EDPs numériques pour l'analyse d'images

Jean-Marie Mirebeau*

January 26, 2021

Abstract

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

Contents

1 Introduction

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

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

1.1 Pertinence des EDPs en traitement d'image

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

- *Physique des dispositifs d'acquisition.* De nombreux dispositifs d'imagerie sont fondés sur la physique des milieux continus (vibrations, absorption lumineuse, ...) qui est décrite de manière directe par des EDPs. En particulier les méthodes de *tomographie*, qui consistent à "reconstruire volume d'un objet à partir d'une série de mesures effectuées depuis l'extérieur de cet objet" (Wikipedia), utilisées en imagerie médicale ou sismique.
  Exemple : La transformée de Radon, correspondant à la "tomographie axiale".
  Exemple : La reconstruction de relief à partir des ombres (Shape from shading).

- *Modèle interprétable et explicatif.* Par analogie avec des modèles physiques, les EDPs permettent de définir un cadre interprétable et explicatif pour le traitement de données.
  Exemple : Description multi-échelle d'une image, via un processus de diffusion.
  Exemple : Les courbes elastica de Euler, qui correspondent à la position de repos d'une barre élastique, sont utilisées pour l'extraction de contours et de courbes dans des images.

- *Reconstruction d'information globale.* On peut considérer une EDP comme un système d'équations couplées, où chaque point du domaine porte (typiquement) une inconnue et une équation. La résolution d'un tel système est un objet global, qui synthétise les contraintes imposées localement.
  Exemple : La résolution de l'équation eikonale permet de déterminer le plus court chemin dans un domaine contenant des obstacles et des zones plus ou moins rapides.

- *Structure mathématique, garanties.* Les EDPs ont une structure mathématique riche qui permet d'établir un certain nombre de garanties : terminaison de l'algorithme, robustesse au bruit, validité sur des cas simples, etc

1.2 Les EDPs considérées dans ce cours

Dans ce cours nous étudierons les schémas de discrétisation, 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

$$\partial_t u = \Delta u, \qquad \|\nabla u\| = 1, \tag{1}$$

avec des conditions au bord appropriées.

- L'équation de la chaleur s'interpête comme le flot gradient (descente de gradient) de l'énergie de Dirichlet $\int_\Omega \|\nabla u\|^2$. Elle lisse les changements brusques dans la fonction u; dans le cas d'une image elle élimine le bruit, mais rend flous les contours d'objets et les textures oscillantes. L'équation de la chaleur permet de séparer, au fil du temps, les différentes échelles d'une image.

- L'équation eikonale caractérise la fonction distance euclidienne u à un point ou à une région source. Le gradient de sa solution $\nabla u(x)$ donne la direction opposée à la source. L'équation eikonale est le modèle le plus simple de propagation de front, et remonter son gradient permet de calculer des plus courts chemins.

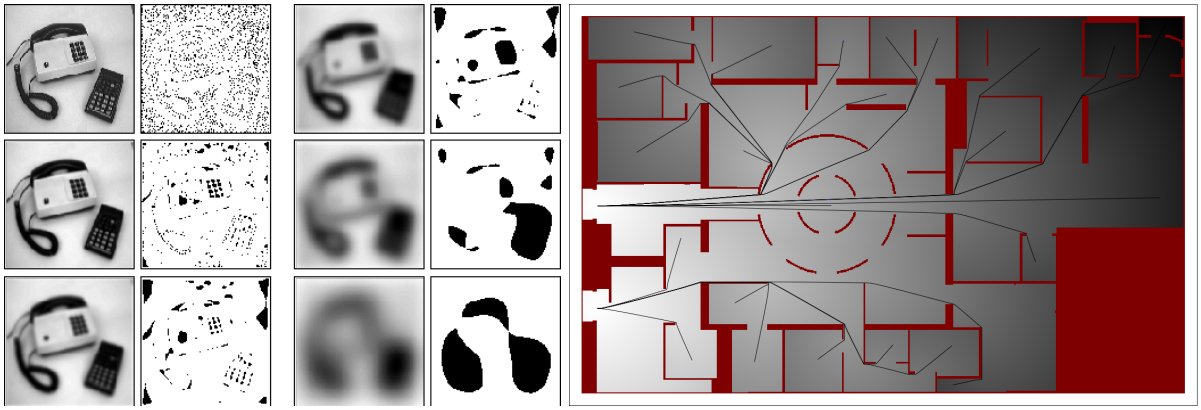

Figure 1: (Gauche) Equation de la chaleur et représentation multi-échelle d'une image. Crédit image: Tony Lindeberg. (Droite) Solution de l'équation eikonale (niveaux de grid) dans un domaine avec obstacles (rouge), et chemins minimaux.

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

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

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

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

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

- *Micro-structure.* Certains milieux physiques sont micro-structurés, ce qui affecte les ondes s'y propageant de manière anisotrope, notamment lorsque leur longueur d'onde excède l'échelle de la micro-structure, par un effet appelé homogénéisation. Ondes sismiques dans un milieu stratifié, ondes lumineuses dans un crystal, ... En traitement de l'image, il est aussi courant de détecter les directions préférentielles de textures oscillantes (comme dane une image d'empreinte digitale) pour les traiter de manière appropriée.

- *Rôle distinct des dimensions du domaine.* Certains domaines mathématiques ne correspondent pas à un milieu physique, mais à un espace d'états dont les dimensions jouent des rôles différents, ce qui crée des structures anisotropes. Par exemple l'espace $\mathbb{R}^2 \times \mathbb{S}^1$ des positions et orientations, qui correspond aux configurations d'un véhicule simple, possède une structure dite sous-riemannienne, permettant en (x_0, x_1, θ) seulement les déplacements engendrés par $(\cos\theta, \sin\theta, 0)$ et $(0, 0, 1)$.

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

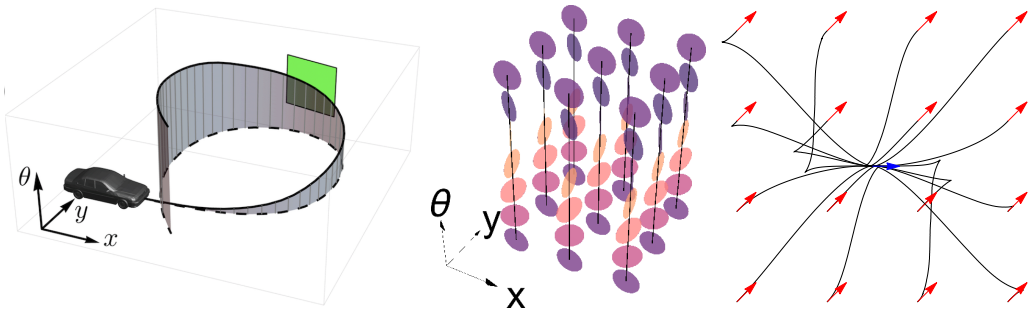

Figure 3: Exemple d'anisotropie liée au rôle distinct des dimensions du domaine. Ici le véhicule de Reeds-Shepp, dont l'espace d'états est $\mathbb{R}^2 \times \mathbb{S}^1$. (Gauche) Une trajectoire admissible. (Centre) Boule unité dans l'espace des vitesses admissibles. (Droite) Projection planaire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

- *Proximité des bords ou de fissures.* La modélisation des écoulements au bord d'un domaine, ou de l'élasticité au voisinage d'une fissure, se fait souvent par l'intermédiaire de modèles réduits traitant de manière spécifique la dimension tangente.
- *Paramétrisation d'un domaine complexe.* La résolution de d'EDPs dans des géométries complexes, surfaces ou volumes, peut se faire par l'intermédiaire de leur paramétrisation par des domaines de référence (rectangles). La paramétrisation induit alors en général une distortion anisotrope de l'EDP sur le domaine de référence, même si elle était isotrope sur la géométrie initiale.
- *Linéarisation d'un problème.* Certaines EDPs linéaires sont obtenues par perturbation d'une EDP non-linéaire au voisinage d'une solution. Selon les cas, l'équation linéarisée peut être anisotrope même si l'équation non-linéaire était isotrope. Par exemple, dans le cas de Monge-Ampère, on a formellement $\det(\nabla^2(u+h)) - \det(\nabla^2 u) = \mathrm{Tr}(D(u)\nabla^2 h) + \mathcal{O}(h^2)$ où $D(u)$ est la comatrice de $\nabla^2 u$.

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

$$\partial_t u = \mathrm{div}(D\nabla u), \qquad\qquad \langle \nabla u, D\nabla u \rangle = 1, \tag{4}$$

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

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 description plus détaillée de la différentiation automatique.

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

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

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

Les différences finies se caractérisent par l'ordre de la dérivée qu'elles approchent (ici 1,1,2), et leur ordre de précision (ici 1,2,2). Une autre approche conceptuellement proche, pour approcher les dérivées d'une fonction définie ponctuellement, consiste à interpoler cette fonction (par éléments finies, 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 discrétisation des équations aux dérivées partielles. Leur structure mathématique simple et (en général) linéaire permet d'étudier mathématiquement les quantités qui en sont dérivées.

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

Différentiation automatique. On appelle différentiation automatique les procédés informatiques permettant de calculer les dérivées de fonctions définies par des programmes, en utilisant les dérivées exactes des fonctions usuelles (exp, $\sqrt{\cdot}$, sin, etc) et en appliquant les règles de composition des différentielles.

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

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

Différentiation 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**[1] **en première approche**, en faveur de la différentiation automatique. En effet, elle rend le programme peu lisible, peu flexible, long à écrire, et introduit de nombreux bugs. Lorsqu'elle nécessaire, ses résultats doivent a minima être contrôlés sur des exemples par une approche alternative[2].

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 $f : \Omega \to F$, où $\Omega \subseteq E$ est ouvert, est différentiable en $x \in \Omega$ s'il existe $L \in \mathcal{L}(E,F)$ telle que*

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

On note $L = \mathrm{d}f_{|x}$, appelée différentielle de f en x.

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

$$\langle \nabla f(x), v \rangle = \mathrm{d}f_{|x}(v)$$

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

Exemple : soit H un Hilbert, et $f : H \to \mathbb{R}$ 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 $f : \Omega \to \mathbb{R}^n$, où $\Omega \subseteq \mathbb{R}^m$, différentiable en $x \in \Omega$. On appelle matrice jacobienne de f en x, notée $\mathrm{D}f_{|x}$, la matrice de $\mathrm{d}f_{|x}$ dans les bases canoniques de $\mathbb{R}^m$ et $\mathbb{R}^n$. Ses composantes sont appelées dérivées partielles de f*

$$\mathrm{D}f_{|x} = \left(\frac{\partial f_i}{\partial x_j}(x) \right)_{\substack{1 \le i \le n \\ 1 \le j \le m}}$$

[1] 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'éxécution 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.

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

En l'absence d'ambiguité (choix de base), on ne se gènera pas pour identifier $\mathrm{d}f_{|x}$ et $\mathrm{D}f_{|x}$.
Exemple: si $f : \mathbb{R}^m \to \mathbb{R}$, alors $\nabla f(x) = (\mathrm{D}f_{|x})^T$

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

$$\mathrm{d}(g \circ f)_{|x} = \mathrm{d}g_{|f(x)} \circ \mathrm{d}f_{|x}.$$

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

2.2 Préliminaire: coût d'un produit matriciel

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

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

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

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

Considérons $A_1, \cdots, A_n$ des matrices de taille $I_0 \times I_1, \cdots, I_{n-1} \times I_n$. 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

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

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

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

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

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

Proof. Soit $\mathcal{A}$ (resp. $\mathcal{B}$) la matrice obtenue en remplaçant les coefficients non-nuls de A (resp. B) par 1. Alors $\mathcal{AB}$ 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 $\mathcal{AB}$ sur la ligne d'indice i, où $1 \leq i \leq I$, vaut

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

Le résultat annoncé s'ensuit. □

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

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

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

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

Interprétation en tant que Jacobienne. Soient $f_1 : \mathbb{R}^{I_1} \to \mathbb{R}^{I_0}$, $\cdots$, $f_n : \mathbb{R}^{I_n} \to \mathbb{R}^{I_{n-1}}$ 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

$$\mathrm{D}(f_1 \circ \cdots \circ f_n)_{|x_n} = \mathrm{D}f_{1|x_1} \cdots \mathrm{D}f_{n|x_n}.$$

Les programmes informatiques décrivent des fonctions complexes comme composées de fonctions élémentaires. Leur différentiation automatique peut donc s'interpréter[3] comme le produit matriciel des matrices Jacobiennes des étapes intermédiaires. Selon la structure de ces facteurs, on préfèrera utiliser un produit associatif à gauche, ou à droite, ou un produit creux; voire une combinaison de ces approches suivant les parties du facteur.

2.3 Trois approches de la différentiation automatique

On peut distinguer trois approches principales de la différentiation automatique: *dense*, *creuse*, et par *rétro-propagation*. Elles correspondent conceptuellement aux trois stratégies détaillées §2.2 pour le calcul d'un produit matriciel: associativité à droite, à gauche, ou produit creux, ce qui permet d'anticiper leurs forces et leurs faiblesses respectives. Cependant leur implémentation numérique s'éloigne de ce cadre, car pour plus de commodité et d'efficacité les matrices Jacobiennes des étapes intermédiaires ne sont en général pas construites explicitement.

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

$$f : \mathbb{R}^m \to \mathbb{R}^n, \tag{9}$$

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

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

Différentiation automatique dense (m petit).

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

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

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

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

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

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

Différentiation automatique creuse (f simple).

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

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

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

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

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

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

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

Par rétro-propagation (n petit).

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

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

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

2.4 Dérivées d'ordre deux et supérieur

Rappels : définitions et propriétés élémentaires.

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

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

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

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

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

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

Exemple: la Hessienne de $f : \mathbb{R}^n \to \mathbb{R}$ definie 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.

- *Différentiation 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 equations de Hamilton, qui jouent un rôle fondamental en physique et en mathématiques

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

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

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

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

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

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

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

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

3 Discrétisation de la diffusion anisotrope et non-linéaire

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

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

sur un domaine $\Omega \subseteq \mathbb{R}^d$, muni d'un champ de matrices symétriques définies positives $D : \Omega \to S_d^{++}$. 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[4] en traitement d'image.

3.1 Flot gradient dans un espace abstrait

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

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

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

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

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

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

De manière équivalente, $y = x - \varepsilon \nabla \mathcal{E}(y)$, ce qui caractérise bien un pas de gradient implicite.

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 $(x^{\varepsilon}_n)_{n \geq 0}$ et une application constante par morceaux $\boldsymbol{x}^{\varepsilon} : [0, \infty[\to X$ par*

$$x^{\varepsilon}_0 = x_0 \qquad x^{\varepsilon}_{n+1} = \mathrm{prox}^{\varepsilon}_{\mathcal{E}}(x^{\varepsilon}_n),\ \forall n \geq 0, \qquad \boldsymbol{x}^{\varepsilon}(t) = x^{\varepsilon}_n,\ \forall t \in [n\varepsilon, (n+1)\varepsilon[.$$

Supposons qu'il existe une suite $\varepsilon_k \to 0$, telle que les fonctions $\boldsymbol{x}^{\varepsilon_k} : [0, \infty[\to X$ convergent localement uniformément vers une limite $\boldsymbol{x} : [0, \infty[\to X$. Alors on dit que $\boldsymbol{x}$ 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 and 3.2 ne permettent d'établir ni l'existence ni l'unicité des objets considérés. L'existence d'un flot gradient et d'un minimiseur pour (11) s'établit sous des hypothèses de compacité. On dit qu'une partie d'un espace métrique est relativement compacte si son adhérence est compacte.

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

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

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

$$d_X(x^{\varepsilon}_k, x^{\varepsilon}_{k+1})^2 \leq \varepsilon\big(\mathcal{E}(x^{\varepsilon}_k) - \mathcal{E}(x^{\varepsilon}_{k+1})\big),$$

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

$$d_X(x^{\varepsilon}_n, x^{\varepsilon}_m)^2 \leq \Big(\sum_{m \leq k < n} d_X(x^{\varepsilon}_k, x^{\varepsilon}_{k+1}) \Big)^2 \leq (n-m) \sum_{m \leq k < n} d_X(x^{\varepsilon}_k, x^{\varepsilon}_{k+1})^2 \leq (n-m)\varepsilon\big(\mathcal{E}(x^{\varepsilon}_m) - \mathcal{E}(x^{\varepsilon}_n)\big).$$

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

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

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

L'unicité et la régularité du flot gradient et de l'opérateur proximal s'établissent sous des hypothèses de convexité[5].

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 $\boldsymbol{x}, \boldsymbol{y}$ issus de points x_0, y_0 satisfont $\|\boldsymbol{x}(t) - \boldsymbol{y}(t)\| \leq \|x_0 - y_0\|$ pour tout $t \geq 0$.*

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

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

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

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

Par convexité, le terme de gauche est positif pour tout $t \in [0, 1]$. Un développement limité du terme de droite lorsque $t \to 0$ 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 $\mathrm{prox}_{\mathcal{E}}^{\varepsilon}$ 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. □

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 Wasserstein* (transport optimal) [JKO98]. Dans cette sous-section, on justifie l'interprétation formelle de l'équation de la chaleur par première approche (la plus ancienne). Ceci permet de justifier son existence, son unicité, et sa régularité par rapport aux paramètres, par application directe des Définitions 3.1 and 3.2 and Propositions 3.3 and 3.4. Soit $\Omega \subseteq \mathbb{R}^2$ un domaine borné de bord régulier, on note dans la suite

$$\mathbb{L}^2 := \mathbb{L}^2(\Omega), \qquad \langle u, v\rangle_{\mathbb{L}^2} = \int_\Omega u(x)v(x)\,\mathrm{d}x. \tag{12}$$

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

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

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

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

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

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

On se propose de vérifier les hypothèses de Propositions 3.3 and 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 $c_{\min}\|v\|^2 \leq \|v\|^2_{D(x)} \leq c_{\max}\|v\|^2$ pour tout $v \in \mathbb{R}^d$, $x \in \overline{\Omega}$. On en déduit pour tout $u \in \mathbb{H}^1(\Omega)$

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

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

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

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

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

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

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

Gradient de l'énergie de Dirichlet. Pour u et h suffisament régulières, nous calculons

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

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

$$\mathrm{d}\mathcal{E}_{|u}(h) = \mathcal{Q}(u,h) = \big\langle h, -\operatorname{div}(D\nabla u)\big\rangle_{\mathbb{L}^2(\Omega)} + \int_{\partial\Omega} h\langle \boldsymbol{n}, D\nabla u\rangle \tag{17}$$

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

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

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

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

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

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

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

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

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

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

Compte tenu de la Définition 3.2 du flot, il suffit de prouver que $v := \mathrm{prox}^{\varepsilon}_{\mathcal{E}}(u)$ 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 discrétisation $h > 0$, et on introduit le domaine discrétisé et le produit scalaire

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

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

$$Q_h(u,u) := h^d \sum_{x\in\Omega_h} Q^x_h(u,u), \qquad \text{telle que} \quad Q^x_h(u,u) := \|\nabla u(x)\|^2 + \mathcal{O}(h^2). \tag{18}$$

La forme locale de Q^x_h en $x \in \Omega_h$ dépend de celle de la matrice $D(x)$. On se contente de donner $Q^x_h(u,u)$, pour la lisibilité, car la reconstruction de la forme polaire se fait par polarisation (14).

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

$$Q_h^x(u,u) = \frac{d(x)}{2h^2} \sum_{1\le i\le d} \Big[\big(u(x+he_i) - u(x)\big)^2 + \big(u(x-he_i) - u(x)\big)^2 \Big], \tag{19}$$

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

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

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

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

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

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

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

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

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

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

où l'on a noté Id la matrice identité, $I_h := h^d\,\mathrm{Id}$ la matrice du produit scalaire $\langle\cdot,\cdot\rangle_h$, et Q_h celle de la forme quadratique du même nom. Le schéma est stable[6], en norme $\mathbb{L}^2(\Omega_h)$, pourvu que $\mathrm{Id} - \delta_t I_h^{-1} Q_h$ ait toute ses valeurs propres dans $[-1,1]$. De manière équivalente

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

[6]Sous la même condition, l'energie définie par Q_h (ou toute puissance de celle-ci) décroit. Le schéma implicite est pour sa part inconditionellement stable.

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

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

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 $u(x)^2$,

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

où

$$C := \max_{x\in\Omega_h} \sum_{1\leq|i|\leq I} (\rho_i(x) + \rho_i(x+he_i)), \qquad \Big(\text{CFL} : C\delta t \leq 2\Big). \tag{22}$$

Pour majorer la constante (22), on note que D soit Lipschitz, et que

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

Par ailleurs 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 $C \lesssim 4\max\{\mathrm{Tr}(D(x)); x \in \overline{\Omega}\}$.

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

3.4 Décomposition matricielle de Selling

La décomposition de Selling fait partie du champ de recherches de la *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 $i,j \in \{0,\cdots,d\}$.

Définition 3.5 (Superbase)**.** *On appelle superbase de $\mathbb{Z}^d$ une famille $(v_0,\cdots,v_d) \in (\mathbb{Z}^d)^d$ telle que $v_0+\cdots+v_d = 0$ et $|\det(v_1,\cdots,v_d)| = 1$. 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 Figure 4 les superbases obtuses associées à des matrices symétriques, paramétrées par

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

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

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

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

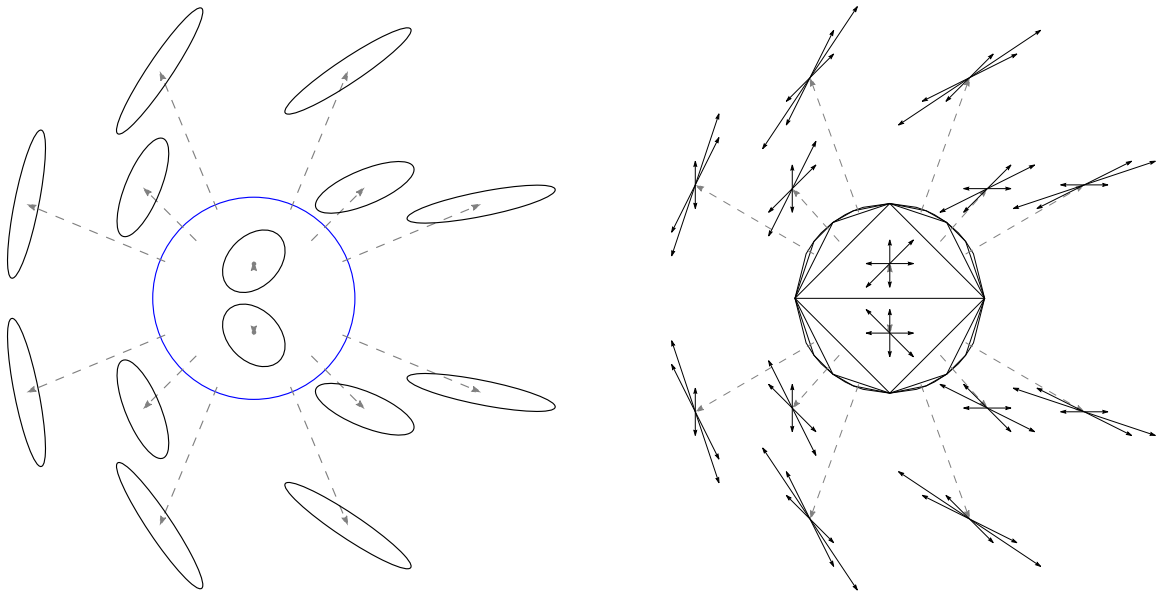

Figure 4: (Gauche) Ellipse définie par $\{v \in \mathbb{R}^2; \langle v, D(x,y)v\rangle = 1\}$, où $D(x,y)$ est définie par la paramétrisation de Pauli (23). (Droite) Superbase $D(x,y)$-obtuse, et son opposée.

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 $(v_1, \cdots, v_d)$ est une base, on a $D = D'$. □

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

Proposition 3.8 (Algorithme de Selling)**.** *Etant donnée une superbase $(v_0, \cdots, v_d)$ 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 $\langle v_i, Dv_j\rangle \geq 0$ alors (resp. en notant $\{i,j,k,l\} = \{0,1,2,3\}$)*

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

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

Proof. Un calcul direct montre que la quantité $\mathcal{E}(v) := \sum_{0\leq i\leq d} \|v_i\|_D^2$ décroit 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. □

Programmation linéaire. Etant donnée $D \in S_d^{++}$, Voronoi définit sa première réduction par le programme linéaire

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

A toute superbase $b = (v_0, \cdots, v_d)$, associons $M_b := \frac{1}{2}\sum_{0\leq i\leq d} v_i v_i^T$. On peut montrer que la matrice M_b satisfait (3.4, droite), et que si b est D-obtuse, alors elle résout (3.4). De plus,

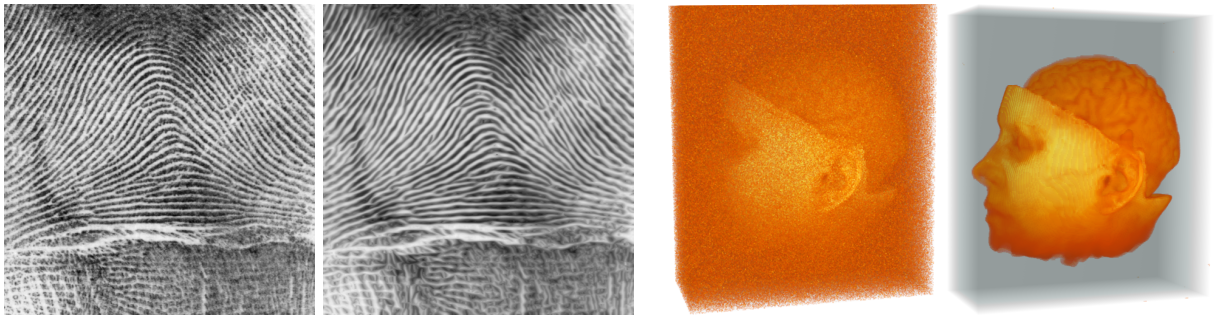

Figure 5: Avant/après l'effet de $\partial_t u = \mathrm{div}(D_u \nabla u)$, avec coefficients de diffusion non-linéaires et anisotropes $D[\nabla u]$ (Coherence enhancing diffusion [Wei98]).

en dimension $d \in \{2, 3\}$, l'algorithme de Selling s'apparente à la méthode du simplexe pour la résolution du programme linéaire (3.4).

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

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

Ainsi, la réduction de Voronoi permet d'étendre la décomposition de Selling en dimension $d > 3$, au prix de complications certaines a 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:

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

où $D[\nabla u](t, x)$ dépend de manière non-linéaire de $\nabla u(t, \cdot)$, et est une matrice symétrique semi-définie positive. Les traitement mathématique, numérique, et les applications de la non-linéarité $D[\nabla u]$ dépendent de deux propriétés principales.

- Dépendance *locale* si $D[\nabla u](x)$ seulement de $\nabla u(x)$. Dépendance *non-locale* si $D[\nabla u](x)$ dépend de $\nabla u(\cdot)$ au voisinage de x, en général via une convolution. Paradoxalement, l'analyse (existence, unicité, stabilité, discrétisation, ... mais aussi contraintes induites sur la forme de $D[\nabla u]$) dans le cas non-local est en général plus simple grâce à l'effet régularisant de la convolution.

- Diffusion *isotrope*[7] si les matrices $D[\nabla u]$ sont en tout point proportionnelles à l'identité, et *anisotrope* sinon. L'anisotropie donne des degrés de liberté appréciables dans les applications. Les formes des équations et schémas numériques isotropes (19) et anisotrope (21) étant assez similaires, l'analyse n'est pas différente.

Deux classes d'équations ont été particulièrement étudiées.

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

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:

$$D[u](x) = g(|\nabla u(x)|)\,\mathrm{Id}\,. \tag{26}$$

De manière à préserver les contours et discontinuités présentes dans l'image, la fonction g promeut une diffusion faible dans les zones de fort gradient. Les modèles suivants ont été particulièrement étudiés, et tirent leur nom de la fonctionnelle de Variation Totale (TV) ou des auteurs Perona et Malik (PM) [PM90]:

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

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

$$\mathcal{E}_f(u) := \int_\Omega f(|\nabla u(x)|)\,\mathrm{d}x \qquad\qquad f'(s) = s\,g(s)$$

Si f est convexe et croissante, de manière équivalente si $s \mapsto sg(s)$ est croissante et positive, alors l'énergie $\mathcal{E}_f$ est convexe. C'est le cas de la variation totale $f_{\mathrm{TV}}(s) = s$, mais pas celui de Perona Malik $f_{\mathrm{PM}}(s) = (2\lambda)^{-1}\ln(1+\lambda s^2)$ pour lequel l'unicité est perdue. Ce phénomène s'illustre visuellement par des instabilités menant à des artefacts *en marche d'escalier*.

Dépendance non-locale et anisotrope. Dans ces travaux, le coefficient de diffusion dépend d'une régularisation par convolution $G_\sigma \star u$ de la solution, où σ désigne l'échelle du bruit dans l'image, et G_σ est typiquement un noyau Gaussien. L'effet régularisant de la convolution permet d'établir l'existence, l'unicité, et la stabilité par rapport aux conditions initiales, à l'aide de techniques relativement standard en analyse parabolique, présentées[8] 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

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

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

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

La fonction de transfer $f : \mathbb{R}^d \to \mathbb{R}^d$ 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].

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

References

[AGS08] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science and Business Media, 2008.

[CLMC92] Francine Catté, Pierre-Louis Lions, Jean-Michel Morel, and Tomeu Coll. Image selective smoothing and edge detection by nonlinear diffusion. *SIAM Journal on Numerical Analysis*, 29(1):182–193, 1992.

[CS92] J H Conway and N J A Sloane. Low-Dimensional Lattices. VI. Voronoi Reduction of Three-Dimensional Lattices. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 436(1896):55–68, January 1992.

[CS13] John Horton Conway and Neil James Alexander Sloane. *Sphere packings, lattices and groups*, volume 290. Springer Science and Business Media, 2013.

[FM14] Jérôme Fehrenbach and Jean-Marie Mirebeau. Sparse non-negative stencils for anisotropic diffusion. *Journal of Mathematical Imaging and Vision*, 49(1):123–147, 2014.

[JKO98] Richard Jordan, David Kinderlehrer, and Felix Otto. The Variational Formulation of the Fokker–Planck Equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, January 1998.

[JS13] Boško S Jovanović and Endre Süli. *Analysis of finite difference schemes: for linear partial differential equations with generalized solutions*, volume 46. Springer Science and Business Media, 2013.

[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.

[Sch09] Achill Schürmann. Computational geometry of positive definite quadratic forms. *University Lecture Series*, 49, 2009.

[Sel74] Eduard Selling. Ueber die binären und ternären quadratischen Formen. *Journal fur die Reine und Angewandte Mathematik*, 77:143–229, 1874.

[Wei98] Joachim Weickert. *Anisotropic diffusion in image processing*, volume 1. Teubner Stuttgart, 1998.