation et généralisation par la programmation linéaire.** Voronoi définit la première réduction [Sch09] d'une matrice  $D \in S_d^{++}$ , en d'autres termes d'une forme quadratique positive, par le programme linéaire suivant:

$$\min_{M \in S_d} \operatorname{Tr}(DM), \quad \text{sous contrainte que } \langle e, Me \rangle \geq 1, \forall e \in \mathbb{Z}^d \setminus \{0\}. \quad (33)$$

A toute superbase  $b = (v_0, \dots, v_d)$ , associons  $M_b := \frac{1}{2} \sum_{0 \leq i \leq d} v_i v_i^\top$ . On peut montrer que la matrice  $M_b$  satisfait (33, droite), et que si  $b$  est  $D$ -obtuse, alors elle résout (33). 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 (33).

Par ailleurs, le programme linéaire dual à (33) prend la forme d'un problème de décomposition matricielle

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

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:

$$\partial_t u = \operatorname{div}(D[\nabla u] \nabla u), \quad (34)$$

où  $D[\nabla u](t, x)$  est une matrice symétrique semi-définie positive, dépendant de manière non-linéaire de  $\nabla u(t, \cdot)$ . L'étude mathématique, l'implémentation numérique, et les applications de l'équation de la chaleur 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.

- 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, voir le tenseur de structure (36). 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 sur la forme de  $D[\nabla u]$  (anisotropie, propriétés de monotonie, ...) sont moins sévères également dans ce cas.

- Diffusion *isotrope*<sup>7</sup> 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:

$$D[\nabla u](x) = g(|\nabla u(x)|) \text{Id}. \quad (35)$$

Sous cette hypothèse, l'équation de la chaleur non-linéaire (34) se simplifie en

$$\partial_t u = \operatorname{div}(g(|\nabla u|)\nabla u).$$

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 (34) s'interprète dans le cas (35) comme le flot gradient de la fonctionnelle

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

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, pour lequel on a existence et unicité de la solution de l'équation de la chaleur non-linéaire (34), dans l'esprit de la Proposition 3.4, voir [ABCM01] pour une preuve détaillée. En revanche l'énergie associée au modèle de Perona et Malik n'est pas convexe, et en effet l'unicité de la solution est perdue. Ce phénomène s'illustre visuellement par des instabilités menant à des artefacts *en marche d'escalier*.

---

<sup>7</sup>Une certaine confusion de terminologie règne dans la littérature [SKB01], venant du fait que l'équation de diffusion (34) peut être reformulée sous forme non-divergence :  $\partial_t u = \langle \operatorname{div}(D[\nabla u]), \nabla u \rangle + \operatorname{Tr}(D[\nabla u]\nabla^2 u)$ , or en insérant l'expression (35) et en appliquant l'opérateur divergence au premier terme on obtient une expression  $\langle \operatorname{div}(D[\nabla u]), \nabla u \rangle = g'(|\nabla u|)\langle \nabla u, \nabla^2 u \nabla u \rangle / |\nabla u|^2$  qui peut être qualifiée d'anisotrope. Dans ce document l'isotropie/anisotropie fait référence à la forme de tenseur de diffusion, lorsque l'EDP est sous forme divergence.

**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  $\sigma$  désigne l'échelle du bruit dans l'image, et  $G_\sigma$  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, présentées<sup>8</sup> 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 sécrit

$$S[\nabla u](x) := G_\rho \star (\nabla u_\sigma \nabla u_\sigma^\top), \quad \text{où } u_\sigma := G_\sigma \star u. \quad (36)$$

Noter que  $\nabla u_\sigma = (\nabla G_\sigma) \star u = G_\sigma \star (\nabla u)$  ne dépend que du gradient de  $u$ . Le paramètre  $\sigma$  correspond à l'échelle de cohérence du bruit (noise scale), et  $\rho$  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 :

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

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<sup>+</sup>15].

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

$$u_\varepsilon - \varepsilon^2 \operatorname{div}(D\nabla u_\varepsilon) = u_0, \quad \left( \text{resp. } \partial_t u = \operatorname{div}(D\nabla u), \quad u|_{t=0} = u_0, \right) \quad (37)$$

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 bord<sup>9</sup> par exemple. Compte tenu de leur caractère linéaire, les solutions de ces équations admettent une représentation intégrale :

$$u_\varepsilon(x) = \int_{\Omega} \mathcal{P}_\varepsilon(x, y) u_0(y) dy, \quad \left( \text{resp. } u(t, x) = \int_{\Omega} \mathcal{Q}_t(x, y) u_0(y) dy \right) \quad (38)$$

---

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

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

Le noyau de Poisson (resp. de la chaleur) est lié à la distance géodésique  $d_M$  sur  $\bar{\Omega}$  associée aux champ de tenseurs inverse  $M = D^{-1}$ , définie en (40). En effet, sous des hypothèses adéquates [Var67], on a pour  $x, y \in \Omega$

$$\mathcal{P}_\varepsilon(x, y) = \exp\left(-\frac{d_M(x, y) + o(1)}{\varepsilon}\right). \quad \left(\text{resp. } \mathcal{Q}_t(x, y) = \exp\left(-\frac{d_M(x, y)^2 + o(1)}{4t}\right).\right) \quad (39)$$

En exploitant cette relation, des méthodes numériques permettent d'estimer des distances géodésiques via la résolution des EDPs linéaires (37).

**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  $C^\infty$ . On suppose également  $D \in C^\infty(\bar{\Omega}, S_d^{++})$ , 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 [CLPQ20], comme la résolution directe de l'équation eikonaile que l'on verra § 6, ou les méthodes de tir fondées sur l'équation d'Euler-Lagrange Proposition 4.5, 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 quadrangulaires, etc)*. En effet, une discréétisation 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 discréétisation 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 eikonaile par un terme d'ordre deux (49).
- Une implémentation du transport optimal [SdGP+ 15], rendue possible par le calcul efficace des convolutions (38).

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

## 4.1 Distance riemannienne

On rappelle la définition de la distance géodésique, sur un ouvert  $\Omega$  de  $\mathbb{R}^d$  muni d'une métrique riemannienne  $M : \Omega \rightarrow S_d^{++}$ , sujets à l'Hypothèse 4.1, et quelques propriétés de régularité. On discute également des équations différentielles ordinaires satisfaites par les géodésiques riemannniennes : l'équation de backtracking qui fait intervenir la fonction distance (44), et l'équation autonome d'Euler-Lagrange Proposition 4.5. La distance riemannienne séparant  $x, y \in \bar{\Omega}$  est définie par

$$d_M(x, y) := \min_{\gamma \in \Gamma_x^y} \int_0^1 \|\gamma'(t)\|_{M(\gamma(t))} dt, \quad (40)$$

où  $\Gamma_x^y$  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  $x \in \bar{\Omega}$  par

$$v_0(x) := \min_{y \in \partial\Omega} d_M(x, y). \quad (41)$$

On étend par convention  $v_0$  à  $\mathbb{R}^d$  par la valeur nulle. 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 (40) et (41).*

*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  $\Omega$ , 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 (40) 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 (40) 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 (41) 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 (41), paramétré à vitesse riemannienne unité avec  $T := v_0(x)$ ,  $\gamma(0) = x$ ,  $\gamma(T) \in \partial\Omega$ . Alors*

$$d_M(x, \gamma(t)) = t, \quad v_0(\gamma(t)) = v_0(x) - t \tag{42}$$

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

$$\gamma(t) \in \operatorname{argmin}_{y \in \overline{\Omega}} \frac{1}{2t} d_M(x, y)^2 + v_0(y) \tag{43}$$

*Proof.* (Preuve de (42).) Soit  $y_* \in \partial\Omega$  le minimiseur de (41), de sorte que  $d_M(x, y_*) = v_0(x) = T$ . Comme  $\gamma$  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 (43).) 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  $\delta$ . 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 (43), que  $\gamma(\varepsilon) = \operatorname{prox}_{v_0}^\varepsilon(x)$ . En d'autres termes, le chemin optimal  $\gamma$  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

$$\gamma'(t) = -V(\gamma(t)), \quad \text{où } V(x) := D(x)\nabla v_0(t). \tag{44}$$

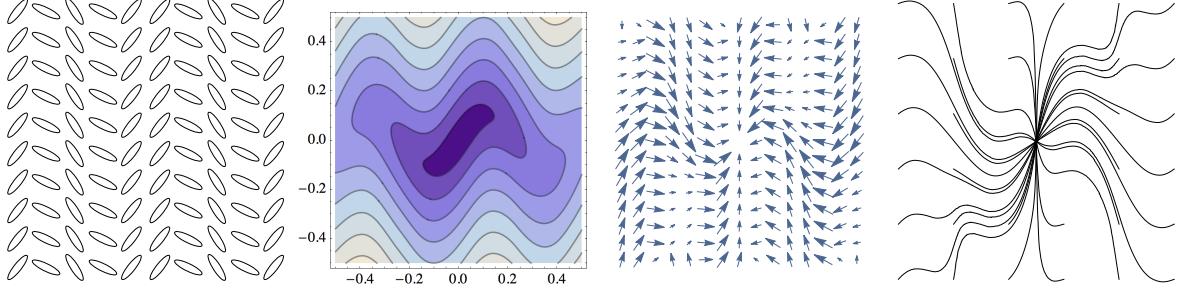


Figure 6: (i) Représentation d'une métrique riemannienne via l'indicatrice de Tissot. (ii) Distance riemannienne au point central (41). (iii) Opposée du gradient riemannien de cette distance (44, droite), et (iv) géodésiques minimisantes vers le point central.

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 (40) et (41) 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 (40) et (41) sont atteints et les chemins optimaux correspondants satisfont, après reparamétrisation à vitesse riemannienne constante,*

$$\frac{d}{dt} \nabla_{\dot{x}} \mathcal{L}(\gamma, \gamma') = \nabla_x \mathcal{L}(\gamma, \gamma'), \quad \text{où } \mathcal{L}(x, \dot{x}) = \frac{1}{2} \|\dot{x}\|_{M(x)}^2. \quad (45)$$

Dans le cas (41), on a de plus  $\gamma'(1) \propto D(y)\mathbf{n}(y)$ , où  $\mathbf{n}$  désigne la normale extérieure à  $\Omega$ .

*Proof.* (On admet la régularité du chemin optimal  $\gamma$ .) 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  $\gamma$  minimise (40) et est paramétré à vitesse constante, ce qui motive l'introduction du lagrangien (45, droite). Considérons un chemin  $\gamma$  optimal pour (40) et une petite perturbation  $\eta$ , supposés réguliers. Un développement limité d'ordre un donne

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

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

$$\int_0^1 \left( \langle \eta, \nabla_x \mathcal{L} \rangle - \langle \eta, \frac{d}{dt} \nabla_{\dot{x}} \mathcal{L} \rangle \right) dt + \langle \eta(1), \nabla_{\dot{x}} \mathcal{L}(\gamma(1), \gamma'(1)) \rangle - \langle \eta(0), \nabla_{\dot{x}} \mathcal{L}(\gamma(0), \gamma'(0)) \rangle. \quad (46)$$

Dans le cas (40) 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 (46), et l'annulation du terme intégral donne (45, gauche). Dans le cas (41), la perturbation satisfait  $\eta(0) = 0$  mais est seulement sujette à  $y + \eta(1) \in \partial\Omega$ , donc  $\nabla_{\dot{x}} \mathcal{L}(\gamma(1), \gamma'(1)) \propto \mathbf{n}(y)$ , ce qui équivaut à  $\gamma'(1) \propto D(y)\mathbf{n}(y)$ .  $\square$

**Remarque 4.6** (Equations de Hamilton). *Les équations d'Euler-Lagrange (45) équivalent à*

$$\frac{d}{dt}\gamma = \nabla_{\hat{x}}\mathcal{H}(\gamma, \hat{\gamma}), \quad \frac{d}{dt}\hat{\gamma} = -\nabla_x\mathcal{H}(\gamma, \hat{\gamma}), \quad \text{où } \mathcal{H}(x, \hat{x}) := \frac{1}{2}\|\hat{x}\|_{D(x)}^2, \quad (47)$$

où  $\hat{\gamma}$  désigne l'impulsion, et  $\mathcal{H}$  le Hamiltonien, qui est le conjugué de Legendre-Fenchel du Lagrangien. Dans le cas riemannien on a  $\hat{\gamma}(t) = M(\gamma(t))\gamma'(t)$  compte tenu de (47, gauche).

Nous terminons cette sous-section par un résultat technique sur la régularité de la fonction distance, utilisé §4.3 pour montrer qu'elle est solution de viscosité d'une EDP adéquate. Rapelons que  $v_0$  est  $C$ -concave, pour une certaine constante  $C$ , ssi  $x \mapsto v_0(x) - C\|x\|^2$  est concave.

**Proposition 4.7.** *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 (45) 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  $C^\infty$  au voisinage de  $\partial\Omega$ .

( $C$ -concavité) Soit  $x \in \Omega$  tel que  $v_0(x) > \varepsilon$ . Soit  $\gamma : [0, T] \rightarrow \overline{\Omega}$  un chemin minimisant pour (41), paramétré à vitesse riemannienne constante, où  $T = v_0(x) > \varepsilon$ ,  $\gamma(0) \in \partial\Omega$ ,  $\gamma(T) = x$ . (Noter que  $\gamma$  et  $T$  dépendent implicitement de  $x$ .) Pour  $h$  assez petit, le chemin  $t \mapsto \gamma(t) + th/T$  prend ses valeurs dans  $\overline{\Omega}$ , et l'on a donc

$$v_0(x + h) \leq F_x(h) := \int_0^T \|\gamma'(t) + h/T\|_{M(\gamma(t)+th/T)} dt$$

La fonction  $F_x$ , définie ci-dessus, est  $C^2$  au voisinage de 0. En bornant  $v_0$  par une fonction  $F_x$  lisse tangente en  $x$ , on établit la propriété de  $C$ -concavité sur une partie compacte  $K \subseteq \Omega$  arbitraire, avec  $C = \sup_{x \in K} \|\nabla^2 F_x(0)\|$ .  $\square$

## 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  $\varepsilon$  joue un rôle plus transparent. Cette technique est introduite dans [Var67] pour l'équation de Poisson et de la chaleur, et fait suite aux travaux de Hopf et Cole portant sur l'équation de Burgers uni-dimensionnelle [Hop50].

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

**Définition 4.8.** *Sous l'Hypothèse 4.1. Pour tout  $\varepsilon > 0$ , on note  $u_\varepsilon : \overline{\Omega} \rightarrow \mathbb{R}$  la solution de*

$$u_\varepsilon - \varepsilon^2 \operatorname{div}(D\nabla u_\varepsilon) = 0 \quad \text{dans } \Omega, \quad u_\varepsilon = 1 \quad \text{sur } \partial\Omega. \quad (48)$$

Sous les hypothèses de la Définition 4.8, on montre [Eva10, théorème 6, section 6.3] que  $u_\varepsilon \in C^2(\overline{\Omega}, ]0, 1[)$  est une solution classique et strictement positive de (48), pour tout  $\varepsilon > 0$ . Notons l'interprétation énergétique

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

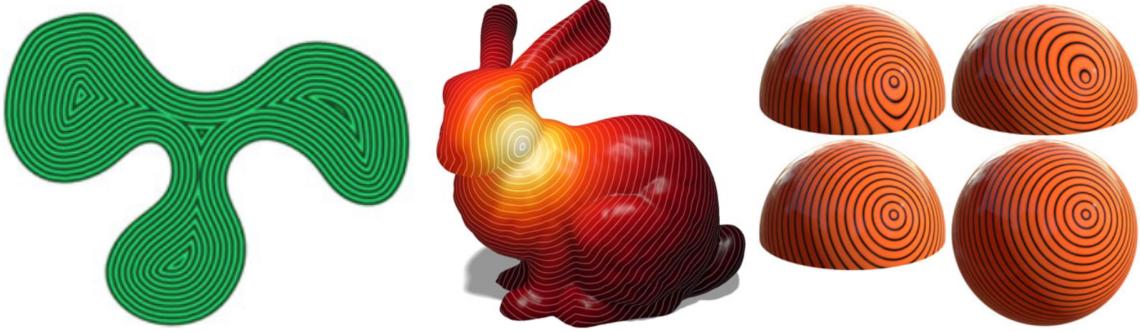


Figure 7: Approximation de la distance géodésique, via le noyau de Poisson ou l'équation de la chaleur. (i) Distance au bord d'un domaine du plan. (ii) Distance à un point d'une variété fermée. (iii) Distance à un point d'une variété à bord, avec différentes conditions au bord (Neumann, Dirichlet, Robin,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

Présentons maintenant la transformation logarithmique et à l'EDP non-linéaire associée. Celle-ci est une “relaxation visqueuse” de l'équation eikonale riemannienne, obtenue dans le cas limite  $\varepsilon \rightarrow 0$ .

**Proposition 4.9.** *Sous les hypothèses de la Définition 4.8. Définissons  $v_\varepsilon \in C^2(\bar{\Omega}, \mathbb{R}_+)$  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*

$$\|\nabla v_\varepsilon\|_D^2 - \varepsilon \operatorname{div}(D\nabla v_\varepsilon) = 1 \text{ dans } \Omega, \quad v_\varepsilon = 0 \text{ sur } \partial\Omega. \quad (49)$$

*Proof.* La fonction  $v_\varepsilon$  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  $\omega$  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 (49), 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 (41).

**Remarque 4.10** (Variantes et améliorations). *Plusieurs heuristiques sont proposées dans [CWW13] pour améliorer la précision l'approximation  $v_\varepsilon$  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, présentés sous la Définition 4.13, au lieu des espaces métriques ou de Hilbert et des considérations énergétiques vus précédemment. Le formalisme discret correspondant est présenté §5. Pour la simplicité, on se limitera aux conditions au bord de Dirichlet.

**Définition 4.11** (Opérateur elliptique). *Un opérateur  $F$  sur un domaine  $\Omega$  est dit dégénéré elliptique si pour tout  $u \in C^2(\Omega)$  et tout  $x \in \Omega$  on a*

$$Fu(x) := \mathcal{F}(x, u(x), \nabla u(x), \nabla^2 u(x)), \quad (50)$$

où  $\mathcal{F} : \bar{\Omega} \times \mathbb{R} \times \mathbb{R}^d \times S_d$  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 usuel<sup>10</sup> 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.

Les opérateurs apparaissant dans l'équation de Poisson (37) et l'équation eikonale relaxée (49) peuvent s'écrire sous la forme (50), comme suit :

$$u - \varepsilon^2(\langle \operatorname{div} D, \nabla u \rangle + \operatorname{Tr}(D\nabla^2 u)), \quad \|\nabla v\|_D^2 - \varepsilon(\langle \operatorname{div} D, \nabla v \rangle + \operatorname{Tr}(D\nabla^2 v)). \quad (51)$$

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 symétriques définies positives, pour tout  $x \in \bar{\Omega}$ . On note que (51, gauche) est elliptique avec  $\delta = 1$ , tandis que (51, droite) est seulement dégénéré elliptique.

La théorie des solutions de viscosité permet de donner sens à l'équation (50) lorsque  $u$  est continue, mais pas forcément différentiable, par l'intermédiaire de fonctions test  $\varphi$  qui lui sont tangentes supérieurement ou inférieurement.

**Définition 4.12.** Soient  $u \in C^0(\Omega)$ ,  $\varphi \in C^2(\Omega)$ , et  $x \in \Omega$ . On dit que  $\varphi$  est tangente supérieurement (resp. inférieurement) à  $u$  en  $x \in \Omega$  si

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

**Définition 4.13.** Soit  $F$  un opérateur dégénéré elliptique, soit  $u \in C^0(\Omega)$  et soit  $x \in \Omega$ . On définit, au sens des solutions de viscosité :

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

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

Si la fonction  $u$  n'admet pas de tangente supérieure (resp. inférieure) en un point  $x$ , alors la condition apparaissant dans la Définition 4.13 est vide, et l'on a donc  $Fu(x) \leq 0$  (resp.  $Fu(x) \geq 0$ ) au sens des solutions de viscosité.

Si une fonction  $u$  est  $C^2$  au voisinage d'un point  $x$ , et si  $F$  est un opérateur elliptique, alors on vérifie facilement que  $Fu(x) \geq 0$  (resp.  $Fu(x) \leq 0$ , resp.  $Fu(x) = 0$ ) au sens des solutions de viscosité si et seulement si ces propriétés sont vraies au sens classique.

Une force du cadre des solutions des viscosités est l'existence d'un principe de comparaison, qui permet de comparer point par point leurs sous-solutions et sur-solutions. Nous le démontrons ici *sous l'hypothèse simplificatrice* que les fonctions à comparer sont  $C^2$ .

---

<sup>10</sup>  $A \preceq B$  ssi  $B - A$  est semi-définie positive.

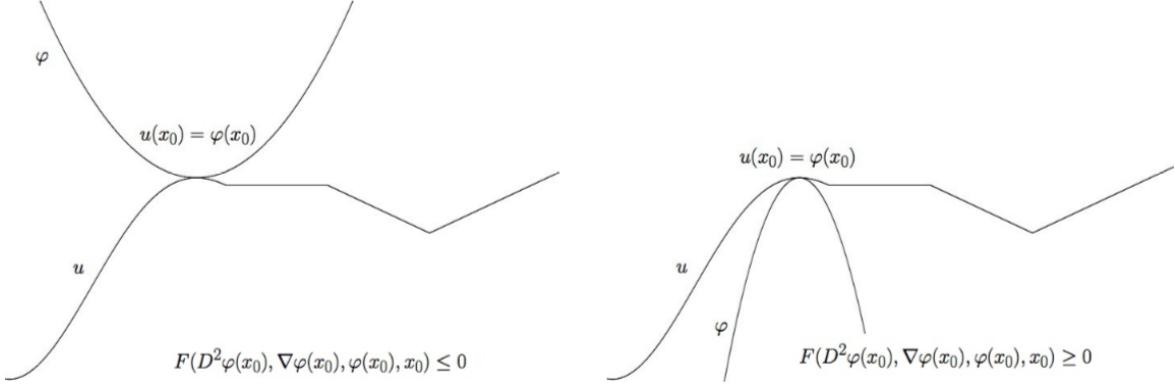


Figure 8: (i) Fonction test tangente *supérieurement* et *sous-solution*, (ii) fonction test tangente *inférieurement* et *sur-solution*, voir la Définition 4.13. Crédits image : [FGN13].

**Lemme 4.14** (Principe de comparaison, cas de fonctions lisses). *Soit  $\bar{u}, \underline{u} \in C^2(\bar{\Omega})$ ,  $f : \Omega \rightarrow \mathbb{R}$ ,  $g : \partial\Omega \rightarrow \mathbb{R}$ . Soit  $F$  un opérateur elliptique sur un ouvert  $\Omega$  borné. Supposons que*

$$\begin{cases} F\bar{u} \leq f, & \text{dans } \Omega, \\ \bar{u} \leq g, & \text{sur } \partial\Omega, \end{cases} \quad \begin{cases} F\underline{u} \geq f, & \text{dans } \Omega, \\ \underline{u} \geq g, & \text{sur } \partial\Omega. \end{cases} \quad (52)$$

Alors  $\bar{u} \leq \underline{u}$  sur  $\bar{\Omega}$ .

*Proof.* Par compacité, il existe un point  $x \in \bar{\Omega}$  où  $\bar{u} - \underline{u}$  atteint son maximum. Supposons pour la contradiction que  $\bar{u}(x) - \underline{u}(x) > 0$ . Alors  $x \in \Omega$ , car  $\bar{u} \leq g \leq \underline{u}$  sur  $\partial\Omega$ . De plus, par les conditions d'optimalité du premier et du second ordre en ce point:

$$\nabla \bar{u}(x) = \nabla \underline{u}(x), \quad \nabla^2 \bar{u}(x) \preceq \nabla^2 \underline{u}(x).$$

Nous en déduisons que  $G\bar{u}(x) \geq G\underline{u}(x)$  pour tout opérateur dégénéré elliptique sur  $\Omega$ , par les propriétés de monotonie par rapport à  $u(x)$  et  $\nabla^2 u(x)$ . En choisissant  $Gu := Fu - \delta u$  on obtient que  $F\bar{u}(x) - F\underline{u}(x) \geq \delta(\bar{u}(x) - \underline{u}(x)) > 0$ , ce qui contredit l'hypothèse que  $F\bar{u}(x) \leq f(x) \leq F\underline{u}(x)$  sur  $\Omega$ . Ainsi  $\bar{u}(x) - \underline{u}(x) \leq 0$ , et la preuve est complète.  $\square$

Il est parfois nécessaire d'appliquer le principe de comparaison à des fonctions  $\bar{u}$  et  $\underline{u}$  seulement continues, par exemple pour comparer la distance au bord  $v_0$  (qui n'est pas différentiable en tout point) avec  $v_\varepsilon$ , voir (53). Pour établir le principe de comparaison dans ce cadre, des hypothèses supplémentaires sont nécessaires, ainsi que des techniques comme le doublement de variables : maximiser  $(x, y) \in \bar{\Omega}^2 \mapsto \bar{u}(x) - \underline{u}(y) - \frac{1}{\rho} \|x - y\|^2$ , où  $\rho > 0$  est petit.

On renvoie à [BR98, Théorème 2.1] pour la preuve du principe de comparaison pour l'équation de Poisson (51, gauche). Comme l'équation eikionale relaxée (51, droite) s'en déduit par un changement de variables monotone, la transformation logarithmique, elle le satisfait également.

**Proposition 4.15.** *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  $\Omega$ , où  $C = C(D, \Omega)$  est une constante, et où  $v_0$  est définie par (41).*

*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  $\Omega$  au sens des solutions de viscosité.

(Preuve de  $-\text{Tr}(D\nabla^2 v_0) \geq -C$ .) Supposons  $\varphi$  tangente inférieurement à  $v_0$  en  $x$ , et rappelons que  $v_0$  est  $C_0$ -concave, voir la Proposition 4.7. Donc

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

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  $\Omega$ . On a montré que  $-\text{Tr}(D\nabla^2 v_0) \geq -C$  sur  $\Omega$ , 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 (41) 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  $\|\nabla u\|_D \geq 1$  sur  $\Omega$  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 (49), 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. Par convention, on étend  $v_0$  à  $\mathbb{R}^d$  par 0.

**Corollaire 4.16.** *Il existe des constantes  $\varepsilon_0 > 0$  et  $C$  telles que pour tout  $0 < \varepsilon \leq \varepsilon_0$ , les fonctions suivantes sont des sous- et sur-solution de viscosité de (49), où l'on note  $\delta := \sqrt{\varepsilon}$*

$$\overline{v_\varepsilon} := (1 - C\delta)G_\delta \star v_0 - C\delta, \quad \underline{v_\varepsilon} := (1 + \varepsilon C)v_0.$$

*Proof.* (Sur-solution.) Soit  $\lambda := 1 + \varepsilon C \geq 1$ . On note que  $\underline{v_\varepsilon} := \lambda v_0$  est nulle sur  $\partial\Omega$ . Soit  $x \in \Omega$ , et soit  $\varphi$  tangente inférieurement à  $v_0$  en  $x$ , de sorte que  $\lambda\varphi$  est tangente supérieurement à  $\underline{v_\varepsilon}$ . Par la Proposition 4.15 on a  $\rho := \|\nabla\varphi(x)\|_D \geq 1$  et  $-\text{Tr}(D\nabla^2\varphi) \geq -C_0$  où  $C_0$  est une constante. Ainsi, en notant  $C_1$  une borne sur  $\|\text{div}(D)\|_{D^{-1}}$  nous obtenons

$$\|\lambda\nabla\varphi\|_D^2 - \varepsilon(\text{Tr}(D\lambda\nabla^2\varphi) + \langle \text{div}(D), \lambda\nabla\varphi \rangle) \geq \lambda^2\rho^2 - \lambda\varepsilon(C_0 + C_1\rho) \geq 1 + \varepsilon(2C - C_0) + \mathcal{O}(\varepsilon^2).$$

En choisissant la constante  $C$  telle que  $2C > C_0$ , et lorsque  $\varepsilon$  soit assez petit, on obtient un membre de droite supérieur ou égal à 1, comme annoncé.

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

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

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  $\lambda = 1 - C\delta$

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

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

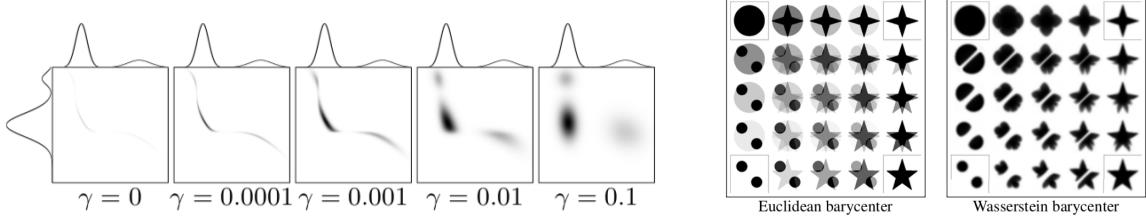


Figure 9: Gauche: effet de la relaxation entropique sur le plan de transport. Droite: interpolation de mesures dans les espaces Euclidiens et de Wasserstein. Credits image : [SdGP<sup>+</sup>15]

Par le principe de comparaison, on obtient  $\overline{v_\varepsilon} \leq v_\varepsilon \leq \underline{v_\varepsilon}$ , pour tout  $0 < \varepsilon \leq \varepsilon_0$  où  $\overline{v_\varepsilon}$  et  $\underline{v_\varepsilon}$  sont définies au Corollaire 4.16. On vérifie par ailleurs que  $\underline{v_\varepsilon} = v_0 + \mathcal{O}(\varepsilon)$  et  $\overline{v_\varepsilon} = v_0 + \mathcal{O}(\sqrt{\varepsilon})$ . Finalement, il existe une constante  $C$  telle que pour tout  $0 < \varepsilon \leq \varepsilon_0$

$$v_\varepsilon - C\varepsilon \leq v_0 \leq v_\varepsilon + C\sqrt{\varepsilon}. \quad (53)$$

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 (39) et la formule de convolution (38). Ces techniques sont présentées en détail dans [SdGP<sup>+</sup>15], ainsi que dans [PCo19] avec d'autres approches.

Soient  $(X, \mu)$  et  $(Y, \nu)$  des ensembles finis munis de mesures de probabilité<sup>11</sup>, 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

$$W_\varepsilon(\mu, \nu) := \min_{\pi \in \Pi(\mu, \nu)} \sum_{x \in X, y \in Y} c(x, y) \pi_{xy} + \varepsilon \text{KL}(\pi; \mu \otimes \nu), \quad (54)$$

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

$$\forall x \in X, \sum_{y \in Y} \pi_{xy} = \mu_x, \quad \forall y \in Y, \sum_{x \in X} \pi_{xy} = \nu_y. \quad (55)$$

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

$$\text{KL}(\pi; \pi^*) := \sum_{z \in Z} \pi_z \ln \left( \frac{\pi_z}{\pi_z^*} \right) \quad \left( = 1 + \sum_{z \in Z} \pi_z \ln \left( \frac{\pi_z}{e\pi_z^*} \right) \right) \quad (56)$$

avec dans notre cas  $Z = X \times Y$  et  $z = (x, y)$ . L'expression (56, droite) de la divergence de Kullback-Lieber  $\text{KL}(\pi; \pi^*)$  est équivalente si  $\pi$  est une mesure de probabilité. On la préfèrera car sa dérivée par rapport à  $\pi$  a une expression un peu plus simple. On note également les propriétés suivantes.

<sup>11</sup>Une 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:  $\sum_{x \in X} \mu_x = 1$ .

**Lemme 4.17.** Pour toutes mesures de probabilité  $\pi, \pi^*$  sur un ensemble fini  $Z$ , on a  $\text{KL}(\pi; \pi^*) \geq 0$  avec égalité ssi  $\pi = \pi^*$ , et  $\pi \mapsto \text{KL}(\pi; \pi^*)$  est convexe.

*Proof.* La fonction  $f : s \in \mathbb{R}_+ \mapsto s \ln s$ , avec par convention  $f(0) = 0$ , est clairement continue et strictement convexe. Par composition,  $\pi \mapsto \text{KL}(\pi; \pi_*)$  est strictement convexe comme annoncé. On conclut par l'inégalité de Jensen

$$\text{KL}(\pi; \pi_*) = \sum_{z \in Z} \pi_z^* f\left(\frac{\pi_z}{\pi_z^*}\right) \geq f\left(\sum_{z \in Z} \pi_z^* \frac{\pi_z}{\pi_z^*}\right) = f(1) = 0. \quad \square$$

Le problème (54) fait intervenir une fonction objectif convexe et  $C^\infty$  sur son domaine de définition  $]0, \infty[^{X \times Y}$ . Ce problème est soumis à  $N_X + N_Y$  contraintes linéaires (55), la contrainte de positivité de  $\pi$  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  $\pi$ , 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 (55) au sein de la fonction objectif. Ainsi (54) s'écrit

$$\begin{aligned} & \min_{\pi \geq 0} \left( \sum_{\substack{x \in X \\ y \in Y}} c(x, y) \pi_{xy} + \varepsilon \text{KL}(\pi; \mu \otimes \nu) + \sup_{\phi} \sum_{x \in X} \phi(x) \left( \mu_x - \sum_{y \in Y} \pi_{xy} \right) + \sup_{\psi} \sum_{y \in Y} \psi(y) \left( \nu_y - \sum_{x \in X} \pi_{xy} \right) \right) \\ &= \min_{\pi \geq 0} \sup_{\phi, \psi} \left( \sum_{x \in X} \phi(x) \mu_x + \sum_{y \in Y} \psi(y) \nu_y + \sum_{\substack{x \in X \\ y \in Y}} \left( c(x, y) - \phi(x) - \psi(y) + \varepsilon \ln \left( \frac{\pi_{xy}}{e \mu_x \nu_y} \right) \right) \pi_{xy} \right) + \varepsilon. \end{aligned} \quad (57)$$

On a utilisé l'expression (56, droite) de la divergence KL, car dans les calculs qui suivent  $\pi$  n'est pas forcément de somme unité. Le théorème du *minimax de Sion*<sup>12</sup> permet d'échanger l'ordre des optimisations dans la dernière expression, et donc d'obtenir

$$\begin{aligned} W_\varepsilon(\mu, \nu) - \varepsilon &= \sup_{\phi, \psi} \left( \sum_{x \in X} \phi(x) \mu_x + \sum_{y \in Y} \psi(y) \nu_y + \min_{\pi \geq 0} \sum_{\substack{x \in X \\ y \in Y}} \left( c(x, y) - \phi(x) - \psi(y) + \varepsilon \ln \left( \frac{\pi_{xy}}{e \mu_x \nu_y} \right) \right) \pi_{xy} \right) \\ &= \sup_{\phi, \psi} \left( \sum_{x \in X} \phi(x) \mu_x + \sum_{y \in Y} \psi(y) \nu_y - \varepsilon \sum_{\substack{x \in X \\ y \in Y}} \mu_x \nu_y \exp\left(\frac{\phi(x) + \psi(y) - c(x, y)}{\varepsilon}\right) \right). \end{aligned} \quad (58)$$

L'expression (58) porte le nom de formulation duale de Kantorovitch du problème du transport optimal, ici discret et avec relaxation entropique. On a utilisé l'identité

$$\min\{(a + b \ln(s/e))s; s > 0\} = b \exp(-a/b),$$

---

<sup>12</sup>Les hypothèses [Kom88] sont la convexité des ensembles d'optimisation, la compacité de l'un deux (ici  $\pi$  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).

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  $\phi$  et  $\psi$  optimaux, le plan de transport  $\pi$  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 (58). 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<sup>+</sup>15].

Différencions  $F(\phi, \psi)$  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

$$\exp(-\phi(x)/\varepsilon) = \sum_{y \in Y} \nu_y \exp\left(\frac{\psi(y) - c(x, y)}{\varepsilon}\right), \quad \text{equiv à: } \Phi(x)^{-1} = \sum_{y \in Y} \nu_y \Psi(y) K(x, y). \quad (59)$$

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 (59, droite), est proche de la convolution (38), 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 permettent en particulier de discréteriser directement l'équation eikionale, sans le terme régularisant rencontré dans la méthode de Varadhan (49), ainsi que de nombreuses autres EDPs issues notamment de la théorie du contrôle optimal.

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 discréterisations 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 croissante en sa troisième variable coordonnée par coordonnée. Un schéma est dit  $\delta$ -elliptique, où  $\delta \geq 0$ , si  $u \mapsto Fu - \delta u$  est DDE.

**Lemme 5.2.** *Si  $F$  est  $\delta$ -elliptique, où  $\delta \geq 0$ , alors pour tout  $u : X \rightarrow \mathbb{R}$  et tout  $\lambda \geq 0$  on a*

$$F(u + \lambda) \geq Fu + \delta\lambda, \quad F(u - \lambda) \leq Fu - \delta\lambda. \quad (60)$$

*Proof.* Soit  $\tilde{F}$  un schéma DDE. Alors par sa définition et ses propriétés de croissance

$$\begin{aligned}\tilde{F}(u + \lambda)(x) &= \tilde{\mathcal{F}}(x, u(x) + \lambda, [u(x) - u(y)]_{y \in X \setminus \{x\}}) \\ &\geq \tilde{\mathcal{F}}(x, u(x), [u(x) - u(y)]_{y \in X \setminus \{x\}}) = \tilde{F}u(x).\end{aligned}$$

Supposons maintenant que  $Fu = \tilde{F}u + \delta u$  est un schéma  $\delta$ -elliptique. Alors  $F(u + \lambda) - Fu = \tilde{F}(u + \lambda) - \tilde{F}u + \delta(u + \lambda - u) \geq \delta\lambda u$  comme annoncé (60, gauche). Le cas de  $u - \lambda$  est similaire.  $\square$

**Remarque 5.3** (Cas d'un opérateur linéaire). *Considérons un schéma  $F$  linéaire et  $\delta$ -elliptique sur l'ensemble  $X = \{1, \dots, N\}$ , avec  $\delta \geq 0$ . Ainsi  $Fu = Au + b$ , où  $A$  est une matrice de taille  $N \times N$ , et où  $b \in \mathbb{R}^N$ . Alors on peut écrire  $A = D - M$  où  $D$  est une matrice diagonale, et  $M$  est une matrice ayant des zéros sur la diagonale. La propriété d'ellipticité de  $F$  implique alors que les entrées de  $D$  et  $M$  sont positives, et satisfont  $D_{ii} \geq \delta + \sum_{1 \leq j \leq N} M_{ij}$  pour tout  $1 \leq i \leq N$ .*

## 5.1 Principe de comparaison, existence d'une solution

A l'instar des opérateurs dégénérés elliptiques présentés §4.3, 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 que dans le cadre continu, et évitent un certain nombre de technicités. Par exemple, il n'est pas nécessaire de passer par des fonctions test pour définir les notions de sous- et sur-solution, comparer les Définitions 4.13 et 5.5. De même, le principe de comparaison s'établit de manière simple et directe le cas discret, voir la Proposition 5.4, alors que la preuve de sa variante continue nécessite des artifices techniques comme le doublement de variables, évoqué après Définition 4.13. Une référence classique sur les schémas DDE est [Obe06].

On commence par énoncer, Proposition 5.4 et Corollaire 5.6, les principes de comparaison associés aux schémas DDE.

**Proposition 5.4** (Principe de comparaison strict). *Soit  $F$  un schéma DDE sur un ensemble fini  $X$ , et soient  $\bar{u}, \underline{u} : X \rightarrow \mathbb{R}$  tels que  $F\bar{u} < F\underline{u}$  sur  $X$ . Alors  $\bar{u} < \underline{u}$  sur  $X$ .*

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

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

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

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

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

**Définition 5.5.** *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  $\bar{u} \leq \underline{u}$  pour toute sous-solution  $\bar{u}$  et toute sur-solution  $\underline{u}$ .*

**Corollaire 5.6** (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  $\bar{u}$  une sous-solution, soit  $\underline{u}$  une sur-solution, et soit  $(\underline{u}_n)_{n \geq 0}$  une suite de sur-solutions strictes convergeant vers  $\underline{u}$  lorsque  $n \rightarrow \infty$ . Alors  $\bar{u} < \underline{u}_n$  sur  $X$  pour tout  $n \geq 0$  par la Proposition 5.4, et donc  $\bar{u} \leq \underline{u}$  sur  $X$  par passage à la limite, comme annoncé. Le cas d'une suite sous-solutions strictes approchant  $\bar{u}$  est similaire.

Si  $F$  est elliptique, et si  $\underline{u}$  est une sur-solution, alors  $\underline{u} + \varepsilon$  est une sur-solution stricte pour tout  $\varepsilon > 0$  par (60, gauche), 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 5.7** (Solution de Perron). *Soit  $F$  un schéma dégénéré elliptique continu et admettant une sous-solution  $\bar{u}$  et une sur-solution  $\underline{u}$ , avec  $\bar{u} \leq \underline{u}$ . Alors  $F$  admet une solution telle que  $\bar{u} \leq u \leq \underline{u}$ , à savoir*

$$u(x) := \max\{v(x); v : X \rightarrow \mathbb{R} \text{ sous-solution et } \bar{u} \leq v \leq \underline{u} \text{ sur } X\}. \quad (61)$$

*Proof.* L'ensemble  $\{v : X \rightarrow \mathbb{R} \text{ sous-solution et } \bar{u} \leq v \leq \underline{u} \text{ sur } X\}$  est compact, car  $X$  est fini et le schéma  $F$  est continu, et il est non-vide, car il contient  $\bar{u}$ . Ainsi  $u : X \rightarrow \mathbb{R}$  est bien définie, satisfait  $\bar{u} \leq u \leq \underline{u}$  sur  $X$ , et le max est atteint dans (61). Soit  $x \in X$ , et soit  $v : X \rightarrow \mathbb{R}$  tel que  $u(x) = v(x)$ . 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  $Fu(x) < 0$ , pour un certain  $x \in X$ . Alors  $u(x) < \underline{u}(x)$  car on aurait sinon  $Fu(x) \geq F\underline{u}(x) \geq 0$  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  $\varepsilon$  assez petit on a  $Fv_\varepsilon(x) \leq 0$  par continuité. De plus  $\bar{u} \leq v_\varepsilon \leq \underline{u}$ , pour  $\varepsilon$  assez petit, donc  $v_\varepsilon$  convient pour (61). Ceci contredit la définition de  $u$ , donc on a  $Fu(x) \geq 0$ , ce qui conclut.  $\square$

**Corollaire 5.8.** *Soit  $F$  un schéma  $\delta$ -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 (60). Ainsi  $F$  admet une solution  $u_*$  telle que  $u - \lambda \leq u_* \leq u + \lambda$  par la Proposition 5.7. De plus  $u_*$  est unique par le principe de comparaison, Corollaire 5.6.  $\square$

## 5.2 Exemples

Les schémas DDE permettent de discréteriser des EDPs dégénérées elliptiques très générales [KT92, BS91, Obe08]. Ils sont cependant limitées en termes de précision, car les schémas DDE ne peuvent être consistants qu'à l'ordre un pour des EDPs d'ordre un, et à l'ordre deux pour des EDPs d'ordre deux. Par ailleurs, certaines approches dites *à deux échelles* [BS91, LN18] font intervenir des stencils de discréétisation très grands, ce qui dégrade la précision numérique des solutions et les rend difficilement utilisables en pratique.

On présente ici quelques discréétisations fondées sur la décomposition de Selling, voir la Proposition 3.7, ce qui permet d'obtenir des schémas DDE relativement compacts et précis, adaptés à une mise en oeuvre numérique. 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 5.9** (Combinaison de schémas). *Si  $F$  et  $G$  des schémas DDE sur un ensemble fini  $X$ , alors c'est aussi le cas des schémas suivants*

$$\alpha F + \beta G, \quad \max\{F, G\}, \quad \min\{F, G\}, \quad \eta \circ F$$

où  $\alpha, \beta \geq 0$  et  $\eta$  est croissante. De plus  $FG$  est DDE si  $F$  et  $G$  sont toujours positifs. Ces propriétés de stabilité valent aussi pour les opérateurs dégénérés elliptiques, dans le cadre continu.

**Remarque 5.10** (Opérateur extrémal). *Il est courant de présenter, ou de reformuler, un opérateur non-linéaire d'intérêt sous la forme*

$$Fu(x) = \max_{\alpha \in A} F^\alpha u(x),$$

où  $A$  est un espace de paramètres, et où  $F^\alpha$  est un opérateur dégénéré elliptique “simple” (par exemple linéaire, ou facile à discréteriser) pour chaque  $\alpha \in A$ . L’opérateur  $F$  est alors lui aussi dégénéré elliptique, s’il est bien défini, par la Proposition 5.9. De plus une stratégie de discréétisation naturelle [BS91], bien que parfois sous-optimale, est de considérer

$$F_h u := \max_{\alpha \in A_h} F_h^\alpha u,$$

où  $h > 0$  est l’échelle de discréétisation,  $A_h \subseteq A$  est un ensemble fini bien choisi approchant l’espace des paramètres, et  $F_h^\alpha$  est une discréétisation de  $F^\alpha$ .

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 considère une matrice  $D \in S_d^{++}$ , dont on se donne une décomposition sous la forme suivante

$$D = \sum_{1 \leq i \leq I} \rho_i e_i e_i^\top, \quad \text{où } \rho_i \geq 0, e_i \in \mathbb{Z}^d, \forall 1 \leq i \leq I. \quad (62)$$

Une telle décomposition est triviale si  $D$  est diagonale. Dans le cas général et en dimension  $d \in \{2, 3\}$ , elle s’obtient typiquement par l’algorithme de Selling, voir la Proposition 3.7.

**Différences finies upwind.** Les différences finies décentrées d’ordre un, souvent dites upwind, sont utilisées notamment pour les discréétisations d’opérateurs DDE des opérateurs d’ordre un. Elles sont définies par

$$\delta_h^{+e} u := \frac{u(x + he) - u(x)}{h} \quad (63)$$

où  $x \in X$  et  $e \in \mathbb{Z}^d$ . Cette expression n’a de sens que si  $x + he \in X$ , voir le dernier paragraphe de cette sous-section pour le traitement des conditions au bord.

**Lemme 5.11.** *Là où il est défini sur  $X$ , le schéma  $-\delta_h^{+e}$  est dégénéré elliptique, et satisfait la propriété de consistante  $\delta_h^{+e} u(x) = \langle \nabla u(x), e \rangle + \mathcal{O}(h)$  pour  $u$  lisse.*

Par combinaison linéaire *non-négative*, on peut discréteriser tout opérateur linéaire d'ordre un. En effet, soit  $\dot{x} = (\alpha_1, \dots, \alpha_d) \in \mathbb{R}^d$ , et soit  $(b_1, \dots, b_n)$  la base canonique de  $\mathbb{R}^d$ , alors

$$\langle \nabla u(x), \dot{x} \rangle = \sum_{1 \leq i \leq d} |\alpha_i| \frac{u(x) - u(x - h \operatorname{sign}(\alpha_i) b_i)}{h} + \mathcal{O}(h). \quad (64)$$

Les opérateurs non-linéaires d'ordre un peuvent souvent s'écrire sous la forme suivante, dite de Bellman:

$$Fu(x) = \max_{\dot{x} \in A} \langle \nabla u(x), \dot{x} \rangle, \quad (65)$$

où  $A \subseteq \mathbb{R}^d$  est un ensemble de paramètres que l'on supposera compact. La Remarque 5.10 s'applique, de sorte que (65) est dégénéré elliptique, et que (64) associé à un échantillonnage fin de  $A$  permet d'en obtenir une discréterisation DDE.

Le cas de l'opérateur eikonal  $Fu(x) = \|\nabla u(x)\|$ , correspondant à  $A = \mathbb{S}^{d-1}$ , admet toutefois une discréterisation alternative particulièrement efficace. Introduite dans [RT92] dans le cadre isotrope, et déjà présentée en (3), elle s'étend au cadre Riemannien anisotrope [Mir19] via la décomposition de Selling (62), comme le montre le résultat suivant.

**Corollaire 5.12.** *Sous l'hypothèse (62), 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 eikonaile Riemannienne,

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

si  $u$  est lisse et si  $x + he_i, x - he_i \in X$  pour tout  $1 \leq i \leq I$ .

*Proof.* Le caractère DDE découle de la Proposition 5.9; noter l'importance du 0, qui assure que le terme mis au carré est non-négatif. Par ailleurs, on note que, pour une fonction  $u$  régulière

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

Ainsi, par linéarité de la trace

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

**Différences finies d'ordre deux.** Les différences finies centrées d'ordre deux sont utilisées notamment pour des discréterisations DDE d'opérateurs d'ordre deux. Elles sont définies par

$$\Delta_h^e u(x) := \frac{u(x + he) - 2u(x) + u(x - he)}{h^2}, \quad (66)$$

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

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

**Corollaire 5.14.** *Sous l'hypothèse (62), 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*

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

*si  $u$  est lisse et si  $x + he_i, x - he_i \in X$  pour tout  $1 \leq i \leq I$ .*

*Proof.* Le caractère DDE 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

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

La Remarque 5.10 associée au Corollaire 5.14, permet d'obtenir des 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  $\det(\nabla^2 u)$  intervenant dans la théorie du transport optimal et certaines applications en optique, on utilise la formule suivante, valable pour toute  $M \in S_d^+$

$$d \det(M)^{\frac{1}{d}} = \inf \{ \text{Tr}(DM); D \in S_d^+, \det(D) = 1 \}. \quad (67)$$

**Différences finies centrées.** Les différences finies centrées d'ordre un sont utilisées notamment pour discrétiser les termes d'ordre un dans les opérateurs elliptiques d'ordre deux, avec un degré élevé de consistance. En effet, et contrairement aux différences finies upwind, elles ne permettent pas *seules* de construire des schémas DDE. Cependant, elles peuvent intervenir dans de tels schémas si elles sont combinées avec des différences finies 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 5.15.** *Là où il est défini sur  $X$ , le schéma  $-\Delta_h^e + \lambda \delta_h^e$  est dégénéré elliptique pourvu que  $|\lambda| h \leq 2$ . De plus on a la propriété de consistance  $\delta_h^e u(x) = \langle \nabla u(x), e \rangle + \mathcal{O}(h^2)$ .*

**Corollaire 5.16.** *Sous l'hypothèse (62), définissons la discrétisation suivante du gradient*

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

*Si  $G$  est  $K$ -Lipschitz, et si  $\varepsilon > 0$ , alors l'opérateur suivant est DDE dès que  $Kh\|D^{-1}e_i\| \leq 2\varepsilon$  pour tout  $1 \leq i \leq I$ :*

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

*Proof.* Pour la consistance, on note que

$$\sum_{1 \leq i \leq I} \rho_i \delta_h^{e_i} u(x) e_i + \mathcal{O}(h^2) = \sum_{1 \leq i \leq I} \rho_i \langle \nabla u(x), e_i \rangle e_i = \left( \sum_{1 \leq i \leq I} \rho_i e_i e_i^\top \right) \nabla u(x) = D \nabla u(x).$$

Le schéma est DDE car la fonction suivante est croissante en chaque  $(w_i)_{1 \leq |i| \leq I}$  pris séparément

$$\sum_{1 \leq i \leq I} \frac{\varepsilon \rho_i}{h^2} (w_i + w_{-i}) + G \left( \sum_{1 \leq i \leq I} \frac{\rho_i}{2h} (w_i - w_{-i}) D^{-1} e_i \right).$$

Noter que formellement  $w_i = u(x) - u(x - he_i)$  (resp.  $w_{-i} = u(x) - u(x + he_i)$ ).  $\square$

**Traitement des conditions au bord.** Les conditions au bord de Dirichlet sont les plus simples à traiter dans le formalisme des solutions de viscosité et des schémas DDE. D'autres types de conditions au bord requièrent d'autres traitements, par exemple dans le cadre de la théorie du transport optimal [Ham19]. Soit  $\Omega \subseteq \mathbb{R}^d$  un domaine, et posons

$$\Omega_h := \Omega \cap h\mathbb{Z}^d, \quad h_x^e := \min\{\eta > 0; x + \eta e \in \Omega_h \cap \partial\Omega\}.$$

Noter que  $h_x^e \in ]0, h]$  par construction. On définit la variante suivante des différences finies upwind

$$\delta_h^{+e} u(x) := \frac{u(x + h_x^e) - u(x)}{h_x^e}, \quad (68)$$

où les valeurs de  $u$  sur  $\partial\Omega$  sont issues de la condition au bord de Dirichlet. Les définitions (63) et (68) coïncident dans l'intérieur du domaine, mais seule (68) reste définie au voisinage du bord. On construit de même les différences finies d'ordre deux et centrées

$$\Delta_h^e u(x) := \frac{2}{h_x^e + h_{-x}^e} (\delta_h^{+e} u(x) + \delta_h^{-e} u(x)), \quad \delta_h^e u(x) := \frac{1}{2} (\delta_h^{+e} u(x) - \delta_h^{-e} u(x)).$$

Les propriétés établies précédemment restent valables, avec la précision suivante concernant le degré de consistance:

$$\Delta_h^e u(x) = \langle e, \nabla^2 u(x)e \rangle + \mathcal{O}(h^r), \quad \delta_h^e u(x) := \langle e, \nabla u(x) \rangle + \mathcal{O}(h^r),$$

où  $r = 2$  si  $x + he, x - he \in X$ , et où  $r = 1$  sinon, c'est à dire au voisinage de  $\partial\Omega$ .

### 5.3 Itérations d'Euler et de Newton

Nous avons établi l'existence de solutions aux schémas DDE, via la méthode de Perron dont les hypothèses sont minimales mais qui n'est pas constructive, voir la Proposition 5.7. On présente ici deux approches alternatives, les méthodes d'Euler et de Newton, applicables sous des hypothèses plus fortes mais dont l'implémentation informatique est en revanche possible et 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 [Obe06].

**Proposition 5.17** (Méthode d'Euler [Obe06]). *Soit  $F$  un schéma  $\delta$ -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  $\{u : X \rightarrow \mathbb{R}; \|Fu\|_\infty \leq \|Fu_0\|_\infty\}$ , et si  $0 < \varepsilon < \delta/K^2$ , alors pour tout  $n \geq 0$*

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

*En particulier,  $\|u_n - u_\infty\|_\infty \leq \delta^{-1} \rho^n \|F(u_0)\|_\infty$ , où  $u_\infty$  est l'unique solution du schéma.*

*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  $F$

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

On en déduit, en utilisant le caractère Lipschitz de  $F$

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

Ainsi  $(1 + K^2\varepsilon^2)\|Fu\|_\infty \geq (1 + \delta\varepsilon)\|Fv\|_\infty$ . On conclut par le Corollaire 5.8.  $\square$

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  $\delta$  est souvent très petite voire nulle. Le pas de temps naturel  $\varepsilon = \delta/(2K^2)$  est ainsi très faible, et le taux de décroissance associé  $\rho = (4K^2 + \delta^2)/(4K^2 + 2\delta^2)$  est très proche de un.

La méthode de Newton avec *damping*, c'est à dire un pas adaptatif voir l'Algorithme 1, est une alternative souvent redoutablement efficace, qui converge typiquement en une dizaine d'itérations. L'ellipticité n'est pas directement une hypothèse de cette méthode, voir la Proposition 5.18, 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 le schéma soit différentiable et que sa jacobienne de  $F$  soit Lipschitz. Notons que cette propriété n'est pas vérifiée par tous les schémas DDE, et notamment pas par ceux qui sont construits comme des max ou min de schémas élémentaires, voir les Proposition 5.9, Corollaire 5.12, et (67).

---

**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 5.18** (Méthode de Newton). *Soit  $F$  un schéma sur un ensemble fini  $X$ , et soit  $u_0 : X \rightarrow \mathbb{R}$ . Supposons que la matrice Jacobienne de  $F$  est inversible, avec  $\|\mathrm{d}F_u\|^{-1} < L$ , et  $K$ -Lipschitz, sur le domaine  $E := \{u : X \rightarrow \mathbb{R}; \|Fu\|_\infty \leq \|Fu_0\|_\infty\}$ . Alors les itérées  $(u_n)_{n \geq 0}$  de l'Algorithme 1 satisfont:*

- (*convergence globale linéaire*)  $\|Fu_n\|_\infty \leq \rho^n\|Fu_0\|_\infty$ , où  $\rho = (4KL^2\|Fu_0\|_\infty)^{-1} < 1$ .
- (*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 tous  $\delta > 0$  et  $n \geq 0$ ,

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

On en déduit, compte tenu du choix de  $v_n$ ,

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

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

Par les arguments précédents, il existe un rang  $n \geq 0$  tel que  $2KL^2\|Fu_n\|_\infty \leq 1$ , et donc  $\delta_n = 1$ . Dans ce cas on obtient  $KL^2\|Fu_{n+1}\| \leq (KL^2\|Fu_n\|)^2$ , puis par récurrence immédiate  $KL^2\|Fu_{n+k}\| \leq (KL^2\|Fu_n\|)^{2^k}$  pour tout  $k \geq 0$  ce qui établit la convergence quadratique locale de  $\|Fu_n\|$  vers 0, donc celle de  $u_n$  car  $[dF]^{-1}$  est borné.  $\square$

## 5.4 Itérations de Jacobi/Gauss-Siedel

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 moins directe que les méthodes d'Euler et de Newton, mais en contrepartie on se passe de certaines de leurs hypothèses restrictives et défauts. Pour Euler: hypothèse d'ellipticité du schéma et lenteur de la convergence. Pour Newton: hypothèse de différentiabilité et nécessité d'inverser des systèmes linéaires de grande taille.

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 5.19.** *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 5.20.** *Soit  $\Lambda$  un opérateur monotone et continu sur  $X$ , et soient  $\bar{u}, \underline{u} : X \rightarrow \mathbb{R}$  tels que  $\Lambda \bar{u} \geq \bar{u}$ ,  $\Lambda \underline{u} \leq \underline{u}$ , et  $\bar{u} \leq \underline{u}$ . Alors les limites suivantes existent et sont point fixes de  $\Lambda$*

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

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

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

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

**Corollaire 5.21** (Méthode de Jacobi). *Soit  $\Lambda, \bar{u}, \underline{u}$  comme dans Proposition 5.20, et supposons de plus que  $\Lambda$  a au plus un point fixe  $u_*$ . Si  $u : X \rightarrow \mathbb{R}$  satisfait  $\bar{u} \leq u \leq \underline{u}$ , alors  $\Lambda^n u \rightarrow u_*$  lorsque  $n \rightarrow \infty$ .*

*Proof.* Par monotonie, on a  $\lambda^n \bar{u} \leq \Lambda^n u \leq \Lambda^n \underline{u}$  pour tout  $n \geq 0$ . Par unicité du point fixe, et par encadrement, on obtient  $\Lambda^n u \rightarrow u_* = \bar{u}_\infty = \underline{u}_\infty$ .  $\square$

Etant donné un schéma numérique  $F$  sur un ensemble fini  $X$ , l'opérateur  $\Lambda$  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 5.22** (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*

$$\lambda \mapsto F(x, \lambda, (\lambda - u(y))_{y \in X \setminus \{x\}}). \tag{69}$$

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

Etant donné un schéma  $F$ , et étant connu l'opérateur  $\Lambda$  associé, on a clairement l'équivalence

$$Fu = 0 \Leftrightarrow \Lambda u = u. \quad (70)$$

On peut donc reformuler la recherche d'un zéro de  $F$  en celle d'un point fixe de  $\Lambda$ . La proposition suivante montre de plus que, si  $F$  est DDE, alors  $\Lambda$  est monotone, ce qui permet d'en trouver un point fixe par itération, voir la Proposition 5.20.

**Proposition 5.23.** *Soit  $F$  un schéma et  $\Lambda$  l'opérateur associé par la Définition 5.22, que l'on suppose exister. Si  $F$  est dégénéré elliptique, alors  $\Lambda$  est monotone. Il est de plus sous-additif sur les constantes, c.a.d  $\Lambda(u + C) \leq \Lambda u + C$  pour tout  $C \geq 0$*

*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  $\lambda$ , donc leurs racines (supposées exister) satisfont  $\Lambda u(x) \leq \Lambda v(x)$ , comme annoncé.

Comme  $F$  est dégénéré elliptique, il est croissante par rapport à sa première variable, donc pour tout  $C \geq 0$

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

d'où l'on tire la sous-additivité sur les constantes.  $\square$

**Remarque 5.24** (Itérations de Gauss-Siedel). *La méthode de Jacobi, présentée Proposition 5.20, est fondée sur une mise à jour globale de l'inconnue  $u$  sur tout l'ensemble  $X$  à chaque itération:  $u_{n+1} = \Lambda u_n$ , pour tout  $n \geq 0$ . Pour les itérations de Gauss-Siedel, on énumère  $X = \{x_1, \dots, x_K\}$  les points de discréttisation, et on pose  $u_n^{k+1}(x_k) = \Lambda u_n^k(x_k)$  pour tout  $0 \leq k < K$  et tout  $n \geq 0$ , ainsi que  $u_{n+1}^0 = u_n^K$ . En d'autres termes, Gauss-Siedel met à jour séquentiellement, plutôt que simultanément, les valeurs de l'inconnue  $u$  aux différents points de  $X$ . Cette flexibilité ouvre sur une série de variantes, discutées §6.2.*

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

**Cas d'un opérateur linéaire.** Nous reprenons les notations de la Remarque 5.3. Par la Définition 5.22 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 discréttisation de l'équation eikonale du Corollaire 5.12. Le problème local résolu pour la mise à jour de Jacobi (69) prend la forme

$$\sum_{1 \leq i \leq I} \rho_i(\lambda - v_i)_+^2 - h^2 = 0, \quad v_i := \min\{u(x + he_i), u(x - he_i)\}, \quad (71)$$

où  $a_+ := \max\{0, a\}$ . Sans nuire à la généralité, on peut supposer  $v_1 \leq \dots \leq v_I$ . Le membre de gauche, vu comme fonction de  $\lambda \in \mathbb{R}$ , 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.

## 6 Résolution numérique et applications de l'équation eikonale

Ce chapitre est dédié à l'analyse numérique et la présentation de l'implémentation informatique de schémas de résolution de l'équation eikonale. On montre la convergence de la discrétisation §6.1, on présente les méthodes de résolution itératives §6.2, et non-itératives §6.3. Quelques applications sont abordées §6.4.

### 6.1 Convergence vers la solution continue

On montre ici la convergence uniforme des solutions l'équation eikonale discrétisée, vers la solution continue. A cet effet, on se donne un domaine ouvert borné  $\Omega \subseteq \mathbb{R}^d$ , muni d'une métrique  $M \in C^0(\overline{\Omega}, S_d^{++})$ . On recherche une solution de l'équation eikonale Riemannienne

$$\|\nabla u\|_D = 1 \quad \text{sur } \Omega, \quad u = 0 \quad \text{sur } \partial\Omega, \quad (72)$$

où  $D := M^{-1}$ . On introduit pour tout  $h > 0$  une discrétisation  $\Omega_h$  du domaine, et une décomposition des tenseurs

$$\Omega_h := \Omega \cap h\mathbb{Z}^d, \quad D(x) = \sum_{1 \leq i \leq I} \rho_i(x) e_i e_i^\top, \quad (73)$$

où  $\rho_i(x) \geq 0$  et  $e_i \in \mathbb{Z}^d$ , pour tout  $1 \leq i \leq I$ . En dimension  $d \in \{2, 3\}$ , on pourra utiliser la décomposition de Selling, voir la Proposition 3.7. La discrétisation de l'équation eikonale est celle déjà introduite dans le Corollaire 5.12: on pose, pour  $u : \Omega_h \rightarrow \mathbb{R}$  et  $x \in \Omega_h$

$$F_h u(x) := h^{-2} \sum_{1 \leq i \leq I} \rho_i(x) \max\{0, u(x) - u(x + he_i), u(x) - u(x - he_i)\}^2. \quad (74)$$

où l'on étend  $u$  par 0 sur le complémentaire de  $\Omega_h$ , de manière consistante avec les conditions au bord (72, droite). On souhaite résoudre  $F_h u = 1$ .

**Lemme 6.1** (Sous et sur-solutions bornées). *Les fonctions suivantes satisfont  $F_h \bar{u} \leq 1$  et  $F_h \underline{u} \geq 1$  sur  $\Omega_h$ , pour tout  $h \leq 1$ :*

$$\bar{u}(x) = 0, \quad \underline{u}(x) = \langle p, x \rangle + q,$$

où  $p \in \mathbb{R}^d$  est tel que  $\|p\|_{D(x)} \geq 1$  pour tout  $x \in \overline{\Omega}$ , et  $q \in \mathbb{R}$  est tel que  $\underline{u}(x + re_i) \geq 0$  pour tout  $x \in \overline{\Omega}$ ,  $1 \leq i \leq I$ ,  $0 \leq r \leq 1$ .

*Proof.* Par construction (74) on a  $F_h \bar{u}(x) = 0$  pour tout  $x \in \Omega_h$ . Par ailleurs

$$F_h \underline{u}(x) \geq \sum_{1 \leq i \leq I} \rho_i(x) \max\{0, -\langle p, e_i \rangle, \langle p, e_i \rangle\}^2 = \sum_{1 \leq i \leq I} \rho_i(x) \langle e_i, p \rangle^2 = \|p\|_{D(x)}^2.$$

Enfin, on remarque que  $p$  existe bien, car la plus petite valeur propre  $\lambda_1(D(x))$  est positive et dépend continument de  $x \in \overline{\Omega}$ . On peut choisir  $q := \max_{x \in \overline{\Omega}} |\langle p, x \rangle| + \max_{1 \leq i \leq I} |\langle p, e_i \rangle|$ .  $\square$

**Corollaire 6.2.** *Pour tout  $h \geq 0$ , il existe une unique solution à  $F_h u_h = 1$ , qui de plus satisfait  $\bar{u} \leq u_h \leq \underline{u}$ , où  $\bar{u}$  et  $\underline{u}$  sont issus du Lemme 6.1.*

*Proof.* L'existence d'une solution découle de la méthode de Perron, voir la Proposition 5.7, et de l'existence de sous- et sur-solutions établie au Lemme 6.1. Si  $F_h u \geq 1$ , alors  $F_h(\lambda u) = \lambda^2 F_h u$  pour tout  $\lambda \geq 0$ , par homogénéité du schéma. Ainsi toute sur-solution de  $F_h = 1$  est limite de sur-solutions strictes, en prenant  $\lambda \rightarrow 1^+$ , et donc  $F_h$  admet un principe de comparaison par le Corollaire 5.6. On en déduit l'unicité de  $u_h$ , et l'encadrement  $\bar{u} \leq u_h \leq \underline{u}$ .  $\square$

On souhaite passer à la limite les solutions discrètes, pour obtenir une solution de l'équation continue. A cet effet, on établit une propriété d'équicontinuité des solutions discrètes, qui permet d'extraire des sous-suites convergentes. On fait dans le Lemme 6.3 l'hypothèse simplificatrice que l'équation eikonalement résolue est isotrope, mais il est possible d'éliminer celle-ci, au prix de l'utilisation de propriétés fines de la décomposition de Selling.

**Lemme 6.3** (Equi-continuité). *On fait l'hypothèse simplificatrice d'une métrique isotrope,*

$$D(x) = c(x)^{-2} \text{Id}, \quad \rho_i(x) = c(x)^{-2},$$

décomposée sur la base canonique  $(e_i)_{i=1}^d$ , avec les poids  $\rho_i(x) = c(x)^{-2}$ ,  $1 \leq i \leq d$ , où  $c : \bar{\Omega} \rightarrow ]0, \infty[$  est continue et strictement positive. Alors la solution notée  $u_h$  de l'équation discrétisée  $F_h u_h = 1$  est  $c_{\max}$ -Lipschitz sur  $h\mathbb{Z}^d \supseteq \Omega_h$ , pour tout  $h > 0$  et vis-à-vis de la distance  $L^1$ , où  $c_{\max} := \max\{c(x); x \in \bar{\Omega}\}$ .

*Proof.* On note  $u := u_h$  pour la lisibilité. Soit  $x \in \Omega_h$ , et soit  $y = x + he_i \in h\mathbb{Z}^d$  l'un de ses voisins immédiats sur la grille cartésienne, où  $1 \leq i \leq d$ . Alors par construction

$$1 = F_h u_h(x) \geq h^{-2} c(x)^{-2} \max\{0, u(x) - u(y)\}^2,$$

donc  $u(x) \leq u(y) + c(x)h \leq u(y) + c_{\max}\|x - y\|_1$ . L'inégalité  $u(x) \leq u(y) + c_{\max}\|x - y\|_1$  reste vraie si  $x \in h\mathbb{Z}^d \setminus \Omega_h$  et si  $y$  est l'un de ses voisins, car alors  $u(x) = 0$  par définition et car  $u(y) \geq 0$  par le Corollaire 6.2. De proche en proche, on obtient  $u(x) \leq u(y) + c_{\max}\|x - y\|_1$  pour tous  $x, y \in h\mathbb{Z}^d$  comme annoncé.  $\square$

**Proposition 6.4.** *Soit  $u_h : \Omega_h \rightarrow \mathbb{R}$  la solution de  $F_h u_h = 1$ . On suppose la famille  $(u_h)_{h>0}$  équi-continue<sup>13</sup>. Alors  $u_h$  converge uniformément lorsque  $h \rightarrow 0$  vers la solution de viscosité de (72).*

*Proof.* On a montré que les fonctions  $u_h$  sont bornées uniformément lorsque  $0 < h \leq 1$ , voir le Corollaire 6.2. Comme cette famille de fonctions est équi-continue, et grâce théorème d'Arzela Ascoli, elle admet une valeur d'adhérence: il existe une suite  $h_n \rightarrow 0$  telle que  $u_{h_n}$  converge uniformément vers une fonction  $u : \bar{\Omega} \rightarrow \mathbb{R}$  lorsque  $n \rightarrow \infty$ . Par convergence uniforme,  $u$  est continue, et comme  $u_h = 0$  hors de  $\Omega_h$  par convention, on obtient  $u = 0$  sur  $\partial\Omega$ .

Soit  $x_* \in \Omega$ , et soit  $\varphi \in C^2(\bar{\Omega}, \mathbb{R})$  tangente supérieurement à  $u$  en  $x_*$ . Quitte à considérer  $x \mapsto \varphi(x) + C|x - x_*|^2$ , avec  $C$  assez grand, on peut supposer que  $x_*$  est un minimum global strict de  $\varphi - u$ . Pour tout  $n \geq 0$ , soit  $x_n$  le minimiseur de  $\varphi - u_{h_n}$  sur  $\Omega_{h_n}$ . Alors  $x_n \rightarrow x_*$  lorsque  $n \rightarrow \infty$ , par convergence uniforme de  $u_{h_n}$  vers  $u$  lorsque  $n \rightarrow \infty$ . Par ailleurs  $\varphi(x_n) - \varphi(y) \leq u_{h_n}(x_n) - u_{h_n}(y)$  pour tout  $y \in \Omega_{h_n}$  par choix de  $x_n$ , donc

$$1 = F_{h_n} u_{h_n}(x_n) \geq F_{h_n} \varphi(x_n) = \|\nabla \varphi(x_n)\|_{D(x_n)}^2 + \mathcal{O}(h_n),$$

d'où l'on déduit  $1 \geq \|\nabla \varphi(x)\|_{D(x)}^2$ , puis que  $u$  est une sous-solution de (72). On montre de même que  $u$  est une sur-solution de l'équation eikonalement, donc une solution. On sait par ailleurs que l'équation eikonalement admet une unique solution de viscosité. Ainsi la famille  $(u_h)_{h>0}$  est équi-continue, bornée, et admet une unique valeur d'adhérence. Elle converge donc vers celle-ci, ce qui conclut.  $\square$

---

<sup>13</sup>Cette propriété découle du Lemme 6.3, dans le cas d'une métrique isotrope du moins.

Le résultat de convergence établi à la Proposition 6.4 est qualitatif, et repose sur un argument de compacité. Il est également possible d'obtenir une estimation quantitative, de la forme

$$u_h - C\sqrt{h} \leq u \leq u_h + Ch \quad (75)$$

où  $C = C(\Omega, M)$  est une constante. On peut rapprocher ce résultat des estimations similaires pour la transformation logarithmique de l'équation de poisson (53), en notant toutefois que les termes d'erreur supérieur et inférieur sont inversés. L'estimation supérieure (75, droite) est facile, et présentée ci-dessous. L'estimation inférieure (75, gauche) est classique également mais un peu plus difficile, voir par exemple [Mir19].

**Lemme 6.5** (Estimation quantitative, côté “facile”). *La solution  $u$  de l'équation eikonale (72) satisfait  $F_h u \leq 1 + \mathcal{O}(h)$  pour tout  $0 < h \leq 1$ . On en déduit  $u \leq u_h + \mathcal{O}(h)$ .*

*Proof.* Pour tous  $x, y \in \bar{\Omega}$ , on a par la Proposition 4.3

$$|u(x) - u(y)| \leq d_M(x, y) \leq \int_0^1 \|x - y\|_{M((1-t)x+ty)} dt \leq \|x - y\|_{M(x)} + \mathcal{O}(\|x - y\|^2).$$

On en déduit  $|u(x) - u(x + he_i)| \leq \|e_i\|_{M(x)} + \mathcal{O}(h)$ , pour tout  $1 \leq i \leq I$ , puis par linéarité de la trace

$$F_h u(x) + \mathcal{O}(h) \leq \sum_{1 \leq i \leq I} \rho_i(x) \|e_i\|_{M(x)}^2 = \sum_{1 \leq i \leq I} \text{Tr}(\rho_i(x) e_i e_i^\top M(x)) = 1.$$

Par homogénéité de  $F_h$ , on obtient  $F_h[(1 - Ch)u] \leq 1$  pour tout  $h$  assez petit, où  $C$  est une constante assez grande. Donc  $(1 - Ch)u \leq u_h$  par le principe de comparaison, ce qui conclut.  $\square$

## 6.2 Méthodes itératives

Dans cette partie, on fixe un opérateur continu monotone  $\Lambda$  sur un ensemble  $X$ , qui est sous-additif sur les constantes<sup>14</sup>, et possède un unique point fixe  $u_* : X \rightarrow \mathbb{R}$ . L'objectif est de présenter des méthodes numériques praticables pour le calcul de ce point fixe.

Les méthodes présentées dans cette section ont vocation à s'appliquer à l'opérateur issu de la résolution locale du schéma de l'équation eikonale (71), qui satisfait ces propriétés, voir la Proposition 5.23. Rappelons que la solution du schéma  $Fu = 0$ , correspond au point fixe de l'opérateur  $\Lambda u = u$ , voir (70), et qu'elle existe et est unique sous des hypothèses basiques, voir les Proposition 5.7 et Corollaire 5.6.

Il n'est a priori pas possible de calculer le point fixe exact  $u_*$  de l'opérateur  $\Lambda$ , en un nombre fini d'opérations, sans hypothèse supplémentaire<sup>15</sup>. Etant donnée une tolérance  $\varepsilon > 0$ , on se donne donc l'objectif de trouver  $u : X \rightarrow \mathbb{R}$  telle que

$$\|u - \Lambda u\|_\infty \leq \varepsilon. \quad (76)$$

Une analyse plus fine que celle présentée ici est nécessaire pour estimer l'erreur  $\|u - u_*\|_\infty$  entre le point fixe exact et une solution de (76), ou encore la vitesse de convergence et la complexité numérique des méthodes. Sur ce dernier point, mentionnons qu'elles convergent en pratique nettement plus vite si elles sont initialisées avec une *sur-estimation de la solution*.

<sup>14</sup>C.à.d  $\Lambda(u + C) \leq \Lambda u + C$  pour tout  $u : X \rightarrow \mathbb{R}$  et tout  $C \geq 0$ , voir la Proposition 5.23

<sup>15</sup>Une telle hypothèse est justement introduite §6.3.

**Itération globale, dite de Jacobi.** Les points fixes d'un opérateur monotone peuvent s'approcher en itérant celui-ci, sous des hypothèses adéquates, voir la Proposition 5.20 et Corollaire 5.21.

**Proposition 6.6.** *Sous les hypothèses de continuité, monotonie, sous-additivité sur les constantes, et unicité du point fixe  $u_*$  de l'opérateur  $\Lambda$  sur un ensemble fini  $X$ , faites en introduction. On a  $\Lambda^n u_0 \rightarrow u_*$  lorsque  $n \rightarrow \infty$ , pour tout  $u_0 : X \rightarrow \mathbb{R}$ .*

*Proof.* Par sous additivité sur les constantes, on a  $\Lambda(u_* + K) \leq \Lambda u_* + K = u_* + K$  pour tout  $K \geq 0$ . Donc  $\underline{u} := u_* + K$  est une sur-solution, c.a.d  $\Lambda \underline{u} \leq \underline{u}$ . De même  $\Lambda \bar{u} \geq \bar{u}$  où  $\bar{u} := u_* - K$ . Le Corollaire 5.21 permet de conclure.  $\square$

L'itération globale de l'opérateur  $\Lambda$  est plus efficace en pratique que l'itération d'Euler présentée Proposition 5.17, qui est d'ailleurs difficilement applicable dans le cas de l'équation eikionale, dont l'opérateur est elliptique mais pas dégénéré elliptique. Par ailleurs l'itération de Newton présentée Proposition 5.18 ne s'applique pas ici, car le schéma n'est typiquement pas différentiable, et sa Jacobienne lorsqu'elle existe est très mal conditionnée voire non-inversible. Cependant, l'itération globale reste coûteuse et loin d'être optimale du point de vue de la complexité numérique, car elle ne reflète pas le phénomène de propagation de propagation d'un front dans le domaine discrétilisé, sous-jacent au modèle de l'équation eikonal.

**Itération localisée, dite de Gauss-Siedel.** Pour diminuer le coût numérique de la résolution de l'équation eikonal, on remplace l'itération globale par des itérations locales choisies judicieusement. Le résultat suivant donne un cadre général permettant de montrer la convergence de ces méthodes.

**Définition 6.7.** *Pour tout  $x \in X$ , on note<sup>16</sup>  $V(x) := \{y \in X; \text{"}\Lambda u(x) \text{ dépend de } u(y)"\}$  le voisinage de dépendance de  $\Lambda$  en  $x$ . Soit aussi  $V[y] := \{x \in X; y \in V(x)\}$  le voisinage inversé.*

**Proposition 6.8.** *Mêmes hypothèses que la Proposition 6.6. Soit  $u_0 : X \rightarrow \mathbb{R}$ , et*

$$u_{n+1}(x) := \begin{cases} \Lambda u_n(x) & \text{si } x \in X_n \text{ et } |\Lambda u_n(x) - u_n(x)| > \varepsilon, \\ u_n(x) & \text{sinon,} \end{cases}$$

*pour tout  $0 \leq n \leq N$ , où  $X_n \subseteq X$  est un sous-ensemble de points. On fait les hypothèses suivantes sur la famille  $(X_n)_{n=0}^N$*

- (*Les voisins d'un point mis à jour sont revisités.*) Pour tout  $0 \leq n \leq N$  et tout  $x \in X_n$ :

$$\text{si } |u_{n+1}(x) - u_n(x)| > \varepsilon \quad \text{alors } V[x] \subseteq \bigcup_{n < m \leq N} X_m.$$

- (*Tous les points sont visités une fois au moins.*) On a  $X \subseteq \bigcup_{n=0}^N X_n$ .

*Alors  $\|u_N - \Lambda u_N\|_\infty \leq \varepsilon$ .*

*Proof.* Soit  $x \in X$ , et soit considérons l'entier  $0 \leq n \leq N$  maximal tel que  $x \in X_n$ ; noter que cet entier existe par la seconde hypothèse. Par la première hypothèse, on a  $u_N(y) = u_n(y)$  pour tout  $y \in V(x)$ , donc  $\Lambda u_N(x) = \Lambda u_n(x)$ . Par construction, on a  $u_{n+1}(x) = u_N(x)$ , et  $|\Lambda u_n(x) - u_{n+1}(x)| \leq \varepsilon$ . Le résultat s'ensuit.  $\square$

<sup>16</sup>Formellement :  $V(x)$  est l'ensemble des  $y \in X$  tel qu'il existe  $u_0, u_1 : X \rightarrow \mathbb{R}$  égales sur  $X \setminus \{y\}$  et telles que  $\Lambda u_0(x) \neq \Lambda u_1(x)$ .

Un grand nombre de méthodes numériques pour la résolution des équations eikonales peuvent se décrire dans le cadre de la Proposition 6.8, et se distinguent par le choix des ensembles  $(X_n)_{n \geq 0}$ . En voici quelques unes, sans exhaustivité, et décrites informellement:

- (Ensembles  $(X_n)_{n \geq 0}$  fixés a priori.) Méthodes de sweeping [Zha05]. Chaque ensemble  $X_n$  décrit une tranche du domaine, verticale de gauche à droite puis de droite à gauche, horizontale de haut en bas puis de bas en haut, et ainsi de suite jusqu'à convergence. Certaines variantes utilisent des tranches diagonales.
- (Propagation à partir des points dont la valeur change.) Iteration adaptative de Gauss-Siedel (AGSI) [BR06]. Lorsque un point  $x \in X$  est mis à jour, ses voisins  $V[x]$  sont placés dans une file d'attente pour mise à jour ultérieure.
- (Propagation à partir des points dont la valeur se stabilise.) Fast Iterative Method (FIM) [JW08]. Lorsqu'un point  $x \in X$  est mis à jour, on le place dans une file d'attente pour une seconde mise à jour ultérieure. Si celle-ci est sans effet, alors ses voisins  $V[x]$  sont placés dans la file d'attente.

**Adaptation aux architectures massivement parallèles** La résolution des équations eikonales sur architecture GPGPU (General Purpose Graphics Processing Unit) permet d'accélérer ces calculs d'un facteur typiquement compris entre  $20\times$  et  $100\times$  [JW08, GHZ18]. Pour en tirer parti, il faut prendre en compte leur architecture nommée Single Instruction Multiple Threads (SIMT). La mise en oeuvre suivante est typique.

- Un code exécuté par le CPU supervise le déroulement des calculs, et fait appel à des *noyaux* exécutés par le GPU.
- Chaque noyau est exécuté par des *threads*, eux mêmes regroupés en *blocs*, et assignés à des unités de calcul du GPU.

A titre indicatif, une Nvidia GTX 1080 dispose de 80 unités de calcul, chacune capable d'exécuter un *warp* de 32 threads simultanément. Si le nombre de blocs demandés excède le nombre d'unités de calcul, alors le GPU les exécutera successivement. Si le nombre de threads au sein d'un bloc excède la taille du warp, alors l'unité de calcul les exécutera successivement.

Les spécificités de l'architecture GPU induisent différents compromis, et en particulier un avantage et un inconvénient discutés ci-dessous.

- (Mémoire partagée par les threads d'un bloc) Si les threads d'un bloc travaillent sur un petit ensemble de données communes, alors il est possible et recommandé de les charger une seule fois dans une mémoire partagée au sein de l'unité de calcul, ce qui permet de diminuer les accès à la mémoire globale. C'est avantageux car dans les méthodes de calcul scientifique parallèles, la bande passante vers la mémoire globale est souvent plus limitante que le coût des opérations numériques
- (Pénalité liée à la divergence des threads) Supposons que les threads d'un warp<sup>17</sup> rencontrent une instruction de branchement “**if** condition **then** instr1 **else** instr2”, et que la *condition* soit vraie pour certains mais pas pour d'autres. Alors les deux branches *instr1* et *instr2* sont exécutées *successivement* en désactivant les threads non concernés; leurs temps de traitement s'additionnent donc.

---

<sup>17</sup>Sous partie d'un block dont la taille est déterminée par l'architecture matérielle

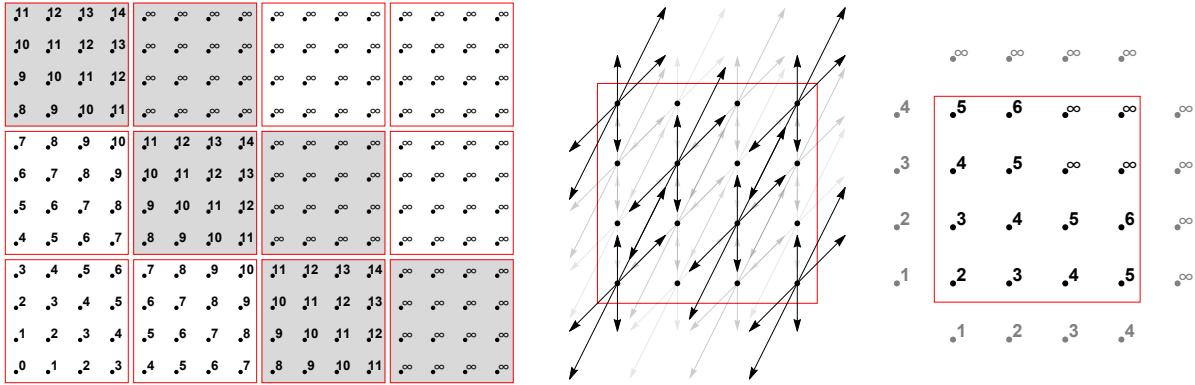


Figure 10: *Gauche:* Décomposition de la grille cartésienne  $X_h$  en tuiles  $X_h^b$ , d'indice de block  $b \in B_h$ . Les blocs grisés sont *actifs*. *Centre:* La mise à jour d'un block  $b \in B_h$  requiert de charger les valeurs de l'inconnue  $u : X_h \rightarrow \mathbb{R}$  à la fois au sein de la tuile  $X_h^b$  et en certains points voisins. *Droite:* Plusieurs mises à jour sont effectuées au sein de la tuile (deux ici), pour permettre au front de s'y propager.

Dans le cadre d'une résolution de l'équation eikionale sur GPU [JW08], le domaine est typiquement découpé en tuiles de taille  $8 \times 8$  ou  $4 \times 4 \times 4$ . Un programme exécuté sur le CPU détermine lesquelles de ces tuiles doivent être mises à jour, en s'inspirant des méthodes (Sweeping, AGSI, FIM) présentées au paragraphe précédent. Chaque tuile est assignée à un bloc de threads, chaque point étant traité par un thread. La mise à jour d'une tuile comporte le chargement des données internes et voisines, plusieurs itérations internes de l'opérateur  $\Lambda$ , puis l'export des nouvelles valeurs. (Voir la Fig. 10, et le TP correspondant.)

### 6.3 Méthode du fast marching

La méthode du Fast Marching se distingue des méthodes itératives par l'introduction d'une hypothèse supplémentaire, dite de causalité, et par une complexité quasi-linéaire en  $\mathcal{O}(KN \ln N)$  indépendamment des paramètres de l'EDP résolue, où  $N$  est le nombre de points de discrétisation et  $K$  le nombre moyen de voisins d'un point dans le schéma de discrétisation [Tsi95, Set96, KS98, SV03, Mir14b, Mir14a, Mir18, Mir19].

La propriété de causalité est l'analogue discret du caractère déterministe du problème de contrôle optimal résolu. Elle exprime le fait que les temps d'arrivée du front ultérieurs à une date donnée dépendent seulement des temps antérieurs à cette date. Etant donnée  $u : X \rightarrow \mathbb{R}$  et  $\lambda \in \mathbb{R}$ , on pose pour tout  $x \in X$

$$u^{<\lambda}(x) := \begin{cases} u(x) & \text{si } u(x) < \lambda, \\ +\infty & \text{sinon.} \end{cases}$$

**Définition 6.9** (Opérateur causal). *On dit qu'un opérateur  $\Lambda$  sur un ensemble fini  $X$  est causal si pour tous  $u, v : X \rightarrow \mathbb{R}$  et tout  $\lambda \in \mathbb{R}$  on a*

$$u^{<\lambda} = v^{<\lambda} \quad \Rightarrow \quad (\Lambda u)^{\leq \lambda} = (\Lambda v)^{\leq \lambda}.$$

La méthode du fast marching est présenté de manière abstraite dans l'Algorithme 2, et sa capacité à calculer le point fixe exact de l'opérateur en un nombre fini d'itérations et présentée dans la Proposition 6.10. Notons que l'opération de mise à jour partielle  $u_{n+1} \leftarrow \Lambda u_n^{\leq t_n}$  qu'il

fait intervenir est peu coûteuse, car seules les valeurs en quelques points doivent en fait être actualisées - les voisins  $y \in V[x]$  des points  $x \in X$  dont la valeur satisfait  $t_{n-1} < u_n(x) \leq t_n$ , que l'on a “acceptés” à l’itération précédente.

---

**Algorithm 2** Algorithme du fast marching dans un cadre abstrait

---

**Initialization**  $n \leftarrow 0$ ,  $t_0 \leftarrow -\infty$ , et  $u_0 \leftarrow +\infty$  sur  $X$ .

**While**  $t_n < \infty$  **do**

$$u_{n+1} \leftarrow \Lambda u_n^{\leq t_n}, \quad t_{n+1} := \min\{u_{n+1}(x); x \in X, u_{n+1}(x) > t_n\}, \quad n \leftarrow n + 1.$$


---

**Proposition 6.10** ([Mir19], Appendix A, Proposition A.2). *Soit  $\Lambda$  un opérateur monotone et causal sur un ensemble fini  $X$ , auquel on applique la méthode du fast marching, voir l’Algorithme 2. Alors il existe  $N < \#(X)$  tel que<sup>18</sup>  $t_N = +\infty$ , de sorte que l’algorithme termine, et que  $\Lambda u_N = u_N$ . De plus on a  $u_n \leq u_{n+1}$  sur  $X$ , et  $u_n^{\leq t_n} = u_{n+1}^{\leq t_n}$ , pour tout  $0 \leq n \leq N$ .*

Si l’opérateur de mise à jour  $\Lambda$  est construit à partir d’un schéma dégénéré elliptique  $F$ , voir la Définition 5.22, alors la propriété de causalité introduite Définition 6.9 se déduit [Mir19, AppendixA, Proposition A.4] de la propriété suivante du schéma:

$$Fu(x) = \mathcal{F}(x, u(x), (u(x) - u(y))_+, y \in X), \quad (77)$$

où  $a_+ := \max\{0, a\}$ . En d’autres termes, il suffit que le schéma ne dépende que des *parties positives* des (opposées des) différences finies. Cette propriété est bien satisfaite pour la discrétisation que nous considérons pour l’équation eikonaile Riemannienne, voir le Corollaire 5.12. La complexité quasi-linéaire du fast marching est essentiellement optimale pour le problème résolu, et les gains en temps de calcul peuvent atteindre  $10\times$  et plus dans certaines applications, notamment liées à l’imagerie médicale [Mir14b, Mir14a], par rapport aux méthodes itératives comme l’AGSI et la FIM exécutées également sur CPU. Cependant deux obstructions limitent ses applications:

- (Causalité) La mise en oeuvre du fast-marching requiert un opérateur satisfaisant la propriété de causalité, voir la Définition 6.9, ou un schéma numérique obéissant à la propriété correspondante (77). Construire une discrétisation causale peut poser des difficultés, en particulier si la métrique a une forme complexe [Mir14b, Mir14a, Mir18, Mir19], ou si le domaine est représenté par un maillage non-structuré [KS98, SV03].
- (Séquentialité) L’algorithme du fast marching requiert d’effectuer les mises à jour des points dans un ordre spécifique, déterminé par une queue de priorité. Il est par nature séquentiel, et ne peut donc pas bénéficier des accélérations permises par les architectures matérielles massivement parallèles de type GPGPU.

## 6.4 Applications de l’équation eikonaile

On donne ici quelques aperçus d’applications de la résolution numérique de l’équation eikonaile. On pourra consulter [PPKC10] pour les applications en traitement de l’image et de la géométrie, et [Set99] pour d’autres domaines de l’ingénierie et du calcul scientifique.

---

<sup>18</sup>Rappelons que le minimum de l’ensemble vide est défini comme  $+\infty$

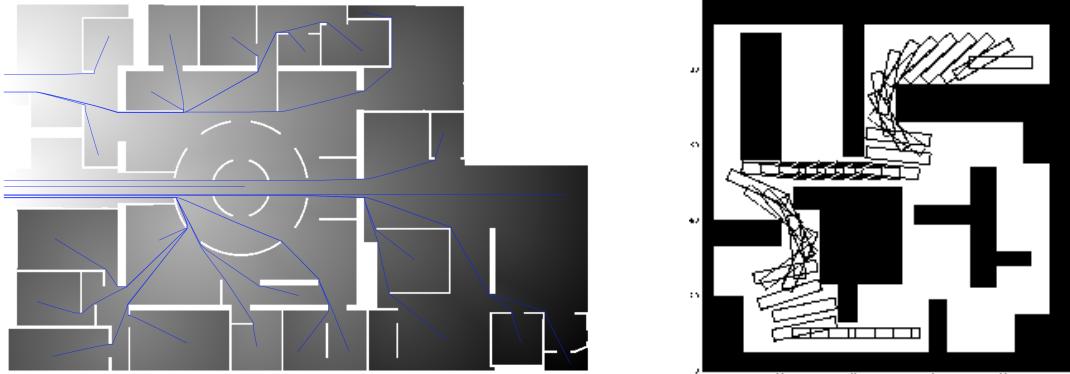


Figure 11: Gauche : Chemins minimaux vers la sortie du centre Pompidou. Droite: Problème du sofa, image J.A. Sethian.

**Shape from shading** La reconstruction de formes à partir des données d’illumination, fait partie des premières applications proposées de la résolution des équations eikonales [RT92]. En effet, sous hypothèse d’un éclairage et d’une prise de vue depuis la verticale, et d’une réflectivité du matériau Lambertien, l’intensité lumineuse réfléchie prend la forme

$$I(x) = \frac{\alpha I_0}{\sqrt{1 + \|\nabla u(x)\|^2}},$$

où  $u$  est une description de la forme par une carte d’élévation,  $I_0$  est l’intensité de la source, et  $\alpha$  est le coefficient de réflectivité. On peut donc reconstruire  $u$  par la résolution d’une équation eikonal, sous certaines hypothèses, notamment l’absence de forme en volcan.

**Planification de mouvements.** La résolution de l’équation eikonal permet de calculer le plus court chemin d’un point à un autre dans un domaine comportant éventuellement des obstacles. Ce calcul peut être utilisé lors de la planification de mouvements robotiques, ou encore pour comme modèle de la vitesse individuelle souhaitée par les piétons d’une foule [MV11]. Certaines applications comme le problème du déménagement de sofa nécessitent calcul dans un espace d’états non physique comme  $\mathbb{R}^2 \times \mathbb{S}^1$ , les positions et orientations. Voir la §7 pour des modèles incorporant des contraintes de manœuvrabilité comme l’impossibilité de déplacer latéralement un véhicule.

**Traitement de données géométriques.** Le calcul de géodésiques et de distances géodésiques sur des maillages intervient à dans divers traitements. Elle repose sur résolution de l’équation eikonal, soit directement, soit via l’équation de Poisson ou de la Chaleur §4. Voir [PPKC10, CWW13].

**Re-distantiation pour les fonctions implicites.** Certaines méthodes de simulation numérique représentent des régions sous une forme implicite  $U = \{x \in \Omega; \psi(x) \leq 0\}$ . La résolution de l’équation eikonal permet de calculer la distance signée à un bord, qui est le choix préférentiel de  $\psi$ .

**Segmentation d’images.** Certains algorithmes de segmentations de régions dans des images reposent sur la propagation de fronts, avec des conditions d’arrêt adéquates, implémentées via



Figure 12: Diagramme de Voronoi Riemannien [PPKC10]. Segmentation d’une région par modèle géodésique de Rander [CCM16].

la résolution de l’équation eikonal. Une approche alternative, considérée dans [CCM16], est de décrire leurs bords en tant que géodésiques pour des métriques adéquates, voir la Fig. 12. Une discussion plus détaillée portant sur l’extraction de structures tubulaires est présentée §7.4.

**Tomographie par temps de trajets.** La reconstitution de l’intérieur d’un domaine, via des mesures extérieures, porte le nom de tomographie. En sismologie, ce sont des ondes élastiques qui se propagent dans un milieu, et dont l’amplitude au cours du temps est enregistrée, en particulier la date de première arrivée. Les ondes sismiques satisfont l’équation des ondes, mais suivant les conditions on peut se satisfaire de l’approximation haute fréquence, de type eikonal [LBBMV17]. On passe ainsi d’une équation linéaire dépendant du temps (des dizaines de milliers de pas de temps sont typiquement nécessaires), à une équation non-linéaire indépendante du temps, ce qui permet de diminuer le coût numérique. Il existe d’autres méthodes encore, comme les méthodes de tir fondées sur l’optique géométrique, donc sur des EDOs plutôt que des EDPs, voir la Fig. 13.

## 7 Contrôle non-holonomique

Cette section est dédiée au calcul numérique de chemins dans un domaine  $\Omega \subseteq \mathbb{R}^2$  minimisant globalement une quantité faisant intervenir la courbure. Plus précisément, l’énergie de  $(\mathbf{x}, \boldsymbol{\theta}) : [0, L] \rightarrow \overline{\Omega} \times \mathbb{S}^1$  est définie de manière informelle<sup>19</sup> par

$$\mathcal{E}(\mathbf{x}, \boldsymbol{\theta}) := \int_0^L \rho(\mathbf{x}, \boldsymbol{\theta}) \mathcal{C}(\boldsymbol{\theta}') \, dl, \quad \text{sous contrainte } \mathbf{x}' = \mathbf{e}_{\boldsymbol{\theta}}. \quad (78)$$

On a noté  $\mathbf{x}'$  et  $\boldsymbol{\theta}'$  les dérivées de temporelles de  $\mathbf{x}$  et  $\boldsymbol{\theta}$ , et introduit le vecteur unitaire  $\mathbf{e}_{\boldsymbol{\theta}} := (\cos \theta, \sin \theta)$ . Le paramètre  $\xi > 0$  sert à moduler l’intensité de la pénalisation de courbure. Dans les applications, la longueur euclidienne  $L$  du chemin est libre, tandis que les positions et tangentes aux extrémités  $\mathbf{x}(0), \boldsymbol{\theta}(0), \mathbf{x}(L)$  et  $\boldsymbol{\theta}(L)$  sont fixées.

La fonction  $\rho : \overline{\Omega} \times \mathbb{S}^1 \rightarrow ]0, \infty[$ , supposée continue, définit un coût local arbitraire. Dans les applications, il sert à favoriser les chemins passant au voisinage de certaines positions avec une certaine direction tangente. Par exemple les chemins s’alignant avec les structures détectées dans une image à segmenter (contours d’organes, vaisseaux sanguins, etc).

La fonction  $\mathcal{C}$  désigne un coût dépendant de la courbure  $\kappa = \dot{\boldsymbol{\theta}}$  du chemin. Les modèles suivants de  $\mathcal{C}(\kappa)$  sont standard, et associés aux noms de Reeds-Shepp, Euler-Mumford, et Dubins:

$$\sqrt{1 + \kappa^2}, \quad 1 + \kappa^2, \quad 1 + \chi_{|\kappa| \leq 1}, \quad (79)$$

<sup>19</sup>Une formulation rigoureuse est donnée §7.1 et 7.2.

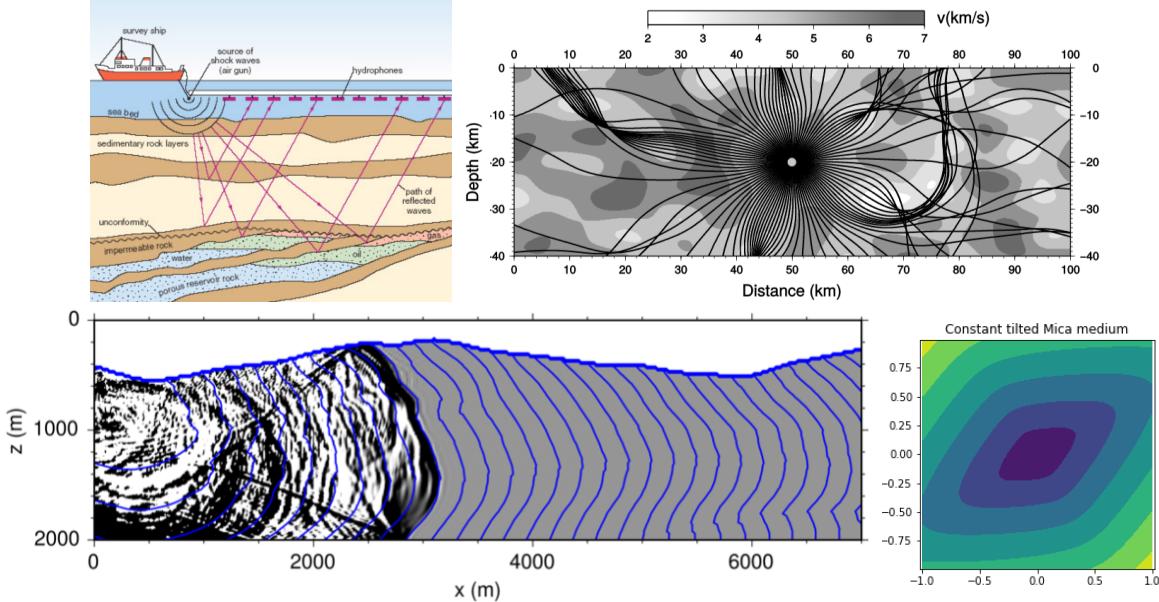


Figure 13: Tomographie sismique par temps de trajet. Haut gauche : contexte. Haut droit : illustration d'une méthode de tir fondée sur l'optique géométrique. Bas gauche : comparaison d'une simulation fondée sur les équations de l'élasticité, et leur asymptotique haute fréquence. Bas droite : exemple de profil de vitesse anisotrope associé à une anisotropie cristalline.

où  $\chi$  désigne la fonction caractéristique d'un ensemble, qui est nulle sur celui-ci et vaut  $+\infty$  ailleurs. Ces quatre modèles, si l'on compte les deux variantes associées au coût de Reeds-Shepp, correspondent à des pénalisations croissantes des fortes courbures et des irrégularités du chemin:

- Le modèle de Reeds-Shepp (reversible) est caractérisé par le coût de Reeds-Shepp (79, gauche) et la possibilité de *passer en marche arrière*, c'est à dire que la contrainte dans (78) est relaxée en  $x' = \pm \mathbf{e}_\theta$ . Ce modèle possède une structure mathématique *sous-Riemannienne* particulièrement riche, et correspond au contraintes de déplacement d'une chaise roulante ainsi qu'au fonctionnement du cortex visuel, voir §7.1.
- Le modèle de Reeds-Shepp (forward) est caractérisé par (79, gauche). Contrairement à ce que laisse supposer (78) au premier abord, ce modèle, correctement défini, autorise les rotations sur place. Ces dernières peuvent être interprétées comme des masses de Diracs de  $\theta'$ , dont le coût est proportionnel à l'angle.
- Le modèle de Euler-Mumford, défini par (79, centre), correspond aux courbes *elastica* qui sont les positions de repos d'une barre élastique. Ses chemins minimaux sont particulièrement lisses et intuitifs.
- Le modèle de Dubins, défini par (79, droite), intègre une borne sur la courbure maximale des trajectoires. Il est populaire en avionique.

**Remarque 7.1** (Modulation de la pénalisation de courbure). *Les pénalisations de courbure de la forme  $\kappa \mapsto \mathcal{C}(\xi(\kappa - \mu))$ , où  $\mathcal{C}$  est l'un des coûts (79) et  $\xi$  et  $\mu$  sont des constantes, sont également traitables par l'approche présentée dans cette section. La constante  $\xi$  a la dimension d'une longueur, et correspond au rayon de courbure typique de la trajectoire. La constante  $\mu$*

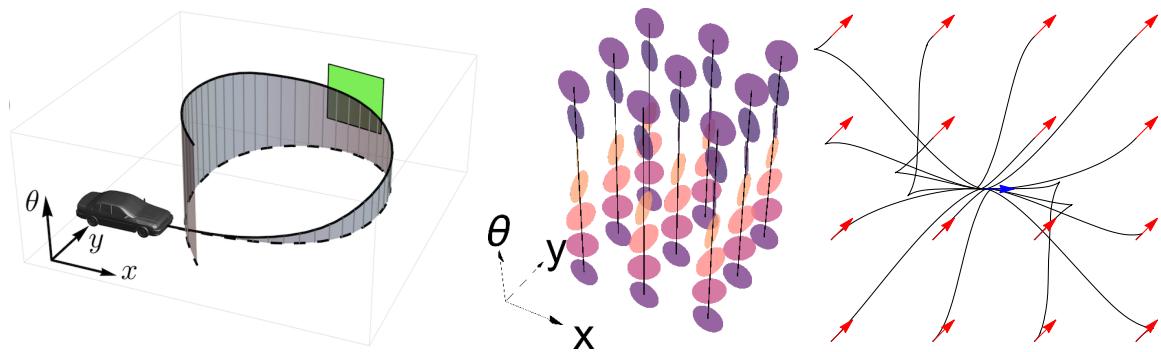


Figure 14: Illustration du modèle de Reeds-Shepp, posé sur  $\mathbb{R}^2 \times \mathbb{S}^1$ .

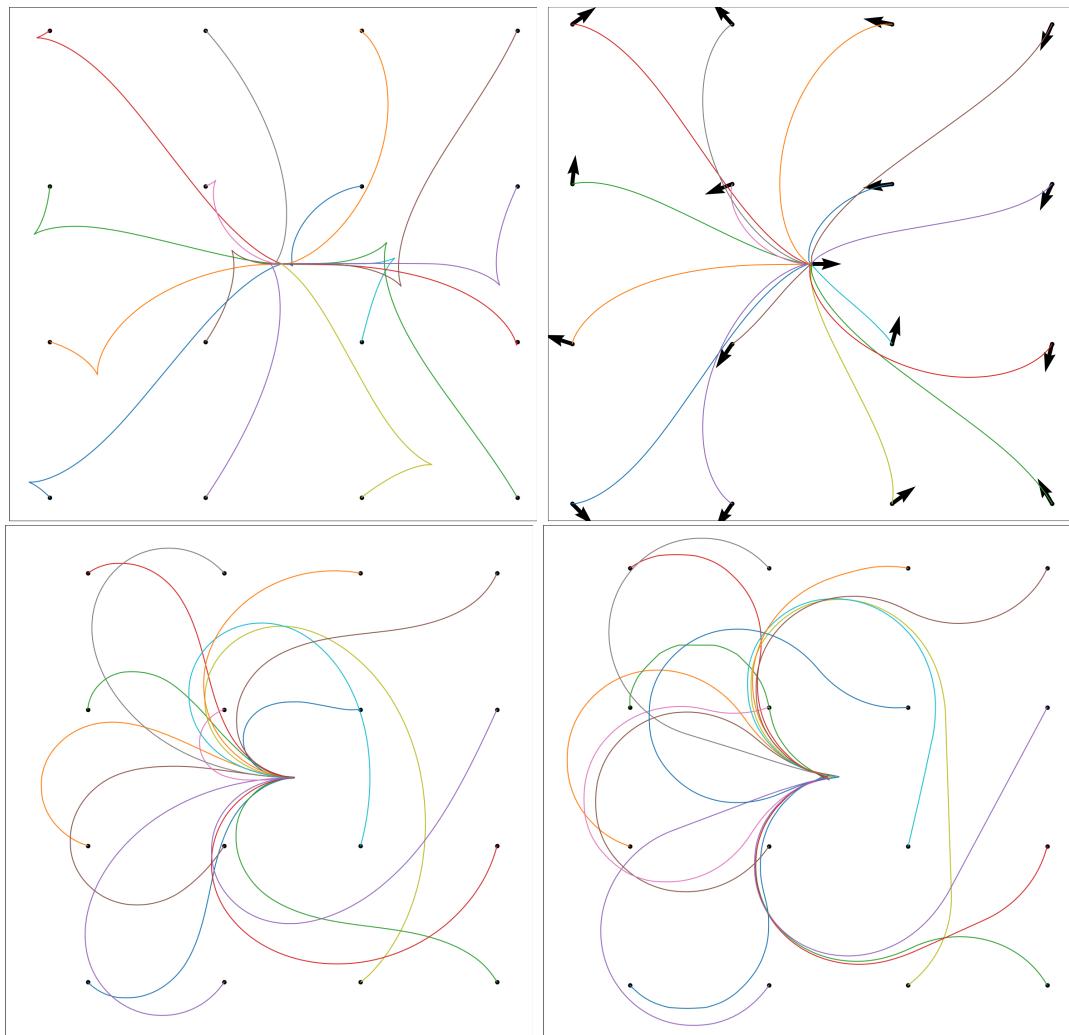


Figure 15: Projections planaires de géodésiques minimisantes pour les modèles de Reeds-Shepp, Reeds-Shepp forward, Elastica et Dubins (haut gauche, haut droit, bas gauche, bas droit). Point source  $(0, 0)$  avec tangente horizontale. Extrémités des géodésiques réparties uniformément dans le domaine, avec des orientations aléatoires (mais identiques pour tous les modèles).

correspond à la courbure neutre, et permet donc d'introduire de l'asymétrie dans le modèle. Si nécessaire,  $\xi = \xi(x, \theta)$  et  $\mu = \mu(x, \theta)$  peuvent dépendre de l'état courant  $(x, \theta) \in \Omega \times \mathbb{S}^1$ .

## 7.1 Modèle sous-Riemannien

Le modèle de Reeds-Shepp réversible possède une structure sous-Riemannienne [Mon06], c'est à dire que l'on peut reformuler l'énergie (78) comme une longueur

$$\mathcal{E}(\mathbf{x}, \boldsymbol{\theta}) = \int_0^T \rho(\mathbf{x}, \boldsymbol{\theta}) \|(\dot{\mathbf{x}}, \dot{\boldsymbol{\theta}})\|_{M_0(\mathbf{x}, \boldsymbol{\theta})}^2 dt, \quad (80)$$

dans une métrique définie par une forme quadratique positive et une contrainte linéaire

$$\|(\dot{\mathbf{x}}, \dot{\boldsymbol{\theta}})\|_{M_0(\mathbf{x}, \boldsymbol{\theta})}^2 = \begin{cases} \langle \dot{\mathbf{x}}, \mathbf{e}_\theta \rangle^2 + \dot{\boldsymbol{\theta}}^2 & \text{si } \det(\dot{\mathbf{x}}, \mathbf{e}_\theta) = 0, \\ +\infty & \text{sinon.} \end{cases}$$

Nous faisons référence à [Mon06] pour la définition formelle d'un modèle sous-riemannien, et notons qu'elle comporte de surcroît une hypothèse de contrôlabilité locale: les champs de vecteurs satisfaisant la contrainte linéaire, et leurs commutateurs jusqu'à une profondeur donnée, doivent engendrer l'espace. Cette hypothèse est bien satisfaite par le modèle de Reeds-Shepp.

On note que la quantité (80) est invariante par reparamétrisation, contrairement à (78) qui suppose une vitesse physique unité  $\|\mathbf{x}'\| = 1$ . Concernant la consistance de (80) avec (78), en restant informel: supposons que  $\mathbf{x}' = \pm \mathbf{e}_\theta$ , alors  $\det(\mathbf{x}', \mathbf{e}_\theta) = 0$  et  $\langle \mathbf{x}', \mathbf{e}_\theta \rangle^2 + \boldsymbol{\theta}'^2 = 1 + \boldsymbol{\theta}'^2 = \mathcal{C}(\boldsymbol{\theta}')^2$  comme souhaité.

**Relaxation Riemannienne.** La géométrie sous-Riemannienne peut être approchée à l'aide de métriques Riemanniennes, en remplaçant la contrainte linéaire par un terme de pénalisation. Dans le cas du modèle de Reeds-Shepp, on peut poser pour tout  $\varepsilon > 0$

$$\|(\dot{\mathbf{x}}, \dot{\boldsymbol{\theta}})\|_{M_\varepsilon(\mathbf{x}, \boldsymbol{\theta})}^2 = \langle \dot{\mathbf{x}}, \mathbf{e}_\theta \rangle^2 + \varepsilon^{-2} \det(\dot{\mathbf{x}}, \mathbf{e}_\theta)^2 + \dot{\boldsymbol{\theta}}^2.$$

Dans les expériences numériques et les applications de ces techniques, le choix  $\varepsilon = 1/10$  permet en général d'obtenir une approximation convaincante et exploitable du modèle de Reeds-Shepp correspondant à  $\varepsilon \rightarrow 0$ , voir [DMMP18, Mir19]. Pour être concret, donnons l'expression matricielle de la métrique qui est diagonale par blocs

$$M_\varepsilon(x, \theta) = \left( \begin{array}{cc|c} \mathbf{e}_\theta \mathbf{e}_\theta^\top + \varepsilon^{-2} \mathbf{e}_\theta^\perp (\mathbf{e}_\theta^\perp)^\top & 0 & 0 \\ 0 & 0 & 1 \end{array} \right) = \left( \begin{array}{cc|c} c^2 + \varepsilon^{-2} s^2 & (1 - \varepsilon^{-2}) c s & 0 \\ (1 - \varepsilon^{-2}) c s & s^2 + \varepsilon^{-2} c^2 & 0 \\ \hline 0 & 0 & 1 \end{array} \right)$$

où l'on a noté  $c = \cos(\theta)$  et  $s = \sin(\theta)$ .

**Équation eikonale.** La distance Riemannienne au bord du domaine satisfait une équation eikonale, dont nous avons étudié la résolution numérique Corollaire 5.12, et dont l'opérateur fait intervenir les matrices inverses de la métrique. Dans le cas du modèle de Reeds-Shepp, cette équation s'écrit

$$\|(\nabla_x u, \nabla_\theta u)\|_{D_\varepsilon(x, \theta)} = \rho(x, \theta) \quad \text{où} \quad \|(\hat{x}, \hat{\boldsymbol{\theta}})\|_{D_\varepsilon(x, \theta)}^2 = \langle \hat{x}, \mathbf{e}_\theta \rangle^2 + \varepsilon^2 \det(\hat{x}, \mathbf{e}_\theta)^2 + \hat{\boldsymbol{\theta}}^2.$$

L'équation eikonale garde son sens lorsque  $\varepsilon \rightarrow 0$ . Cependant, sa discréétisation via la décomposition de Selling, voir §3.4 et Corollaire 5.12, requiert des tenseurs définis positifs, donc  $\varepsilon > 0$ .

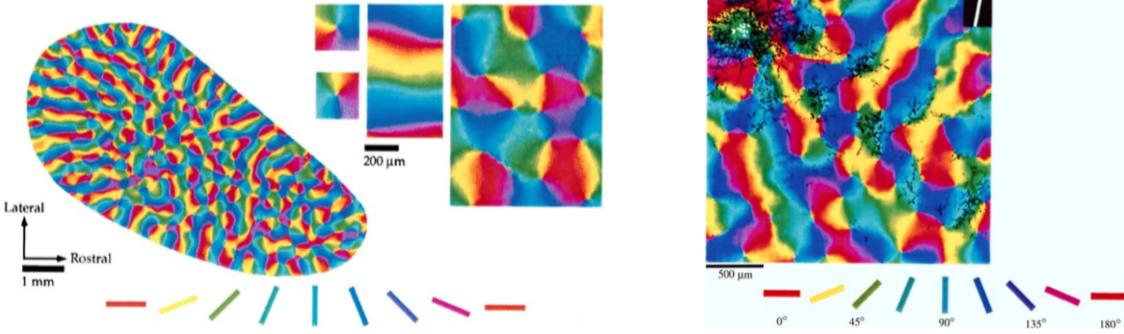


Figure 16: Gauche: orientation détectée par les neurones du cortex visuel V1. Droite: connexions au sein de V1.

**Le véhicule de Reeds-Shepp.** Considérons un véhicule muni de deux roues fixées aux extrémités d'un axe, comme un segway® ou une planche à roulettes de type "hoverboard", ou encore une chaise roulante (en négligeant les roulettes).

Les contraintes de modèle de Reeds-Shepp réversible correspondent bien à ces dispositifs: impossibilité des déplacements latéraux, possibilité d'avancer et de reculer. Par ailleurs (80) donne le nombre de tours de roues du véhicule le long de la trajectoire; en choisissant  $\rho$  constant et déterminé par le diamètre des roues, et  $\xi$  constant et déterminé par la longueur de l'entraxe.

**Le cortex visuel V1.** Le premier niveau du cortex visuel, nommé V1, est une structure du cerveau humain située au fond du crane, à l'extrémité du nerf optique issu du fond d'oeil. Les neurones du cortex visuel V1 réagissent préférentiellement à des stimuli ayant une certaine position dans le champ visuel, et orientés selon une certaine direction [BZSF97]. Pour cette raison, on peut leur associer un point du domaine tridimensionnel  $\Omega \times \mathbb{P}^1$ , où  $\Omega \subseteq \mathbb{R}^2$  représente le champ visuel, et  $\mathbb{P}^1 = [0, \pi[$  avec conditions au bord périodiques. En pratique, le neurone de coordonnées  $(x, \theta)$  s'active si un stimulus visuel de forme allongée est présent à la position  $x$  dans le champ visuel, et aligné avec l'orientation<sup>20</sup>  $\theta$ . Bien que paramétrés par l'espace tridimensionnel  $\Omega \times \mathbb{P}^1$ , les neurones du cortex visuel V1 sont arrangés sur une surface bi-dimensionnelle dans le cerveau humain, voir la (16, gauche).

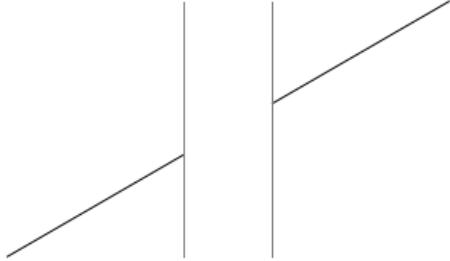
Il a été montré expérimentalement [BZSF97] que chaque neurone de paramètre  $(x, \theta)$  est lié préférentiellement à des neurones correspondant à une orientation proche:  $(x, \theta + \delta\theta)$ , et des neurones correspondant à une position proche dans la direction  $\theta$  donnée:  $(x + \delta t \mathbf{e}_\theta, \theta)$ , voir la Fig. 16, droite. Ces connections neuronales sont interprétées comme l'implémentation biologique de la structure sous-Riemannienne de Reeds-Shepp [Pet03]. Cette approche permet d'expliquer un certain nombres d'hallucinations visuelles, comme des défauts d'alignement apparents, appelés illusions de Poggendorf [FMCS17].

## 7.2 Autres modèles non-holonomes

Les énergies de chemin définies par (78) rentrent dans le cadre de la géométrie sous-Finslerienne. Le préfixe "sous" signifie que certaines directions dans l'espace des états ont un coût infini, c'est à dire que le modèle est non-holonomique, comme dans le cadre sous-Riemannien. Le terme

<sup>20</sup>L'utilisation de l'espace projectif  $\mathbb{P}^1 = [0, \pi[$  au lieu de la sphère  $\mathbb{S}^1 = [0, 2\pi[$  découle de l'impossibilité de discerner le bâtonnet d'orientation  $\theta$  du même bâtonnet retourné, d'orientation  $\theta + \pi$ .

First Poggendorf illusion



Round Poggendorf illusion

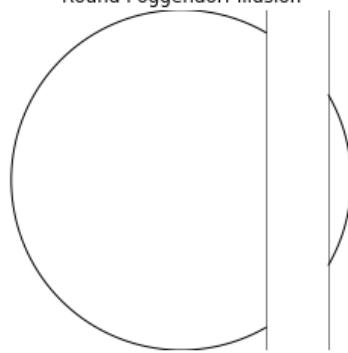
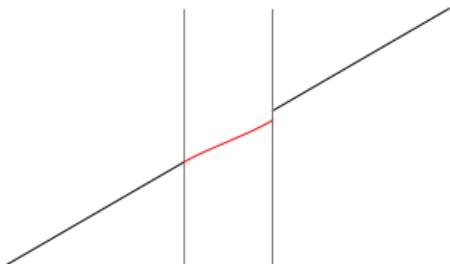


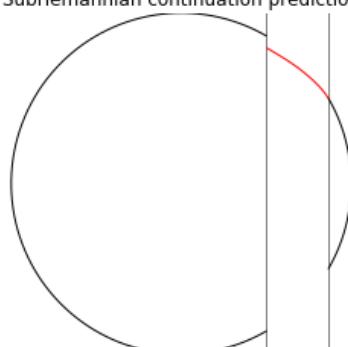
Figure 17: Haut : Deux illusions dues à Poggendorf. Les segments de droite semblent mal alignés, le morceaux de cercle ne semblent pas se rejoindre.

Bas : Interprétation de l'illusion via le modèle de Petitot-Citti-Sarti (nom du véhicule de Reeds-Shepp dans le contexte visuel). L'oeil humain cherche une continuation des segments géodésique (sous riemannienne) ayant les tangentes données.

Subriemannian continuation prediction



Subriemannian continuation prediction



“Finslerien” signifie que la métrique n’a pas une structure quadratique, contrairement au cadre Riemannien ou sous-Riemannien.

Les deux premiers paragraphes introduisent des éléments de géométrie sous-Finslerienne: définition des longueurs, des distances, de l’EDP eikonale et de l’EDO de backtracking permettant de retrouver les géodésiques. Le troisième paragraphe décrit les modèles d’intérêt dans ce cadre, Reeds-Shepp forward, Euler-Mumford, Dubins, tandis que le dernier paragraphe contraste les propriétés de contrôlabilité.

### Géométrie sous-Finslerienne

**Définition 7.2.** Une jauge est une application  $F : \mathbb{R}^d \rightarrow [0, \infty]$ , qui est convexe, 1-homogène, s’annule exclusivement en 0, et telle que  $B(F) := \{v \in \mathbb{R}^d; F(v) \leq 1\}$  est compact.

Une métrique sous-Finslerienne est une application  $\mathcal{F} : \overline{\Omega} \times \mathbb{R}^d$ , où  $\Omega \subseteq \mathbb{R}^d$  est un domaine, telle que  $v \in \mathbb{R}^d \mapsto \mathcal{F}_p(v)$  est une jauge pour tout  $p \in \overline{\Omega}$ , et telle que  $p \in \overline{\Omega} \mapsto B(\mathcal{F}_p)$  est continue pour la distance de Hausdorff.

Soit  $\mathcal{F}$  une métrique sous-Finslerienne sur un domaine  $\Omega$ , et soit  $\rho \in C^0(\overline{\Omega}, ]0, \infty[)$  une fonction coût. On pose pour tous  $p, q \in \overline{\Omega}$

$$\begin{aligned} \text{length}_{\mathcal{F}}(\gamma) &:= \int_0^1 \rho(\gamma(t)) \mathcal{F}_{\gamma(t)}(\gamma'(t)) dt, & \text{où } \gamma \in \Gamma := \text{Lip}([0, 1], \overline{\Omega}). \\ \text{dist}_{\mathcal{F}}(p, q) &:= \inf_{\gamma \in \Gamma_p^q} \text{length}_{\mathcal{F}}(\gamma), & \text{où } \Gamma_p^q := \{\gamma \in \Gamma; \gamma(0) = p, \gamma(1) = q\}. \end{aligned}$$

**Équation eikonale.** La distance depuis le bord du domaine  $\Omega$ , associée à une métrique sous-Finslerienne  $\mathcal{F}$ , est définie par

$$u(p) := \inf_{q \in \partial\Omega} d_{\mathcal{F}}(q, p).$$

Cette fonction n’est pas forcément continue, ni sur la frontière  $\partial\Omega$ , ni dans l’intérieur de  $\Omega$ . Cependant, elle satisfait au sens des solutions de viscosité discontinues [BCD08] la généralisation suivante de l’équation eikonale

$$\mathcal{F}_p^*(\nabla u(p)) = \rho(p), \quad \mathcal{F}_p^*(w) := \max\{\langle w, v \rangle; \mathcal{F}_p^*(v) \leq 1\}, \quad (81)$$

avec  $u = 0$  sur  $\partial\Omega$ . Le chemin minimal d’un point  $p \in \overline{\Omega}$  au bord peut s’extraire, sous réserve de différentiabilité, par la résolution d’une EDO rétrograde dite de backtracking

$$\gamma'(t) = V(\gamma(t)), \quad V(p) := d\mathcal{F}_p^*(\nabla u(p)),$$

avec condition terminale  $\gamma(T) = p$  où  $T = u(p)$ . Comme dans le cas Riemannien, voir la Proposition 4.4, elle peut s’interpréter en tant que descente de gradient implicite de la distance au bord  $u$ , dans la métrique  $\mathcal{F}$ , voir .

**Cas des véhicules de Reeds-Shepp (forward), Euler-Mumford et Dubins.** Ces modèles de véhicules sont posés sur un sous domaine de  $\mathbb{R}^2 \times \mathbb{S}^1$ , typiquement de la forme  $\Omega \times \mathbb{S}^1$ , dont les points prennent la forme  $p = (x, \theta)$ . L’expression du coût (78) suppose un paramétrage de la courbe  $(\mathbf{x}, \boldsymbol{\theta})$  dont la partie physique  $\mathbf{x}$  est de vitesse unité. Pour l’écrire dans le cadre sous-Finslerien, il faut introduire la métrique 1-homogène correspondante. Pour tout point  $p = (x, \theta)$ ,

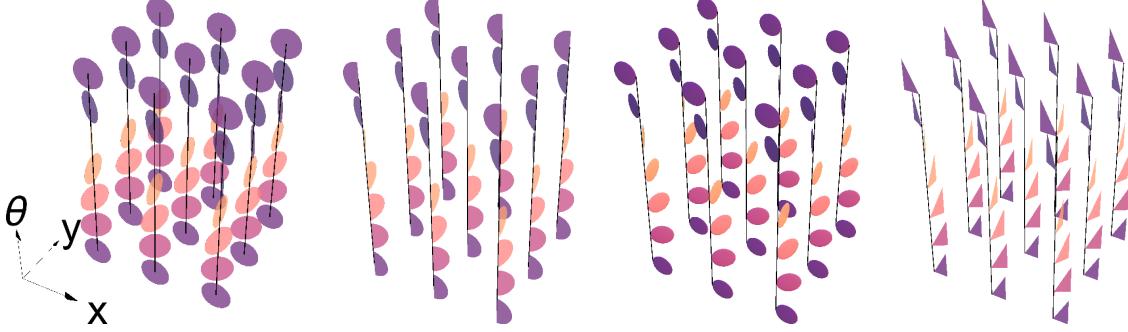


Figure 18: Ensembles de contrôle  $B(\mathcal{F}_{(x,\theta)})$  associés aux modèles de Reeds-Shepp réversible, Reeds-Shepp forward, Euler-Mumford et Dubins. Tous sont d'intérieur vide, ce qui reflète la non-holonomie des modèles.

et tout vecteur tangent  $\dot{p} = (\dot{x}, \dot{\theta}) \in \mathbb{R}^2 \times \mathbb{R}$

$$\mathcal{F}_p(\dot{p}) = \mathcal{F}_{(x,\theta)}(\dot{x}, \dot{\theta}) = \begin{cases} \lambda \mathcal{C}(\dot{\theta}/\lambda) & \text{si } \dot{x} = \lambda \mathbf{e}_\theta \text{ avec } \lambda > 0, \\ 0 & \text{si } \dot{x} = 0 \text{ et } \dot{\theta} = 0, \\ +\infty & \text{sinon.} \end{cases}$$

On peut vérifier que la fonction  $\dot{p} \mapsto \mathcal{F}_p(\dot{p})$  est convexe, pour chacun des coûts usuels (79). La métrique  $\mathcal{F}$  est aussi semi-continue inférieure, donc donne lieu à un problème de minimisation d'énergie bien posé, sauf dans le cas de Reeds-Shepp (forward) où la modification  $\mathcal{F}_{(x,\theta)}(0, \dot{\theta}) := |\dot{\theta}|$  est nécessaire.

L'équation eikionale, que nous souhaitons résoudre numériquement, fait intervenir la métrique duale (81). Celle-ci est définie par un problème d'optimisation, qui admet les simplifications suivantes

$$\mathcal{F}_{(x,\theta)}^*(\hat{x}, \hat{\theta}) = \sup_{(\dot{x}, \dot{\theta}) \neq (0,0)} \frac{\langle \hat{x}, \dot{x} \rangle + \hat{\theta} \dot{\theta}}{\mathcal{F}(\dot{x}, \dot{\theta})} = \sup_{\substack{(r, \dot{\theta}) \neq (0,0) \\ r \geq 0}} \frac{\langle \hat{x}, r \mathbf{e}_\theta \rangle + \hat{\theta} \dot{\theta}}{r \mathcal{C}(\dot{\theta}/r)} = \sup_{\dot{\theta} \in \mathbb{R}} \frac{\langle \hat{x}, \mathbf{e}_\theta \rangle + \hat{\theta} \dot{\theta}}{\mathcal{C}(\dot{\theta})} = f(\langle \hat{x}, \mathbf{e}_\theta \rangle, \hat{\theta}).$$

On a noté

$$f(\hat{s}, \hat{\theta}) := \sup_{\dot{\theta} \in \mathbb{R}} \frac{\hat{s} + \hat{\theta} \dot{\theta}}{\mathcal{C}(\dot{\theta})}.$$

Pour les coûts considérés (79), une étude de fonction permet de calculer ce supremum explicitement [MP19, Appendice A], et aboutit aux expressions explicites suivantes de  $f$ ,

$$\sqrt{\max\{0, \hat{s}\}^2 + \hat{\theta}^2} \quad (\hat{s} + \sqrt{\hat{s}^2 + \hat{\theta}^2})/2 \quad \max\{0, \hat{s} + \hat{\theta}, \hat{s} - \hat{\theta}\} \quad (82)$$

Noter que l'équation eikionale généralisée (81) correspond au choix  $\hat{s} := \langle \mathbf{e}_\theta, \nabla_x u \rangle$  et  $\hat{\theta} := \partial_\theta u$ , et s'écrit donc

$$f(\langle \mathbf{e}_\theta, \nabla_x u \rangle, \partial_\theta u) = \rho. \quad (83)$$

**Propriétés de contrôlabilité.** Les modèles de Reeds-Shepp réversible, Reeds-Shepp forward, Euler-Mumford et Dubins représentent des véhicules sujets à des contraintes de manœuvrabilité de plus en plus fortes. On peut traduire ces propriétés en termes de contrôlabilité, voir [AS13]

dans le cas général, et par exemple [DMMP18, Mir18] pour la discussion de ces modèles en particulier. Pour être concret, considérons les points de  $p_0, p_\delta$  du domaine  $\Omega_r \subseteq \mathbb{R}^2 \times \mathbb{S}^1$  définis par

$$p_0 := (0, 0, 0), \quad p_\delta := (0, \delta, 0), \quad \Omega_r := B(0, r) \times \mathbb{S}^1,$$

où on suppose toujours  $0 < \delta < r \leq 2$ . On peut montrer les estimations suivantes, qui sont optimales, sur le coût de déplacement  $d_{\mathcal{F}}(p_0, p_\delta; \Omega_r)$  de  $p_0$  à  $p_\delta$  via un chemin restreint au domaine  $\Omega_r$ .

- Les modèles Riemanniens, quelle que soit leur métrique  $\mathcal{M}$  (supposée continue), sont localement équivalents à la métrique euclidienne. On a donc, pour  $0 < \delta < r \leq 2$

$$d_{\mathcal{M}}(p_0, p_\delta; \Omega_r) = \mathcal{O}(\delta).$$

- Le modèle de Reeds-Shepp reversible est sous-Riemannien, et pour cette raison il est *localement contrôlable*. Il satisfait

$$d_{RS}(p_0, p_\delta; \Omega_r) = \mathcal{O}(\sqrt{\delta}).$$

Un déplacement admissible entre  $p_0$  et  $p_\delta$ , et de coût optimal à une constante multiplicative près, s'obtient par commutation des contrôles du véhicule : tourner, avancer, tourner dans le sens inverse, reculer. Le théorème de Chow [Mon06] étend cette estimation et cette construction aux modèles sous Riemanniens généraux.

- Le modèle de Reeds-Shepp forward, en revanche, requiert un coût borné inférieurement pour se déplacer entre les points considérés, pourtant arbitrairement proches : il n'est pas localement contrôlable. On a quel que soient  $0 < \delta < r \leq 2$

$$d_{RSF}(p_0, p_\delta; \Omega_r) = \mathcal{O}(1).$$

Un mouvement admissible consiste à tourner de  $\pi/2$ , avancer de  $\delta$ , puis tourner de  $-\pi/2$ . Ce mouvement est contenu dans  $\overline{B}_{\mathbb{R}^2}(0, \delta) \times \mathbb{S}^1 \subseteq \Omega_r$ , c'est à dire que la coordonnée de position  $x$  reste locale, mais pas celle d'orientation  $\theta$ .

- Le modèle de Euler-Mumford peut, comme le précédent, manœuvrer dans le domaine spatial restreint défini par  $\Omega_r$ . En revanche, le coût de déplacement augmente lorsque l'espace disponible se réduit : pour tous  $0 < \delta < r \leq 2$

$$d_{EM}(p_0, p_\delta; \Omega_r) = \mathcal{O}(1/r).$$

- Le modèle de Dubins, initié au centre  $p_0$  d'une boule ouverte  $\Omega_r = B(0, r) \times \mathbb{S}^1$  de rayon  $r$ , ne peut y manœuvrer pour rejoindre  $p_\delta$  que si  $r > 1$ . En effet, le rayon de courbure de ses trajectoires est borné inférieurement par 1. Ainsi pour tous  $0 < \delta < r \leq 2$ ,

$$d_{EM}(p_0, p_\delta; \Omega_r) = \begin{cases} \mathcal{O}(1) & \text{si } r > 1, \\ +\infty & \text{sinon.} \end{cases}$$

### 7.3 Discrétisation de l'équation eikonale généralisée

On présente une approche de la discrétisation des équations eikonales généralisées associées aux modèles considérés dans ce chapitre, issue de [DMMP18, Mir18, MP19]. On note  $p = (x, \theta) \in \mathbb{R}^2 \times \mathbb{S}^1$  un élément de l'espace des positions et orientations, et  $\dot{p} = (\dot{x}, \dot{\theta}) \in \mathbb{R}^2 \times \mathbb{R}$  un vecteur tangent à cet espace.

**Une classe de schémas monotones et causaux.** Pour discréteriser de manière monotone et causale une équation eikonale généralisée  $\mathcal{F}_p^*(\nabla u(p)) = \rho(p)$ , on choisit de chercher une approximation de la métrique duale sous la forme [MP19]

$$\mathcal{F}_p^*(\hat{p})^2 = \max_{1 \leq k \leq K} \left( \sum_{1 \leq i \leq I} \alpha_{ik} \langle \hat{p}, e_{ik} \rangle_+^2 + \sum_{1 \leq j \leq J} \beta_{jk} \langle \hat{p}, f_{jk} \rangle^2 \right). \quad (84)$$

Dans cette écriture, les meta-paramètres  $I, J, K$  sont des entiers non-négatifs, les poids  $\alpha_{ik}, \beta_{jk} \in \mathbb{R}_+$  sont positifs, et les offsets  $e_{ik}, f_{jk} \in \mathbb{Z}^3$  sont à coordonnées entières. Parfois, un terme d'erreur  $\mathcal{O}(\varepsilon^2)$  est présent, lié à un paramètre de relaxation utilisé dans le cadre de la discréterisation. On en déduit de la forme (84) de la métrique le schéma aux différences finies

$$\begin{aligned} \mathcal{F}_p^*(\nabla u(p))^2 &= \max_{1 \leq k \leq K} \left( \sum_{1 \leq i \leq I} \alpha_{ik} \max \left\{ 0, \frac{u(p) - u(p - he_{ik})}{h} \right\}^2 \right. \\ &\quad \left. + \sum_{1 \leq j \leq J} \beta_{jk} \max \left\{ 0, \frac{u(p) - u(p + hf_{jk})}{h}, \frac{u(p) - u(p - hf_{jk})}{h} \right\}^2 \right) + \mathcal{O}(rh), \end{aligned} \quad (85)$$

où  $r := \max\{\|e_{ik}\|, \|f_{jk}\|; \forall i, j, k\}$  est le diamètre du stencil. Ce schéma est bien monotone, car c'est une fonction croissante de différences finies de la forme  $(u(p) - u(p + he))_{e \neq 0}$ . Il est causal car il ne dépend que des parties positives de ces différences finies. (Comme discuté §6.3, l'utilité principale de la causalité est d'appliquer la méthode de résolution en une passe nommée fast marching.)

La discréterisation de l'équation eikonale Riemannienne présentée au Corollaire 5.12 est bien de cette forme, avec  $K = 1$ ,  $I = 0$ , et  $J = d(d+1)/2$  si la décomposition de Selling est utilisée. Le schéma usuel pour l'équation eikonale isotrope [RT92] rentre également dans ce cadre (3), avec  $K = 1$ ,  $I = 0$ , et  $J = d$  et l'utilisation de la base canonique comme offsets  $(f_{j1})_{j=1}^d$ . De même pour la variante considérée dans [Set96] qui utilise  $K = 1$ ,  $I = 2d$  et  $J = 0$ .

**Modèle de Reeds-Shepp.** Considérons l'équation eikonale associée au modèle de Reeds-Shepp forward, qui se déduit de (82).

$$\mathcal{F}_p(\nabla u(p))^2 = f(\langle \nabla_x u, \mathbf{e}_\theta \rangle, \partial_\theta u)^2 = \max\{0, \langle \nabla_x u, \mathbf{e}_\theta \rangle\}^2 + (\partial_\theta u)^2$$

Le résultat suivant permet d'approcher le terme  $\max\{0, \langle \nabla_x u, \mathbf{e}_\theta \rangle\}^2$ .

**Lemme 7.3** (Proposition 4.14 dans [Mir18]). *Soit  $\mathbf{e} \in \mathbb{R}^d$  un vecteur unité, et soit  $\varepsilon > 0$ . Définissons*

$$D := \mathbf{e}\mathbf{e}^\top + \varepsilon^2 P_{\mathbf{e}}, \quad \text{où } P_{\mathbf{e}} := \text{Id} - \mathbf{e}\mathbf{e}^\top.$$

*Noter que  $P_{\mathbf{e}}$  est le projecteur orthogonal sur  $\{\mathbf{e}\}^\perp$ . Considérons une décomposition, comme celle de Selling, telle que*

$$D = \sum_{1 \leq i \leq I} \rho_i e_i e_i^\top,$$

*avec  $\rho_i \geq 0$  et  $e_i \in \mathbb{Z}^d$ . On suppose de plus  $\langle e_i, \mathbf{e} \rangle \geq 0$  pour tout  $1 \leq i \leq I$ . Alors pour tout  $w \in \mathbb{R}^d$*

$$\max\{0, \langle w, \mathbf{e} \rangle\}^2 \leq \sum_{1 \leq i \leq I} \rho_i \max\{0, \langle w, e_i \rangle\}^2 \leq \max\{0, \langle w, \mathbf{e} \rangle\}^2 + \varepsilon^2 \|P_{\mathbf{e}}w\|^2.$$

En choisissant  $\mathbf{e} = \mathbf{e}_\theta$  et  $w = \nabla_x u$  dans le Lemme 7.3 on obtient

$$\max\{0, \langle \nabla_x u, \mathbf{e}_\theta \rangle\}^2 = \sum_{1 \leq i \leq I} \rho_i \max\left\{0, \frac{u(x, \theta) - u(x - he_i, \theta)}{h}\right\}^2 + \mathcal{O}(\varepsilon^2 + rh),$$

où  $r = \max_{i=1}^I \|e_i\|$ . De plus, dans le cas de la décomposition de Selling, on a  $r \leq C \text{Cond}(D) = C\varepsilon^{-1}$ . En rajoutant l'approximation du terme angulaire

$$(\partial_\theta u)^2 = \max\{0, u(x, \theta) - u(x, \theta - h), u(x, \theta) - u(x, \theta + h)\}^2/h^2 + \mathcal{O}(h),$$

on obtient une discrétisation du modèles de Reeds-Shepp forward avec  $K = 1, I = 3, J = 1$ .

**Modèle de Dubins.** L'équation eikonale généralisée associée au modèle de Dubins se déduit de (82), et prend la forme

$$\mathcal{F}_p^*(\nabla u(p))^2 = \max\{\langle \nabla u(p), v^+(\theta) \rangle_+^2, \langle \nabla u(p), v^-(\theta) \rangle_+^2\},$$

où  $a_+ := \max\{0, a\}$ , et où on a noté

$$v^\pm(\theta) := (\cos \theta, \sin \theta, \pm 1).$$

Le Lemme 7.3 permet de discrétiser cette EDP, en choisissant  $\mathbf{e} = v^\pm(\theta)/\|v^\pm(\theta)\|$ . Le schéma obtenu a pour paramètres  $I = 6, J = 0$ , et  $K = 2$ .

**Modèle de Euler-Mumford.** L'équation eikonale généralisée associée au modèle d'Euler-Mumford fait intervenir l'opérateur

$$2\mathcal{F}_p^*(\nabla u(p)) = \langle \nabla_x u(p), \mathbf{e}_\theta \rangle + \sqrt{\langle \nabla_x u(p), \mathbf{e}_\theta \rangle^2 + (\partial_\theta u(p))^2},$$

qui se déduit de (82). Il existe une seconde expression, équivalente, de cet opérateur sous la forme d'une intégrale [Mir18].

$$\mathcal{F}_p^*(\nabla u(p))^2 = \frac{3}{4} \int_{-\pi/2}^{\pi/2} \max\{0, \langle \nabla u(p), \mathbf{e}(\theta, \varphi) \rangle\}^2 \cos \varphi d\varphi,$$

où  $\mathbf{e}(\theta, \varphi) := (\cos \theta \cos \varphi, \sin \theta \cos \varphi, \sin \varphi)$ . On peut discrétiser cette intégrale, un utilisant par exemple la formule de quadrature de Fejer à  $L$  points, qui permet de la remplacer par une somme sur des angles  $\varphi_l = (l - \frac{L+1}{2})\pi/L$ ,  $1 \leq l \leq L$ , munis de poids  $(\omega_l)_{l=1}^L$  adéquats. En utilisant de nouveau Lemme 7.3 pour approcher la quantité intégrée, on obtient un schéma de paramètres  $I = 6L, J = 0, K = 1$ , voir [Mir18]. En pratique, choisir  $L = 5$  donne une approximation adéquate du modèle pour les applications considérées.

## 7.4 Applications

Le calcul de chemins optimaux avec pénalisation de courbure trouve un certain nombre d'applications.

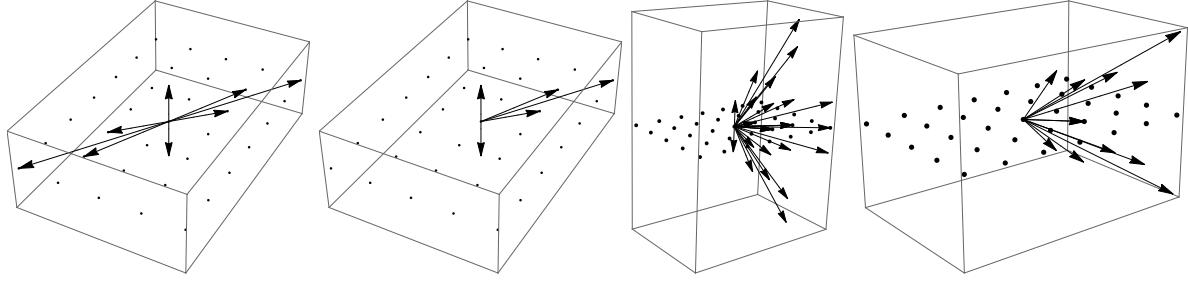


Figure 19: Discretization stencils used for the Reeds-Shepp reversible, Reeds-Shepp forward, Euler-Mumford, and Dubins models. Note the sparseness and anisotropy of the stencils. Model parameters:  $\theta = \pi/3$ ,  $\xi = 0.2$ ,  $\varepsilon = 0.1$ .

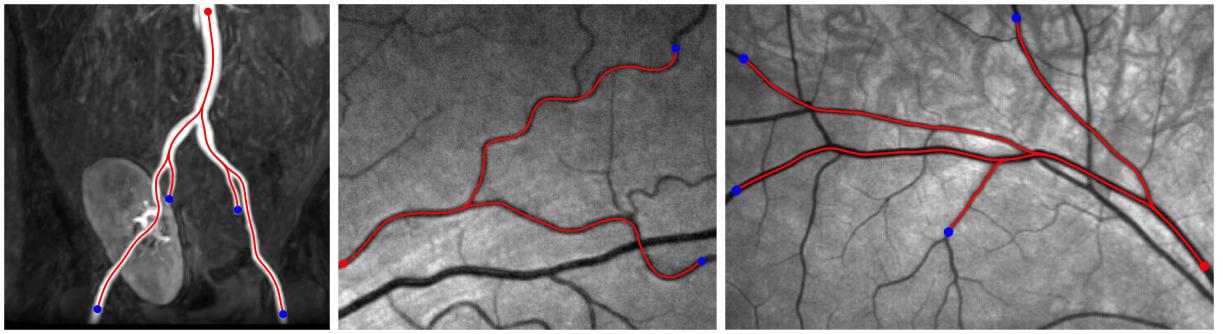


Figure 20: Segmentation de structures tubulaires, via le modèle Euler-Mumford elastica et un coût approprié.

**Segmentation de structures tubulaires, et de contours d'objets.** La pénalisation de la courbure est naturelle dans le cadre de ces problèmes, et est d'ailleurs un ingrédient des premières approches proposées [KWT88]. Cependant celles-ci font intervenir la minimisation de fonctionnelles non-convexes, par des méthodes d'optimisation locale, qui sont très sensibles à l'initialisation et aux perturbations induites par le bruit dans l'image. Ces fonctionnelles sont typiquement la somme d'un terme régularisant, et d'un terme d'attache aux données.

L'utilisation de modèles de chemins plus simples, initialement isotropes et sans pénalisation de courbure, mais pouvant être optimisés globalement grâce au formalisme de l'équation eikionale, a permis des progrès importants en robustesse [CK97]. Le développement des techniques numériques a permis de redonner des degrés de libertés dans la conception des énergies, notamment par l'introduction d'anisotropie [BC11]. Une technique complémentaire est le relèvement dans des espaces de paramètres reflétant certaines propriétés des structures extraites, comme le rayon [LY07, CMC16], le niveau de gris [CC15], ou l'orientation [PKP09].

Le calcul de chemins minimisant globalement une énergie faisant intervenir la courbure [DMMP18, CMC17] permet, paradoxalement, de revenir à des modèles proches de ceux proposés initialement, mais avec la garantie d'une optimisation globale.

**Etude des illusions visuelles.** Comme discuté plus en détail § 7.1, le fonctionnement du cortex visuel V1 a certaines analogies avec le modèle de Reeds-Shepp [Pet03], ce qui permet d'interpréter certaines illusions visuelles [FMCS17], voir la Fig. 17.

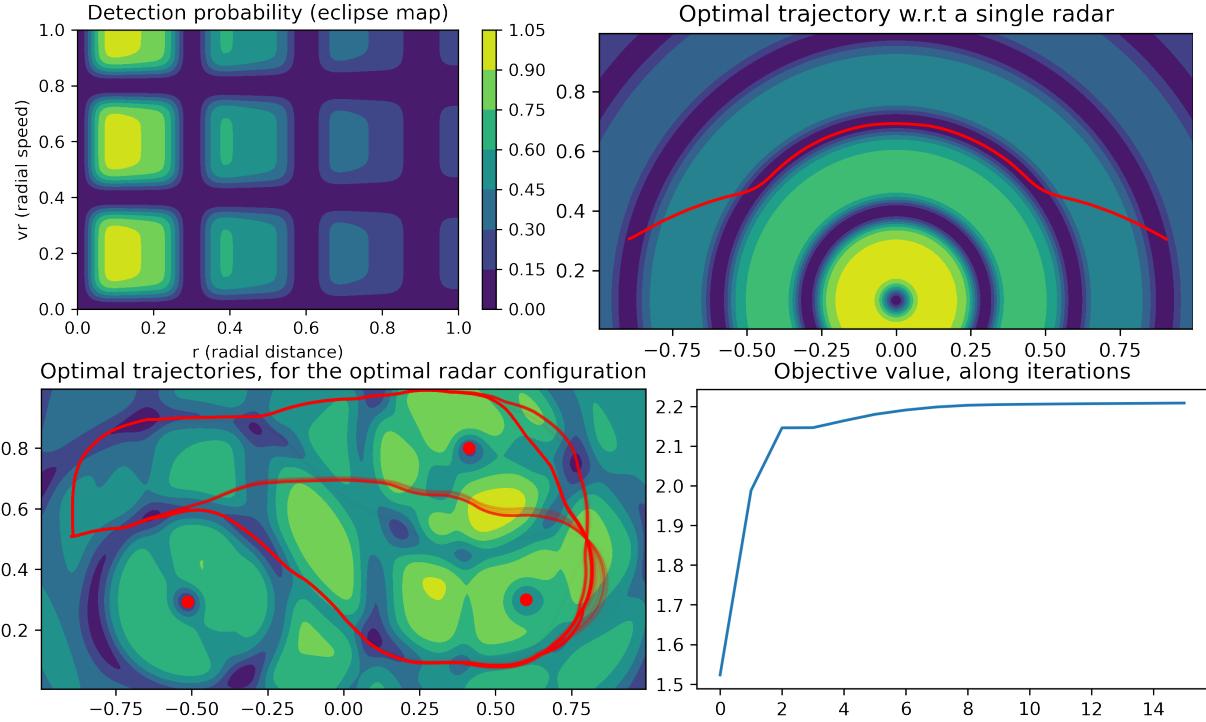


Figure 21: (Haut gauche) Modèle de probabilité de détection d'un véhicule par un radar, en fonction de sa distance radiale et de sa vitesse radiale. Noter que certaines distances et vitesses radiales, distribuées périodiquement, sont *aveugles*. (Haut droite) Trajectoire minimisant la probabilité de détection entre deux points, en présence d'un radar unique. C'est, de manière approchée, une concaténation de cercles, dont le rayon est une distance aveugles, et de spirales, dont la vitesse radiale est aveugle pour le radar. (Bas gauche) Configuration de trois radars, localement optimisée, pour détecter des trajectoires aller-retour du point source gauche au point cible droit. (Bas droite) Valeur de la fonction objectif, au fil de l'optimisation par montée de gradient.

**Planification de mouvement ou interception de véhicules.** Le véhicule de Dubins, dont les trajectoires ont un rayon de courbure borné, est un modèle populaire en avionique. Sa résolution permet de planifier des approches, ou à l'inverse de concevoir des dispositifs de détection de véhicules souhaitant passer inaperçus mais soumis à ces contraintes de mouvement [MD17, DDBM19], voir la Fig. 21.

## References

- [ABCM01] Fuensanta Andreu, Coloma Ballester, Vicent Caselles, and José M Mazón. Minimizing total variation flow. *Differential and integral equations*, 14(3):321–360, 2001.
- [AF11] John Ashburner and Karl J Friston. Diffeomorphic registration using geodesic shooting and Gauss–Newton optimisation. *NeuroImage*, 55(3):954–967, 2011.
- [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.

- [All92] Grégoire Allaire. Homogenization and two-scale convergence. *SIAM Journal on Mathematical Analysis*, 23(6):1482–1518, 1992.
- [AS13] Andrei A Agrachev and Yuri Sachkov. *Control theory from the geometric viewpoint*, volume 87. Springer Science and Business Media, 2013.
- [BC11] Fethallah Benmansour and Laurent D. Cohen. Tubular structure segmentation based on minimal path method and anisotropic enhancement. *International Journal of Computer Vision*, 92(2):192–210, 2011.
- [BCD08] Martino Bardi and Italo Capuzzo-Dolcetta. *Optimal control and viscosity solutions of Hamilton-Jacobi-Bellman equations*. Springer Science & 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.
- [BR06] Folkmar Bornemann and Christian Rasch. Finite-element Discretization of Static Hamilton-Jacobi Equations based on a Local Variational Principle. *Computing and Visualization in Science*, 9(2):57–69, June 2006.
- [Bre11] Haïm Brezis. *Functional analysis, Sobolev spaces and partial differential equations*, volume 2. Springer, 2011.
- [BS91] Guy Barles and Panagiotis E Souganidis. Convergence of approximation schemes for fully nonlinear second order equations. *Asymptotic analysis*, 4(3):271–283, 1991.
- [BZSF97] William H Bosking, Ying Zhang, Brett Schofield, and David Fitzpatrick. Orientation selectivity and the arrangement of horizontal connections in tree shrew striate cortex. *Journal of neuroscience*, 17(6):2112–2127, 1997.
- [CC15] Da Chen and Laurent D. Cohen. Interactive retinal vessel centreline extraction and boundary delineation using anisotropic fast marching and intensities consistency. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 4347–4350. IEEE, 2015.
- [CCM16] Laurent D. Cohen, Da Chen, and Jean-Marie Mirebeau. Finsler Geodesics Evolution Model for Region based Active Contours. In Edwin R Hancock Richard C Wilson and William A P Smith, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 22.1–22.12. BMVA Press, September 2016.
- [CIL92] Michael G Crandall, Hitoshi Ishii, and Pierre-Louis Lions. User’s guide to viscosity solutions of second order partial differential equations. *27(1):1–67*, 1992.
- [CK97] Laurent D. Cohen and R Kimmel. Global minimum for active contour models: A minimal path approach. *International Journal of Computer Vision*, 24(1):57–78, 1997.

- [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.
- [CLPQ20] Keenan Crane, Marco Livesu, Enrico Puppo, and Yipeng Qin. A Survey of Algorithms for Geodesic Paths and Distances. *arXiv preprint arXiv:2007.10430*, 2020.
- [CMA<sup>+</sup>20] Paul Cupillard, Wim Mulder, Pierre Anquez, Antoine Mazuyer, and J Barthélémy. The Apparent Anisotropy of the SEG-EAGE Overthrust Model. In *82nd EAGE Annual Conference and Exhibition*, pages 1–5. European Association of Geoscientists and Engineers, 2020.
- [CMC16] Da Chen, Jean-Marie Mirebeau, and Laurent D. Cohen. Vessel tree extraction using radius-lifted keypoints searching scheme and anisotropic fast marching method. *Journal of Algorithms & Computational Technology*, 10(4):224–234, July 2016.
- [CMC17] Da Chen, Jean-Marie Mirebeau, and Laurent D. Cohen. Global Minimum for a Finsler Elastica Minimal Path Approach. *International Journal of Computer Vision*, 122(3):458–483, 2017.
- [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, Clarisse 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.
- [DDBM19] Johann Dreo, François Desquillet, Frédéric Barbaresco, and Jean-Marie Mirebeau. Netted multi-function radars positioning and modes selection by non-holonomic fast marching computation of highest threatening trajectories. In *International RADAR’19 conference*, 2019.
- [DMMP18] Remco Duits, Stephan PL Meesters, Jean-Marie Mirebeau, and Jorg M Portegies. Optimal paths for variants of the 2D and 3D Reeds-Shepp car with applications in image analysis. *Journal of Mathematical Imaging and Vision*, pages 1–33, 2018.
- [DS13] Alexander Munro Davie and Andrew James Stothers. Improved bound for complexity of matrix multiplication. *Proceedings of the Royal Society of Edinburgh Section A: Mathematics*, 143(2):351–369, 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.
- [Fig17] Alessio Figalli. The Monge–Ampère equation and its applications. European Mathematical Society Publishing House, 2017.

- [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.
- [FMCS17] B Franceschiello, A Mashtakov, G Citti, and A Sarti. Modelling of the Poggendorff Illusion via Sub-Riemannian Geodesics in the Roto-Translation Group. In *International Conference on Image Analysis and Processing*, pages 37–47. Springer, 2017.
- [GHZ18] Daniel Ganellari, Gundolf Haase, and Gerhard Zumbusch. A massively parallel Eikonal solver on unstructured meshes. *Computing and Visualization in Science*, 19(5-6):3–18, 2018.
- [GW08] Andreas Griewank and Andrea Walther. *Evaluating derivatives: principles and techniques of algorithmic differentiation*. SIAM, 2008.
- [Ham19] Brittany Froese Hamfeldt. Convergence Framework for the Second Boundary Value Problem for the Monge–Ampère Equation. *SIAM Journal on Numerical Analysis*, 57(2):945–971, 2019.
- [Hop50] Eberhard Hopf. The partial differential equation  $ut + uux = \mu xx$ . *Communications on Pure and Applied Mathematics*, 3(3):201–230, 1950.
- [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.
- [JW08] Won-Ki Jeong and Ross T Whitaker. A Fast Iterative Method for Eikonal Equations. *SIAM Journal on Scientific Computing*, 30(5):2512–2534, July 2008.
- [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.
- [KS98] R Kimmel and James A. Sethian. Computing geodesic paths on manifolds. *Proceedings of the National Academy of Sciences*, 95(15):8431–8435, July 1998.
- [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.
- [KWT88] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331, January 1988.
- [LBBMV17] P Le Bouteiller, M Benjema, L Métivier, and J Virieux. An accurate discontinuous Galerkin method for solving point-source Eikonal equation in 2-D heterogeneous anisotropic media. *Geophysical Journal International*, 212(3):1498–1522, 2017.

- [Lin13] Tony Lindeberg. *Scale-space theory in computer vision*, volume 256. Springer Science and Business Media, 2013.
- [LN18] Wenbo Li and Ricardo H Nochetto. Optimal Pointwise Error Estimates for Two-Scale Methods for the Monge–Ampère Equation. *SIAM Journal on Numerical Analysis*, 56(3):1915–1941, 2018.
- [LY07] H Li and A Yezzi. Vessels as 4-D curves: Global minimal 4-D paths to extract 3-D tubular surfaces and centrelines. *IEEE Transactions on Medical Imaging*, 26(9):1213–1223, 2007.
- [MD17] Jean-Marie Mirebeau and Johann Dreо. Automatic differentiation of non-holonomic fast marching for computing most threatening trajectories under sensors surveillance. In *International Conference on Geometric Science of Information*, pages 791–800. Springer, 2017.
- [Mir14a] Jean-Marie Mirebeau. Anisotropic Fast-Marching on cartesian grids using Lattice Basis Reduction. *SIAM Journal on Numerical Analysis*, 52(4):1573–1599, January 2014.
- [Mir14b] Jean-Marie Mirebeau. Efficient fast marching with Finsler metrics. *Numerische Mathematik*, 126(3):515–557, 2014.
- [Mir18] Jean-Marie Mirebeau. Fast-marching methods for curvature penalized shortest paths. *Journal of Mathematical Imaging and Vision*, 60(6):784–815, 2018.
- [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.
- [Mon06] Richard Montgomery. *A Tour of Subriemannian Geometries, Their Geodesics and Applications*. American Mathematical Soc., August 2006.
- [MP19] Jean-Marie Mirebeau and Jorg Portegies. Hamiltonian fast marching: A numerical solver for anisotropic and non-holonomic eikonal pdes. *Image Processing On Line*, 9:47–93, 2019.
- [MV11] Bertrand Maury and Juliette Venel. A discrete contact model for crowd motion. *ESAIM: Mathematical Modelling and Numerical Analysis*, 45(1):145–168, 2011.
- [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.

- [Pet03] Jean Petitot. The neurogeometry of pinwheels as a sub-Riemannian contact structure. *Journal of Physiology-Paris*, 97(2-3):265–309, March 2003.
- [PKP09] M Pechaud, R Keriven, and G Peyré. Extraction of tubular structures over an orientation domain. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)*, pages 336–342. IEEE, 2009.
- [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.
- [PPKC10] G Peyré, M Pechaud, R Keriven, and Laurent D. Cohen. Geodesic methods in computer vision and graphics. *Foundations and Trends® in Computer Graphics and Vision*, 5(3-4):197–397, 2010.
- [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<sup>+</sup>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 für die Reine und Angewandte Mathematik*, 77:143–229, 1874.
- [Set96] James A. Sethian. A fast marching level set method for monotonically advancing fronts. *Proceedings of the National Academy of Sciences*, 93(4):1591–1595, 1996.
- [Set99] James A. Sethian. *Level Set Methods and Fast Marching Methods: Evolving Interfaces in Computational Geometry, Fluid Mechanics, Computer Vision, and Materials Science*. Cambridge University Press, 1999.
- [SKB01] N Sochen, R Kimmel, and A M Bruckstein. Diffusions and confusions in signal and image processing . *Journal of Mathematical Imaging and Vision*, 14(3):195–209, 2001.
- [SV03] James A. Sethian and Alexander Boris Vladimirs. Ordered upwind methods for static Hamilton-Jacobi equations: theory and algorithms. *SIAM Journal on Numerical Analysis*, 41(1):325–363, 2003.
- [Tsi95] J.N. Tsitsiklis. Efficient algorithms for globally optimal trajectories. *IEEE transactions on Automatic Control*, 40(9):1528–1538, September 1995.
- [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.
- [XXC12] Junyuan Xie, Linli Xu, and Enhong Chen. Image denoising and inpainting with deep neural networks. In *Advances in neural information processing systems*, pages 341–349, 2012.
- [Zha05] Hongkai Zhao. A fast sweeping method for eikonal equations. *Mathematics of computation*, 74(250):603–627, 2005.