Fast Marching method for the computation of first-arrival travel time of seismic waves in anisotropic media

François Desquilbet

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : MSTII - Mathématiques, Sciences et technologies de l'information, Informatique
Spécialité : Mathématiques Appliquées
Unité de recherche : Laboratoire Jean Kuntzmann

Méthode de Fast Marching pour le calcul du premier temps d'arrivée d'ondes sismiques en milieux anisotropes

Fast Marching method for the computation of first-arrival travel time of seismic waves in anisotropic media

Présentée par :

François DESQUILBET

Direction de thèse :

Ludovic METIVIER Directeur de recherche, CNRS DELEGATION ALPES	Directeur de thèse
Jean-Marie MIREBEAU Directeur de recherche, CNRS DELEGATION ILE-DE-FRANCE SUD	Co-directeur de thèse

Rapporteurs :

JEAN-DAVID BENAMOU
Directeur de recherche, INRIA CENTRE DE PARIS
SERGEY FOMEL
Professeur, The University of Texas at Austin

Thèse soutenue publiquement le **14 octobre 2022**, devant le jury composé de :

EMMANUEL MAITRE Professeur des Universités, GRENOBLE INP	Président
JEAN VIRIEUX Professeur des Universités émérite, UNIVERSITE GRENOBLE ALPES	Examinateur
JEAN-DAVID BENAMOU Directeur de recherche, INRIA CENTRE DE PARIS	Rapporteur
SERGEY FOMEL Professeur, The University of Texas at Austin	Rapporteur

Invités :

JEAN-MARIE MIREBEAU
Directeur de recherche, CNRS DELEGATION ILE-DE-FRANCE SUD
LUDOVIC METIVIER
Directeur de recherche, CNRS DELEGATION ALPES

Abstract

The first arrival traveltime for the propagation of a wave, in the high frequency approximation, is described by the eikonal equation, or possibly a variant with coefficients depending on the medium properties. We present numerical schemes for the computation of the solution to such eikonal equations. These numerical schemes are based on the Fast Marching method (FMM), generalized to complex and non-Riemannian anisotropy settings in 3D media. The FMM is a single pass method, in which the propagation front is discretized and followed throughout the medium, leading to fast computation time. We also explore an opposite paradigm for high performance computation, based on a massively parallel GPU solver.

In particular, we consider the case of seismic pressure waves propagating inside a geophysical medium, with a propagation speed defined by an anisotropic Hooke tensor. In this context of geophysics, we propose two numerical schemes, generalizing ideas from previous schemes and referred to as "semi-Lagrangian" scheme and "Eulerian" scheme.

The semi-Lagrangian scheme can handle anisotropy of fully general shape, but with a limitation based on the strength of the anisotropy, defined as the ratio between the fastest and slowest speed achievable depending on the orientation. A review of the known and tabulated anisotropy properties of geological materials suggests that the method is applicable in most scenarios of interest. We also consider how the limitation of the semi-Lagrangian scheme can be removed in 2D by designing geometric 2D stencils adapted to the anisotropy, and we study the worst case and average case for the cardinality of the stencils designed by this algorithm.

On the other hand, the Eulerian scheme is limited to anisotropy coming from a Tilted Transversely Isotropic (TTI) medium and cannot handle more complex elastic parameters, but it does not have any limitation on the strength of the anisotropy. It works by expressing the TTI eikonal equation as a maximum or minimum of a family of Riemannian eikonal equations, for which efficient discretizations are known. We also consider an implementation of the Eulerian scheme to massively parallel architectures, leading to a computation fifty times faster than the sequential FMM implementation, using a single GPU node. Besides, leveraging similar numerical methods in a different context, we study an application of the Eulerian scheme to an inverse problem involving motion planning for the optimization of the configuration of a radar network.

Abstract (en français)

Le temps de première arrivée pour la propagation d'une onde, dans l'approximation haute fréquence, est décrit par l'équation eikonale, ou éventuellement une variante dont les coefficients dépendent des propriétés du milieu. Nous présentons des schémas numériques pour le calcul de la solution de ces équations eikonales. Ces schémas numériques reposent sur la méthode du Fast Marching (FMM), généralisée à des contextes complexes mettant en jeu de l'anisotropie non riemannienne dans des milieux 3D. La FMM est une méthode en une seule passe, dans laquelle le front de propagation est discrétisé et suivi dans tout le milieu, ce qui permet un temps de calcul rapide. Nous explorons également un paradigme opposé pour un calcul de haute performance, qui repose sur un solveur GPU massivement parallèle.

En particulier, nous considérons le cas d'ondes de pression sismiques se propageant dans un milieu géophysique, avec une vitesse de propagation définie par un tenseur de Hooke anisotrope. Dans ce contexte de géophysique, nous proposons deux schémas numériques, généralisant les idées des schémas précédents et appelés schéma "semi-lagrangien" et schéma "eulérien".

Le schéma semi-lagrangien peut traiter une anisotropie de forme complètement générale, mais avec une limitation liée à la force de l'anisotropie, définie comme le rapport entre la vitesse la plus rapide et la plus lente réalisable en fonction de l'orientation. Un examen des propriétés d'anisotropie connues et répertoriées des matériaux géologiques suggère que la méthode est applicable dans la plupart des scénarios d'intérêt. Nous examinons également comment la limitation du schéma semi-lagrangien peut être supprimée en 2D en concevant des stencils géométriques 2D adaptés à l'anisotropie, et nous étudions le pire cas et le cas moyen pour la cardinalité des stencils conçus par cet algorithme.

D'autre part, le schéma eulérien est limité à l'anisotropie provenant d'un milieu TTI (Tilted Transversely Isotropic) et ne peut pas gérer des paramètres élastiques plus complexes, mais il n'a aucune limitation sur la force de l'anisotropie. Il fonctionne en exprimant l'équation eikonale TTI comme un maximum ou un minimum d'une famille d'équations eikonales riemanniennes, pour lesquelles des discrétisations efficaces sont connues. Nous considérons également une mise en œuvre du schéma eulérien sur des architectures massivement parallèles, conduisant à un calcul cinquante fois plus rapide que la mise en oeuvre séquentielle de la FMM, en utilisant un seul noeud de GPU. Enfin, en utilisant des méthodes numériques similaires dans un contexte différent, nous étudions une application du schéma eulérien à un problème inverse impliquant la planification du mouvement pour l'optimisation de la configuration d'un réseau radar.

Contents

Résumé (en français)

Equation eikonale et temps de première arrivée

L'équation eikonale est une équation aux dérivées partielles non linéaire, qui a été considérée pour la première fois en optique géométrique [Bru95]. Dans ce contexte, elle constitue une généralisation de la loi de Snell pour la propagation d'un rayon de lumière dans un milieu continu avec indice de réfraction variable. L'équation eikonale peut être considérée dans un contexte plus général, pour décrire n'importe quelle sorte d'onde qui se propage dans un domaine muni d'une métrique (i.e. d'une notion de vitesse). Cette équation caractérise, pour une propagation d'onde, le temps de première arrivée de cette onde à chaque endroit du domaine. Les trajets les plus courts au sein du domaine, depuis la source vers n'importe quelle position, peuvent aussi s'en déduire en se déplaçant perpendiculairement aux lignes de niveaux du temps de première arrivée.

En géophysique, une équation eikonale peut être obtenue à partir de l'approximation haute fréquence de l'équation des ondes élastiques, et la métrique correspondante est définie à partir des propriétés élastiques du milieu géologique. La solution de cette équation eikonale correspond au temps de première arrivée de l'onde sismique. Comparée à l'équation eikonale, l'équation des ondes élastiques fournit une plus grande quantité d'informations sur le phénomène de propagation d'onde : en effet, la solution de l'équation des ondes élastiques décrit l'amplitude du champ vectoriel de déplacement à chaque instant et à chaque position, et il est possible d'en déduire les temps de première arrivée comme étant, pour chaque position, le premier instant à partir duquel l'amplitude est non nulle (voir Figure 1 pour une illustration). Cependant, le calcul de la solution de l'équation des ondes dans des milieux complexes en trois dimensions a un coût numérique élevé : l'échelle de la grille de discrétisation doit être significativement plus petite que la longueur d'ondes des oscillations pour éviter la dispersion numérique, et le pas de temps est majoré par la condition de stabilité de Courant-Friedrichs-Levy. En revanche, l'équation eikonale est une équation aux dérivées partielles statique, dont la solution est non-oscillante. Ainsi, le calcul des solutions de l'équation eikonale peut être effectué pour un coût bien inférieur.

Dans le cadre de cette thèse, je m'intéresse à des équations eikonales pour des métriques anisotropes, i.e. pour lesquelles la vitesse de propagation de l'onde dépend non seulement de la position, mais aussi de l'orientation du front d'onde. L'implémentation de solveurs numériques pour les équation eikonales anisotropes est un problème mathématique délicat.

Etude de milieux anisotropes

Il existe différents modèles pour représenter le sous-sol comme un milieu élastique, avec des degrés de complexité variables. Du modèle le plus simple au modèle le plus complexe, on peut citer :

- Milieu élastique isotrope, défini par exemple à partir des deux paramètres de Lamé,
- Milieu elliptique (aussi appelé riemannien),

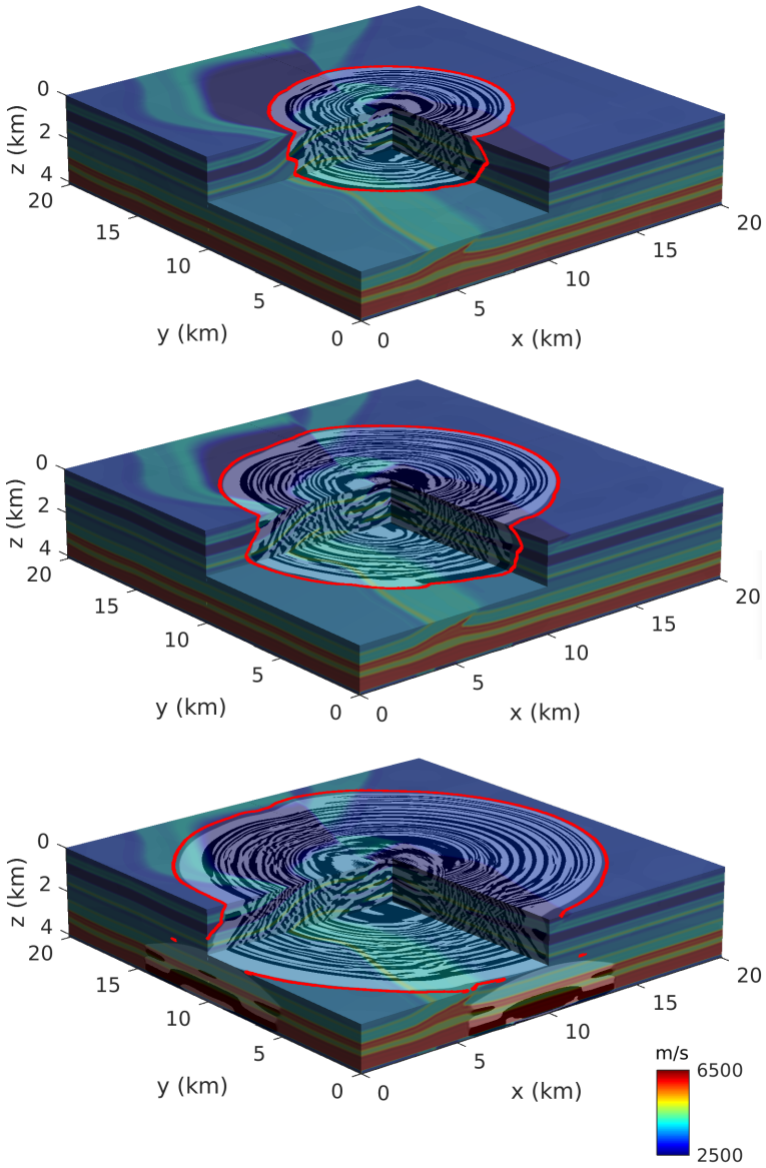

Figure 1: Superposition de la solution de l'équation des ondes élastiques (en noir et blanc) et de la solution de l'équation eikonale (en rouge), où l'arrière-plan représente la vélocité de l'onde. Les différents aperçus ont été obtenus aux temps de propagation $t = 1.5s$ (haut), $t = 2s$ (milieu) and $t = 2.5s$ (bas), pour une source localisée au centre du domaine.

- Milieu transversalement isotrope, qui possède un axe de symétrie par rotation qui est le plus souvent vertical (VTI), ou bien incliné (TTI),
- Milieu orthorhombique, qui possède trois plans de symétries mutuellement perpendiculaires, verticaux ou inclinés,
- Milieu triclinique, qui est la forme d'anisotropie la plus complexe pour les modèles étudiés dans le cadre des milieux élastiques en géophysique.

Pour des ondes sismiques qui se propagent à l'intérieur de la Terre, l'anisotropie doit être prise en compte pour une meilleure modélisation [BC91]. On identifie deux origines principales de cette anisotropie :

- *Anisotropie intrinsèque :* L'anisotropie peut venir naturellement de la forme des minéraux, en particulier de la structure des cristaux à l'échelle atomique. Par exemple, le cristal d'olivine peut être trouvé dans la partie supérieure du manteau sous les océans et conduit à une direction préférée jusqu'à 25% plus rapide que les autres directions, avec un profil de vitesse correspondant à une anisotropie orthorhombique [Hes64].

- *Anisotropie extrinsèque :* des structures fines de matériaux isotropes, telles que les couches sédimentaires, peuvent également affecter le front d'onde de la même manière que le ferait un milieu anisotrope. Pour les couches sédimentaires, ce processus d'homogénéisation conduit généralement à un milieu présentant une symétrie de rotation le long de l'axe de la couche, ce qui mène à un milieu transverse isotrope. Les couches sont généralement horizontales, ce qui conduit à un milieu VTI. Certains déplacements de l'axe de symétrie peuvent également se produire avec les mouvements tectoniques, conduisant à un milieu TTI. Dans le cas de fractures ou de sitations tectoniques complexes, l'anisotropie équivalente dérivée du processus d'homogénéisation peut être encore plus complexe, conduisant à des milieux orthorhombiques ou même tricliniques [CC18].

Pour décrire l'anisotropie d'une métrique, on introduit dans cette thèse deux concepts :

- *Force de l'anisotropie :* à une position donnée, la force de l'anisotropie est le rapport entre la vitesse la plus élevée et la vitesse la plus faible en fonction de l'orientation.

- *Complexité de l'anisotropie:* il s'agit du nombre de paramètres nécessaires pour caractériser la métrique : 1 paramètre est nécessaire pour les métriques isotropes, 3 paramètres pour les métriques elliptiques (6 paramètres en cas d'inclinaison pour définir la rotation), 5 paramètres pour les métriques transverses isotropes (7 si inclinaison), 9 paramètres pour les métriques orthorhombiques (12 si inclinaison), et 21 pour les milieux tricliniques qui constituent la forme d'anisotropie la plus générale pour les milieux élastiques.

Ces deux concepts posent des défis distincts pour la réalisation de solveurs eikonaux : la plupart des solveurs numériques sont limités à une complexité particulière (jusqu'aux milieux transverses isotropes pour la plupart des schémas numériques existants), et peuvent également échouer si la force de l'anisotropie est trop prononcée. En outre, ces deux concepts de force et de complexité sont indépendants : une anisotropie de forme elliptique peut être plus forte qu'une anisotropie de forme orthorhombique, bien que moins complexe.

Etat de l'art

Des algorithmes efficaces pour résoudre l'équation eikonale ont été développés à partir du concept de "level sets" [Set96]. Les méthodes numériques qui en résultent peuvent être divisées en deux classes : *méthodes itératives* et *méthodes en une passe*, qui généralisent respectivement les algorithmes de Bellman-Ford et de Dijkstra pour le calcul du temps de première arrivée dans les graphes.

La méthode itérative la plus connue est sans doute la méthode de Fast Sweeping (FSM). Initialement introduite dans un cadre isotrope [Zha05], le FSM a été étendu à l'anisotropie elliptique 2D [TCOZ03]. Dans le contexte de la géophysique, la FSM a été étendue aux métriques 2D transverse isotrope inclinée [LCZ14], 3D transverse isotrope inclinée [PWZ17] avec un schéma de type Lax-Friedrich du troisième ordre, 3D orthorhombique inclinée [WYF15] en la traitant comme un problème itératif sur de l'anisotropie elliptique, et plus récemment [LBLM] pour la métrique 3D orthorhombique inclinée avec une précision d'ordre élevé par une méthode Galerkin discontinue. Récemment, des méthodes itératives pouvant tirer parti d'une architecture de calcul massivement parallèle, notamment les GPU, ont été proposées dans le cadre isotrope [JW08], et pour l'anisotropie elliptique [GHZ18].

D'autre part, la méthode en une passe la plus connue est sans doute la méthode de Fast Marching (FMM) [Tsi95, Set96], mais l'extension de la FMM aux géométries anisotropes s'est avérée plus difficile. Les premières études [KS98, SV01, AM12] impliquent des schémas numériques avec de larges stencils, ce qui entraîne une augmentation des temps de calcul et une réduction de la précision, et annule donc les avantages de la FMM. Plus récemment, dans [Wah20], un algorithme utilisant la FMM a été développé pour l'anisotropie 3D transverse isotrope inclinée : son principe de base repose sur l'interprétation de cette anisotropie comme une correction d'une anisotropie elliptique, et un algorithm de point fixe est proposé pour implémenter cete correction. Bien que les auteurs illustrent numériquement que l'algorithme peut converger lorsque l'anisotropie considérée est proche d'une anisotropie elliptique verticale, il n'y a pas de preuve formelle de la convergence de l'itération du point fixe qu'ils mettent en œuvre.

Ces dernières années, des extensions de la FMM à l'anisotropie elliptique 2D ont été proposées [Mir14b]. D'autres améliorations ont été apportées à l'anisotropie elliptique 3D [Mir14a, Mir19], et à divers types d'anisotropie anelliptique dégénérée liée à la pénalisation de la courbure [Mir18]. Grâce aux techniques issues de la géométrie des réseaux, la taille du stencil de discrétisation est sous contrôle, évitant ainsi toute perte de temps de calcul et de précision, même dans le cas d'une très forte anisotropie (avec une vitesse de propagation potentiellement dix fois plus rapide dans la direction rapide par rapport aux directions lentes).

Contributions

Dans cette thèse, nous généralisons les outils utilisés dans ces schémas numériques pour développer des solveurs pour l'équation eikonale basés sur la FMM, qui peuvent être appliqués dans le cadre de l'anisotropie rencontrée en géophysique. La prise en compte de l'anisotropie est nécessaire pour produire des modèles réalistes de l'intérieur de la Terre. La force et la complexité de l'anisotropie apportent cependant des difficultés techniques pour la conceptions de schémas numériques qui vérifient les propriétés de monotonie et de causalité, qui sont nécessaires pour l'utilisation de la FMM.

Le premier schéma numérique, présenté en Section 6 [DCC+21], repose sur la méthode semi-lagrangienne. Il permet de prendre en compte une anisotropie qui provient d'un tenseur de Hooke 3D triclinique, ce qui constitue la forme d'anisotropie la plus générale pour des milieux élastiques. Cependant, il y a une contrainte liée à la force de l'anisotropie, avec une limitation qui dépend du stencil utilisé dans le schéma de discrétisation numérique aux différences finies : les stencils doivent vérifier une propriété dite d'angle aigu par rapport à la métrique pour que le schéma numérique correspondant soit causal, et donc puisse être résolu en une passe par la FMM. Des exemples de stencils utilisés sont présentés en Figure 2. Dans le cadre d'une anisotropie trop forte pour être traitée par ces stencils, l'utilisation de stencils encore plus larges pourrait être considérée, mais augmenterait notablement les temps de calcul de la FMM. Dans ce cas, il est préférable de se reporter sur une méthode itérative telle que la FSM, ce qui perd les garanties liées à la FMM mais permet de calculer la solution à partir d'un stencil plus simple. Cependant, nous avons vérifié que la plupart des matériaux usuellement rencontrés en géophysique peuvent être traités par les stencils présentés en Figure 2.

Pour l'utilisation de la méthode semi-lagrangienne dans le cadre de la géophysique, une difficulté technique est liée au calcul de la métrique locale. En effet, celle-ci est naturellement écrite comme la maximisation d'une forme linéaire soumise à une contrainte non convexe. Pour pouvoir calculer efficacement sa valeur, on reformule la contrainte sous une forme fortement convexe, plus simple à manipuler.

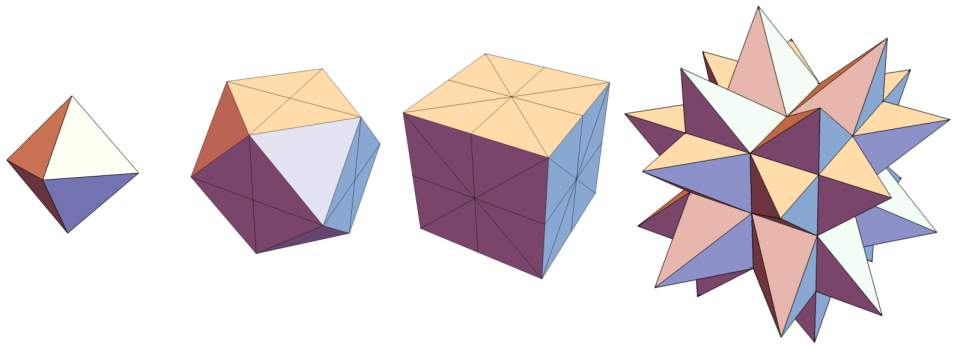

Figure 2: 3D stencils de complexité croissante, utilisés dans le schéma aux différences finies pour l'équation eikonale utilisant la méthode semi-lagrangienne.

Dans le cadre de la dimension 2, on montre comment la limitation du schéma précédent, liée à la force de l'anisotropie, peut être surmontée. On présente en Section 7 [DMM22] un algorithme pour calculer efficacement des stencils 2D strictement aigus par rapport à une métrique. Grâce à l'utilisation de stencils strictement aigus, il est possible de garantir que le schéma numérique correspondant est monotone et causal, même pour des petites perturbations du schéma potentiellement causées par la factorisation à la source ou la recherche d'un ordre élevé de convergence. Dans ce travail, on calcule la cardinalité des stencils construits par l'algorithme en fonction de l'anisotropie de la métrique, en moyenne et dans le pire des cas selon les rotations de la métrique par rapport à la grille de discrétisation.

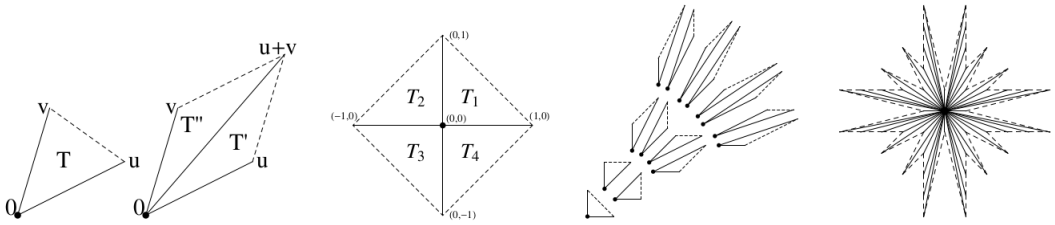

Figure 3: Raffinement d'un triangle (gauche). Stencil 2D initial (centre gauche). Premiers niveaux de raffinement récursifs de T_1 (centre droit). Stencil défini par un raffinement de profondeur 5 (droite).

En Section 8 [DMM22], on présente un autre schéma numérique, qui repose sur une méthode eulérienne, pour résoudre l'équation eikonale dans le cadre de la géophysique. Cette fois, l'anisotropie considérée ne peut pas être plus complexe que celle des milieux transverses isotrope inclinés. Cependant, il n'y a aucune limitation sur la force de l'anisotropie comme dans la méthode semi-lagrangienne. De plus, grâce à la simplicité de la structure eulérienne, le schéma peut aussi être implémenté avec une accélération GPU. Ce schéma numérique repose sur des propriétés géométriques des surfaces de lenteur des milieux transverses isotropes inclinés, qui sont des représentations locales de l'inverse de la vitesse en fonction de l'orientation. On montre que la surface de lenteur d'une métrique TTI peut toujours être représentée localement comme une enveloppe d'ellipses, voir Figure 4. Or, une surface de lenteur elliptique correspond à une métrique riemannienne, pour laquelle l'équation eikonale correspondante peut être résolue efficacement grâce aux outils de la méthode eulérienne. Le schéma numérique pour résoudre l'équation eikonale dans le cas transverse isotrope incliné peut alors s'écrire localement comme un problème d'optimisation sur les métriques riemanniennes qui correspondent aux ellipses de l'enveloppe.

L'implémentation GPU est présentée en Section 9 [MGB+21]. La FMM n'est pas adéquate pour le calcul GPU à cause de sa nature séquentielle, et on utilise plutôt un schéma itératif sur des blocs de points définis sur la grille numérique discrète. Les différents blocs sont traités dans un ordre qui suit l'évolution du front, de manière semblable à la FMM, mais sans avoir la garantie de calcul en une seule passe. L'accélération GPU

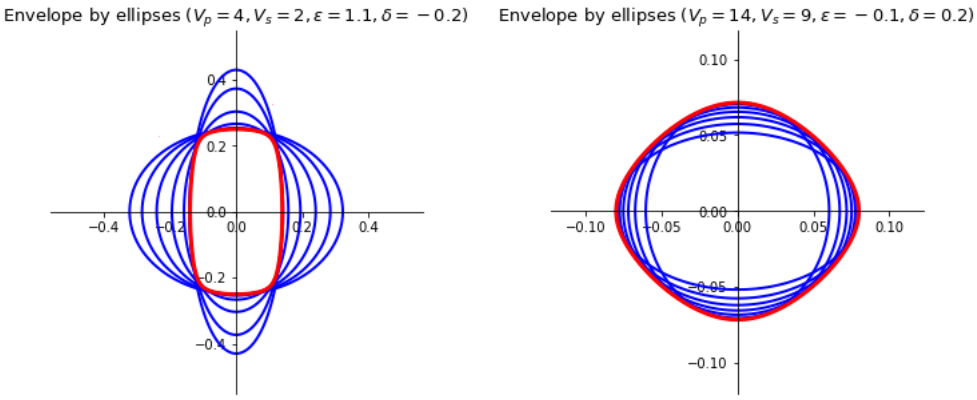

Figure 4: Surfaces de lenteur (en rouge). Le schéma numérique utilise une approximation de ces surfaces par une intersection d'ellipses (gauche) ou une union d'ellipses (droite), représentée en bleues.

permet des temps de calcul 50 fois plus rapides que pour une implémentation séquentielle, avec l'utilisation d'un noeud GPU.

Enfin, on présente une application dans un autre contexte liée à la planification de mouvement, avec un algorithme qui calcule la configuration optimale d'un réseau de radars contre les trajectoires menaçantes au sein de ce réseau, voir Figure 5. Le problème est vu comme un jeu à deux joueurs entre l'attaquant et le réseau de radar. Les trajectoires optimales sont déterminées par l'équation eikonale, qui est définie par une métrique anisotrope liée à la probabilité de détection par le réseau de radars. Ce cadre fournit un exemple de problème inverse lié à une métrique anisotrope : le choix de la configuration optimale du réseau de radars correspond à l'optimisation de paramètres qui définissent la métrique.

Perspectives

Plusieurs extensions du travail réalisé en thèse sont en considération. Tout d'abord, une extension de la méthode eulérienne pour l'anisotropie orthorhombique semble possible, alors que l'anisotropie qu'on peut actuellement considérer doit être transverse isotrope incliné ou moins complexe. En effet, les coupes 2D d'un tenseur de Hooke orthorhombique sont des tenseurs transverses isotropes inclinés. En généralisant le procédé, il faudrait alors résoudre à chaque point de la grille un problème d'optimisation qui est soit une minimisation en 2D, une maximisation en 2D, ou un problème de point-selle en min-max, à la place de l'actuel problème de maximisation ou de minimasation en 1D.

Il est aussi possible d'utiliser notre solveur eikonale pour des applications dans des problèmes inverses en imagerie sismique. L'anisotropie joue un rôle crucial dans la croûte terrestre, et des solveurs eikonaux anisotropes peuvent être utilisés dans le cadre de la tomographie des temps de première arrivée et d'algorithmes de stéréotomographie [Nol08], pour prendre en compte l'anisotropie non seulement dans la modélisation, mais aussi dans

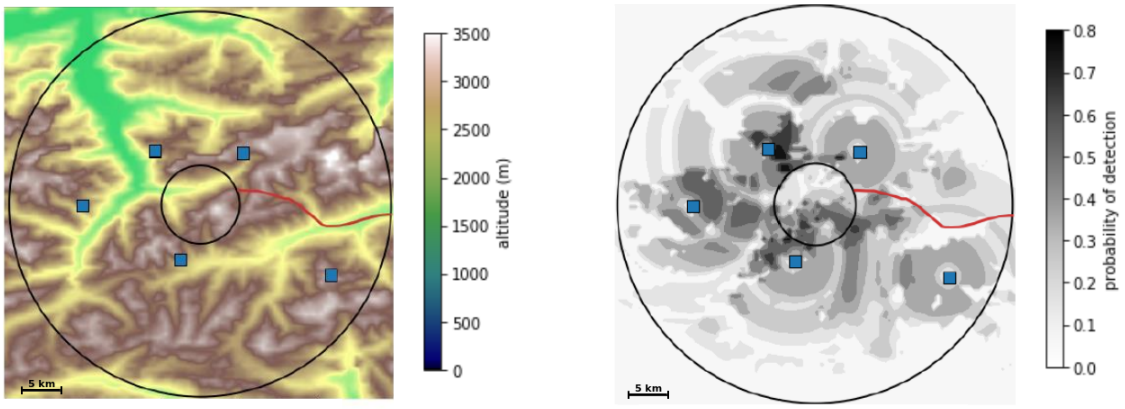

Figure 5: En bleu, position des radars. En rouge, trajectoire la plus menaçante (par rapport à la probabilité de détection le long de la trajectoire) parmi toutes les trajectoires depuis l'anneau extérieur à l'anneau intérieur. La Figure de gauche montre la carte d'élevation de la région, et la Figure de droite montre la probabilité de détection par le réseau de radars.

le problème d'inversion. De plus, l'implémentation GPU réduit grandement les temps de calcul pour le solveur eikonal, ce qui peut permettre de résoudre plusieurs problèmes inverses simultanément, et aller vers l'étude des propriétés statistiques telles que les estimations d'incertitudes pour le problème inverse [TBM19a].

Enfin, un schéma numérique aux différences finies complètement original pour l'équation des ondes élastiques est en considération. Il repose sur la décomposition de Selling du tenseur de Hooke, qui permet de séparer le tenseur en une somme de termes pour lesquels une approximation aux différences finies est possible, de manière similaire à la façon dont la décomposition de Selling a été utilisée dans le cadre de la méthode eulérienne pour décomposer une matrice qui définit une métrique riemannienne. Des tenseurs de Hooke avec une anisotropie complètement générale peuvent être considérés, sans limitation. Des algorithmes pour l'équation des ondes élastiques anisotrope sont déjà considérablement étudiés, mais ce nouveau schéma numérique peut être intéressant en tant que premier schéma aux différences-finies pour la prise en compte de l'anisotropie générale dans l'équation des ondes élastiques.

Preamble

This PhD manuscript is based on several publications which have been written during my PhD training. Some are already published, while some are still under review. For the reader to have an idea of the general content of this work, this manuscript is structured as follows: we start with a general overview of the work performed during the PhD, including a state-of-the-art on numerical solvers for the eikonal equations and applications in geophysics, and a summary of my contributions, before presenting conclusions and perspectives. The detail of my contribution is presented afterwards in dedicated chapters based on the different publications which were realized.

Notations and abbreviations

In the physical space, we use the following notations:

- $\mathbb{E} := \mathbb{R}^n$ is the ambient space (usually with $n = 3$).
- $\Omega \subseteq \mathbb{E}$ is the physical domain in which the wave propagation occurs, with a subset $S \subseteq \overline{\Omega}$ corresponding to the source of the propagation (and usually $S = \{0\}$).
- Letters $x, y \in \Omega$ are used for the position in the physical domain.
- Letters $v, w \in \mathbb{E}$ are used for the velocity vector (or the orientation in case of a normalized vector) in the physical domain.
- $u : \mathbb{E} \to [0, +\infty]$ is the first arrival traveltime of the wave.

In the discretized space, we use the following notations:

- $X \subseteq h\mathbb{Z}^n$, with $h > 0$, is the discretized domain corresponding to Ω, and ∂X is a discretization of S corresponding to the set of source points for the propagation.
- Letters $p, q \in X$ are used for the position in the discretized domain.
- $\Lambda : \mathbb{R}^X \to \mathbb{R}^X$ is the update operator, considered in the finite-difference setting as $\Lambda U = U$, see (22).
- $\mathfrak{F}$ is a scheme, considered for the finite-difference setting $\mathfrak{F}U = 0$.
- $U : X \to [0, +\infty]$ is the numerical approximation of the first arrival traveltime of the wave, see (31).

We also use the following mathematical notations:

- S_n denotes the set of $n \times n$ symmetric matrices, S_n^+ for symmetric semi-definite matrices, and S_n^{++} for symmetric positive definite matrices.
- $\|v\|_M := \sqrt{\langle Mv, v\rangle}$ is the Riemannian norm of a vector v with respect to $M \in S_n^{++}$, with $\langle ., .\rangle$ being the Euclidean scalar product.
- $\|v\|_2 := \sqrt{\sum_i v_i^2}$ is the L^2 norm of a vector v.

In the context of geophysics, we use the following notations:

- c_{ijkl}, with $i, j, k, l \in \{1, 2, 3\}$ is the Hooke tensor and ρ is the density, defining the elastical properties of a 3D medium, see (11).

- We define the matrix $m_c(v) \in S_3$, with c a Hooke tensor and $v \in \mathbb{R}^3$, by $m_c(v)_{ik} := \sum_{j,l} c_{ijkl} v_j v_l$, see (14).

In addition, we use the following abbreviations:

- FMM: Fast Marching method
- FSM: Fast Sweeping method
- GPU: Graphics Processing Unit
- TI: transversely isotropic

  VTI: vertically transversely isotropic

  TTI: tilted transversely isotropic
- TOR: tilted orthorhombic

Part I
General overview of the PhD dissertation

Contents

1 Generalities on the eikonal equation

1.1 Eikonal equation

The eikonal equation is a non-linear partial differential equation which has been first considered in geometric optics [Bru95]. In that context, it represents a generalization of Snell's law for the propagation of a light ray in a continous medium with varying refractive indices. The eikonal equation can be considered in more general settings, for any kind of wave propagation inside a domain equipped with a metric (i.e. a notion of speed). With this equation, one can calculate the first arrival traveltime of the wave everywhere in the domain. The shortest path from the origin of the propagation to another position can also be deduced by moving perpendicularly to the level sets of the first arrival traveltime. In this work, we consider metrics that are anisotropic, meaning that the path length they define depends not only on the position, but also on the orientation of the path at each time.

Before writing the eikonal equation in its general form, we present and discuss some useful mathematical definitions:

- A **gauge** is a 1-homogenous, convex and lower semi-continuous application $F : \mathbb{E} \to [0, \infty]$, with $F(0) = 0$ and positive otherwise. A gauge is the association of a travel cost to a velocity. In applications to seismic models, all the considered gauges are finite and symmetric (i.e. $F(-x) = F(x)$). In contrast, infinite and non-symmetric gauges are encountered in Section 10 for vehicle models.

- The **unit ball for a gauge** F is the compact convex set $B := \{v \in \mathbb{E},\ F(v) \leq 1\}$. The unit ball is the set of all possible velocities for a cost less than or equal to 1. A gauge is said **isotropic** if its unit ball is a disc, meaning that all directions are equivalent. Otherwise, it is **anisotropic** if there are preferential directions. Some examples are presented in Figure 6 in the isotropic, Riemannian and transverse isotropic cases, and Figure 7 for a non-symmetric gauge and a gauge related to vehicle models.

- A **metric** on Ω is an application $\mathcal{F}$ of the form:

$$\mathcal{F} = \begin{cases} \Omega \times \mathbb{E} \to [0, \infty] \\ (x, v) \mapsto \mathcal{F}_x(v) \end{cases} \tag{1}$$

  such that $\mathcal{F}_x$ is a gauge for all $x \in \Omega$, with the corresponding unit balls varying continuously with x according to the Hausdorff distance[1].

- A **path** from $x \in \Omega$ to $y \in \Omega$ is a locally Lipschitz application $\gamma : [0, 1] \to \Omega$ with $\gamma(0) = x$ and $\gamma(1) = y$.

[1]The Hausdorff distance d_H between two non-empty subsets X and Y of a metric space (M, d) is defined as:

$$d_H(X, Y) = \max\{\sup_{x \in X} d(x, Y),\ \sup_{y \in Y} d(X, y)\}$$

- The **cost of the path** γ is defined as

$$\mathcal{C}(\gamma) := \int_0^1 \mathcal{F}_{\gamma(t)}(\gamma'(t))\mathrm{d}t. \tag{2}$$

- The **distance** $d_{\mathcal{F}}$ between x and y corresponds to the minimal cost to travel from x and y: $d_{\mathcal{F}}(x,y) := \min\{\mathcal{C}(\gamma),\ \gamma \text{ path from } x \text{ to } y\}$. The paths which correspond to this minimum are the shortest paths from x to y. Formally, $d_{\mathcal{F}}$ is a quasi-distance: it is homogenous and satisfies the triangular inequality, but is not necessarily symmetric and can have infinite values. In this work, the non-symmetry and infinite cost issues only occur for the vehicle models in Section 10.

- The **first arrival traveltime**, starting from the origin subset $S \subset \overline{\Omega}$, is the function:

$$u : \begin{cases} \Omega \to [0,\infty] \\ x \mapsto d_{\mathcal{F}}(S,x) \end{cases} \tag{3}$$

  We often choose $S = \{0\}$. In other words, the propagation front starts from a single point 0, which is the origin of the coordinate system.

The first arrival traveltime u is not always differentiable, even for smooth metrics. In particular, u is non-differentiable if there are several optimal paths going to the same position, and this position is called "cut-locus", see Figure 8 for an example with a smooth isotropic metric.

The eikonal equation, however, is not written with the metric $\mathcal{F}$, but with what we can define as the dual $\mathcal{F}^*$ of the metric, defined as:

- The dual of a gauge F is the gauge F^* defined by:

$$F^*(v) = \sup\{\langle v, w\rangle, w \in \mathbb{E}, F(w) \leq 1\} \tag{4}$$

  The dual gauge is the association of a travel cost to a slowness (i.e. the inverse of a velocity).

- The dual of a metric $\mathcal{F}$ is the metric $\mathcal{F}^*$ which associates the dual gauge at each position.

This notion of duality can be seen as a special case of Legendre-Fenchel duality[2]. Note also that this is an involution: $F^{**} = F$.

[2]The Legendre-Fenchel dual of a convex function $f : E \to]-\infty,\infty]$ is defined as

$$f^{*LF}(x) = \sup_y \langle x, y\rangle - f(y).$$

The relation with norm duality, considered in (4) and applicable only to 1-positively-homogeneous F, is that

$$\frac{1}{2}F^*(x)^2 = \sup_y \langle x, y\rangle - \frac{1}{2}F(y)^2.$$

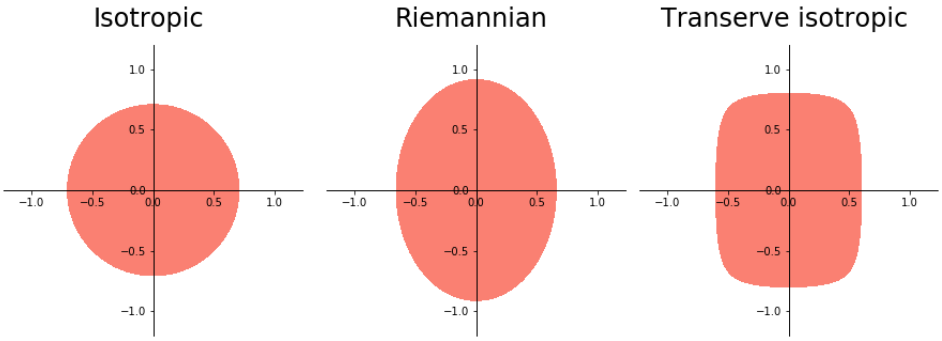

Figure 6: Examples of unit balls of gauges in the isotropic and Riemannian cases, and one corresponding to the dual gauge in the transverse isotropic case (see Section 2.3).

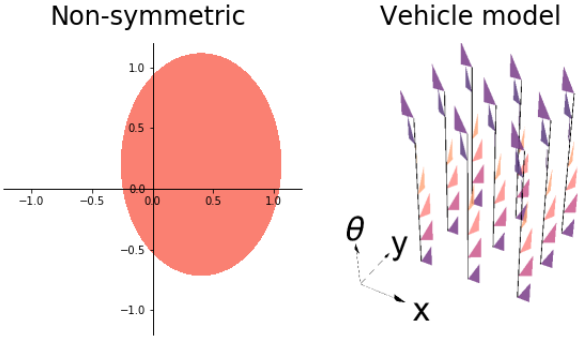

Figure 7: Examples of unit balls of gauges for a non-symmetric gauge, and for a gauge taking infinite values related to vehicle model in a 3D space with the vertical axis representing the orientation of the vehicle (see Section 10).

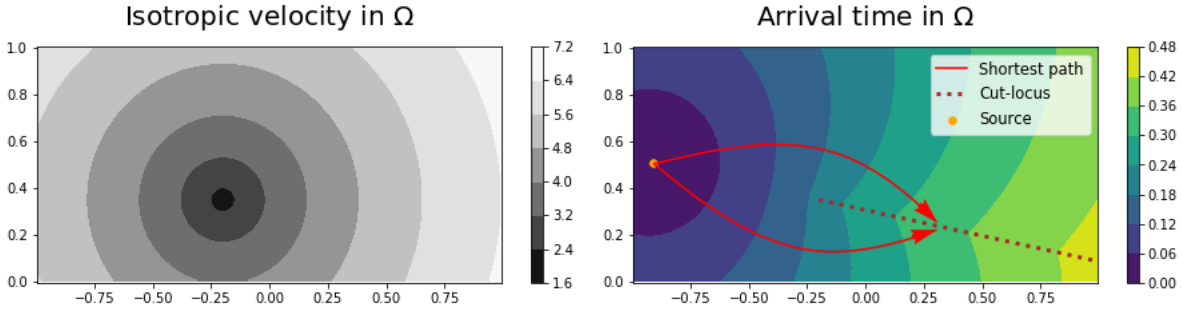

Figure 8: *Left:* visual representation (in a 2D physical space) of an isotropic metric through a color map, with the color indicating the radius of the unit ball of the gauge at each position, i.e. the isotropic velocity of the wave (for a cost equal to 1). *Right:* corresponding first arrival traveltime in this physical space, with a source point to the left. A cut-locus can be observed behind (relative to the source point) the area with low velocity, and corresponds to the location in which two distinct shortest paths are possible, as represented by the two example of shortest paths.

We can now present the eikonal equation: the first arrival traveltime u is solution to the eikonal equation (in the sense of viscosity solutions [CEL84]), which is a partial

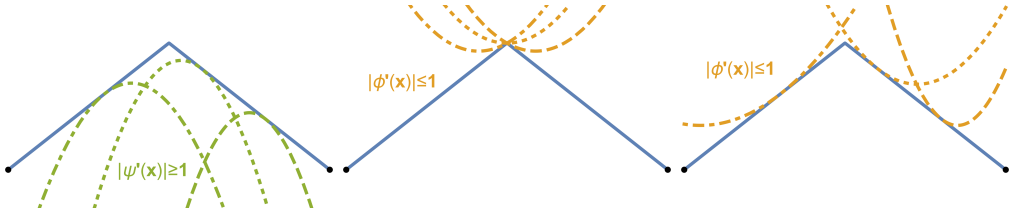

Figure 9: Example of viscosity solution u in blue, with test-functions $\psi \in C^2(\Omega)$ in green (from below) and $\phi \in C^2(\Omega)$ orange (from above), considering here the isotropic metric $\mathcal{F}_x^*(v) = \|v\|_2$.

differential equation of Hamilton-Jacobi-Bellman type, static and of first-order, and is written through the dual metric

$$\mathcal{F}_x^*(\nabla u(x)) = 1, \tag{5}$$

for all $x \in \Omega \setminus S$, where S is the collection of source points. The point source boundary conditions $u(x) = 0$ is applied at the sources points $x \in S$, and outflow boundary conditions are used on $\partial\Omega$.

We briefly present the concept of viscosity solutions: it represents a way to apply a differential operator to a non-differentiable function, through the use of smooth test-functions approaching the solution from below and from above. u is solution of $\mathcal{F}_x^*(\nabla u(x)) = 1$ *in the sense of viscosity solutions* should be understood as follows: let $\phi \in C^2(\Omega)$ arbitrary. If $u - \phi$ attains its minimum at $q \in \Omega$, then $\mathcal{F}_q^*(\nabla u(q)) \geq 1$. If $\phi - u$ attains its minimum at $q \in \Omega$, then $\mathcal{F}_q^*(\nabla u(q)) \leq 1$. Illustrations of these two cases (approach from below and from above) are presented in Figure 9 for an isotropic metric. With these requirements, local minima are not allowed (except for source points, which we consider as part of the boundary of the domain), and the solution must be continuous.

In Table 1, we present a characterization of the various concepts (gauge, unit ball of the gauge, metric, dual metric and eikonal equation) in specific settings (isotropic, Riemannian and general cases).

An essential property of the first arrival traveltime is the fact that a subpath of an optimal path is also an optimal path, which is called the Bellman's optimality principle and is helpful to establish the Lagrangian numerical scheme as in Section 3.2. It can be written as such:

$$u(x) = \inf_{y \in \partial\mathcal{V}(x)} (d_{\mathcal{F}}(y, x) + u(y)), \tag{6}$$

for $\mathcal{V}(x)$ any neighbourhood of x which does not intersect the starting area $\partial\Omega$.

From the knowledge of the first arrival traveltime, it is also possible to deduce the shortest paths by backtracking, i.e. by performing a gradient descent with respect to the metric, which is an ordinary differential equation of the form:

$$\gamma'(t) = \mathrm{d}\mathcal{F}^*_{\gamma(t)}(\nabla u(\gamma(t))), \tag{7}$$

	Isotropic	Riemannian	General
Gauge	$F(v) = \frac{1}{c}\|v\|_2$	$F(v) = \|v\|_M$	$F(v)$ gauge
Unit ball of the gauge	Disk with radius c	Ellipse defined from M	B compact convex
Metric	$\mathcal{F}_x(v) = \frac{1}{c(x)}\|v\|_2$	$\mathcal{F}_x(v) = \|v\|_{M(x)}$	$\mathcal{F}_x(v)$ metric
Dual metric	$\mathcal{F}_x^*(v) = c(x)\|v\|_2$	$\mathcal{F}_x^*(v) = \|v\|_{M^{-1}(x)}$	$\mathcal{F}_x^*(v)$
Eikonal equation	$\|\nabla u(x)\|_2 = \frac{1}{c(x)}$	$\|\nabla u(x)\|_{M^{-1}(x)} = 1$	$\mathcal{F}_x^*(\nabla u) = 1$

Table 1: Characterization of the isotropic and Riemannian metrics, and a general one. For the isotropic case, we define: $c(x) \in \mathbb{R}_+^*$, and for the Riemannian case, we define: $M(x)$ symmetric definite matrix (and the ellipse defined from M is the ellipse with semi-axes of direction X_λ and length $\lambda^{-\frac{1}{2}}$, where X_λ and λ are the eigenvectors and corresponding eigenvalues of M).

where $\mathrm{d}\mathcal{F}^*$ is denoting the differential of the dual metric $(x, v) \mapsto \mathcal{F}_x^*(v)$ with respect to v.

1.2 Applications of the eikonal equation

Computing the first arrival traveltime, as well as the corresponding optimal paths, is a task of interest in various domains, such as: medical image segmentation [Mir14b], modeling of bio-physical phenomena [SKD+07] and motion planning control problems [AM12]. This latter case is studied in Section 10, for the computation of threatening trajectories for vehicles trying to move undetected in a zone monitored by a radar network. Illustrations are presented in Figure 10. We study the best configuration of this radar network against the threatening trajectories.

Several applications also exist in the context of geophysics: eikonal solvers can be used for earthquake hypocenter relocation through backpropagation of the data recorded at the surface by seismic stations[MvN92], asymptotic approximation of Green's functions for Kirchhoff migration to build high resolution images in seismic exploration [Bey87, Ble87, LOP+03], tomographic inversions to determine seismic wave velocities from global and regional scale [Nol08] to exploration and near surface scale targets [BL98, TNCC09, LVF13]. An example of application in seismic tomography is shown in Figure 11: the Earth interior cannot be known physically, but it can be understood through the study of seismic waves propagating inside the Earth. From that, the position and properties of the different layers of the Earth have been assessed.

Computing first arrival traveltimes for wave propagation as in the subsurface leads naturally to consider anisotropic metrics. These applications are the main motivation for the work performed during this PhD, which is presented afterwards.

1.3 State-of-the-art of eikonal solvers

We propose a short review of existing strategies to compute numererically first arrival traveltimes. First, we mention an alternative option to the solution of the eikonal equations: the ray-tracing method [Cer05], which consists in directly computing optimal paths

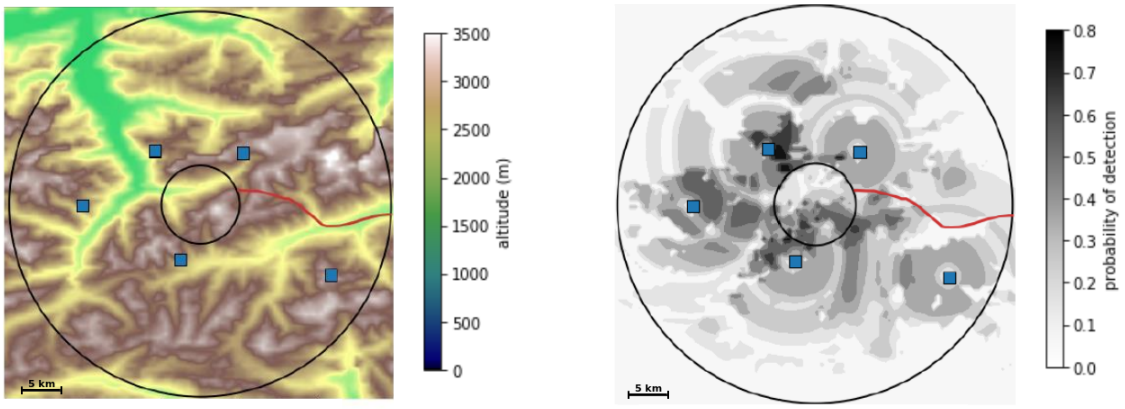

Figure 10: Most threatening trajectory in a radar network. In blue, the location of radars. In red, the most threatening trajectory (with regard to the probability of detection along the path) among all paths from the outside ring to the inner ring. Left Figure shows the elevation map of the region, and right Figure shows the probability of detection from the radar network. In Section 10, we study how to optimize the parameters of the radar network against the most threatening trajectories.

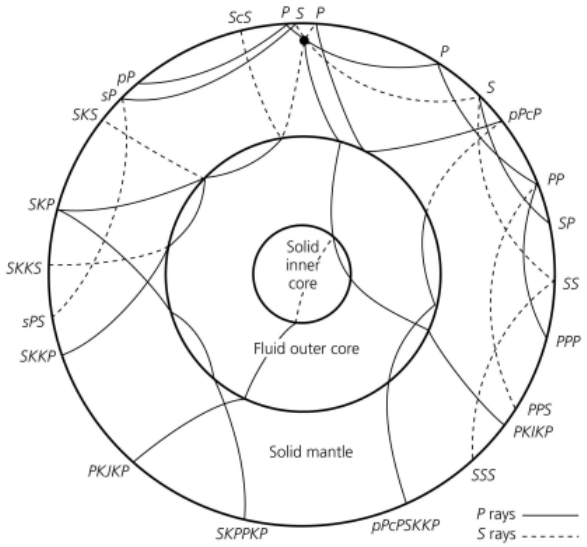

Figure 11: Simplified view of the Earth interior with examples of seismic phases (with names relative to to the phase types P and S, and the number of reflections at discontinuities: inner/outer core, core/mantle, surface), from [SW09].

in the domain. However, several drawbacks have been identified: one ray does not necessarily correspond to the first arrival traveltime, the computation time increases strongly when many travel paths to many points are needed, and calculations can be difficult in shadow zones which can occur even in smooth media. These issues no longer occur when considering numerical schemes for the eikonal equation. On the other hand, computing later arrival traveltimes with an eikonal solver is a difficult problem [RS04], and the ray

method is better suited in that case.

For the eikonal equation, the first finite-difference scheme has been developed by Vidale [Vid88]. It works only for isotropic metrics, and with first-order accuracy. This solver works by induction on the boundary of a square expanding from the source point. It has later been extended to anisotropy [Lec93]. However, the correct solution is not necessarily computed if the anisotropy is too strong, or in case of heterogeneities: causality cannot be guaranteed whenever a ray goes back into the expanding square.

In [OS91], the isotropic eikonal equation is solved by treating it as a dynamic (time-dependent) Hamilton-Jacobi equation, with an "essentially non-oscillatory" (ENO) scheme. This approach has been extended to VTI metrics and high order accuracy in [DS97], with the "down & out" (DNO) strategy. A post-treatment (PS) is added in [KC99], with second-order accuracy, resulting in the ENO-DNO-PS scheme, which was extended to TTI metrics in [Kim99]. However, the method is computationally expensive, and algorithms for the static (no time dependency) eikonal equation have been found to be more efficient.

More efficient algorithms for the static eikonal equation have been developed thanks to the level-set framework [Set96]. The resulting numerical methods can be divided into two classes: *iterative* methods and *single pass* methods, which respectively generalize the algorithms of Bellman-Ford and of Dijkstra for the computation of the first arrival traveltime in graphs.

The best known iterative method is presumably the fast sweeping method. Originally introduced in isotropic settings [Zha05], the fast sweeping method (FSM) has been extended to 2D elliptic anisotropy [TCOZ03]. In the context of geophysics (with metrics detailed in the next section), the FSM has been extended to 2D TTI metrics [LCZ14], 3D TTI metrics [PWZ17] with a third-order Lax-Friedrich fast sweeping scheme, 3D TOR metric [WYF15] by treating it as an iterative problem on elliptic anisotropy, and more recently [LBLM] for the 3D TOR metric with high-order accuracy with a discontinuous Galerkin method. Other iterative methods include the adaptive Gauss-Seidel iteration [BR06], or the buffered fast marching method [Cri09]. Recently, iterative methods which can take advantage of massively parallel computational architecture, GPU in particular, have been proposed in the isotropic settings [JW08], and for elliptic anisotropy [GHZ18].

On the other hand, the best known single pass method is presumably the fast marching method (FMM) [Tsi95, Set96], but the extension of the FMM to anisotropic geometries has proved more difficult. Early studies [KS98, SV01, AM12] involve numerical schemes with wide stencils, leading to increased computation times and reduced accuracy, and therefore negating many of the advantages of the FMM. More recently in [Wah20], an algorithm using the FMM has been developed for the 3D TTI anisotropy: it works by solving a fixed point problem on non-tilted elliptic anisotropy. While the authors illustrate numerically that the algorithm can converge when the considered anisotropy is close from a non-tilted elliptic anisotropy, there is no formal proof of convergence of the fixed

point iteration they implement.

In the recent years, extensions of the FMM to 2D anelliptic anisotropy have been proposed [Mir14b]. Other improvements have been done with 3D elliptic anisotropy [Mir14a, Mir19], and for various types of degenerate anelliptic anisotropy related with curvature penalization [Mir18]. With techniques from lattice geometry, the size of the discretization stencil is kept under control, thus preventing any loss in computation time and accuracy, even in case of a very strong anisotropy (with propagation speed potentially ten times faster in the fast direction compared with the slow directions), see Section 7. In this thesis, we generalize the tools used in this later numerical scheme to develop numerical schemes based on the FMM for metrics with different types of anisotropy, encountered in geophysics.

2 Eikonal equation in the frame of geophysical applications

2.1 First arrival traveltime

In the context of geophysics, an eikonal equation can be obtained as the high-frequency approximation of the elastic wave equation, with the underlying metric defined by the elastic properties of the geological medium. The solution to this eikonal equation corresponds to the first arrival traveltime of the wave. Compared with the eikonal equation, the elastic wave equation gives more information on the behaviour of a seismic wave. Indeed, the solution to the elastic wave equation gives the amplitude of the wave at any time and any position, and the first arrival traveltime can be deduced from it as, for each position, the first time that the amplitude is non-zero, see Figure 12 & 13 for illustrations. However, computing the solution to the wave equation in three dimensional complex media can be numerically expensive: the scale of the discretization grid must be significantly smaller than the oscillation wavelength to prevent numerical dispersion, and the time step is bounded by the Courant-Friedrichs-Levy stability condition. In contrast, the eikonal equation is a static partial differential equation, with a non-oscillatory solution. For these reasons, solutions to the eikonal equation can typically be computed at a much lower cost.

The elastic properties of a medium are usually characterized by the density $\rho(x)$ and a fourth-order elasticity tensor, called Hooke tensor and denoted by $c_{ijkl}(x)$, where $i, j, k, l \in \{1, 2, 3\}$. We present the properties of the Hooke tensor in more details in Section 2.3.

The amplitude displacement vector of a seismic wave is denoted by $\mathcal{U}_i(x, t)$ (along the i-th coordinate axis). First, from Newton's law, we have:

$$\rho \partial_{tt} \mathcal{U}_i = F_i + \sum_j \partial_j \sigma_{ij}, \tag{8}$$

where F is the source field and σ is the stress tensor. From Hooke's law, we also have

$$\sigma_{ij} = \sum_{k,l} c_{ijkl} \epsilon_{kl}, \tag{9}$$

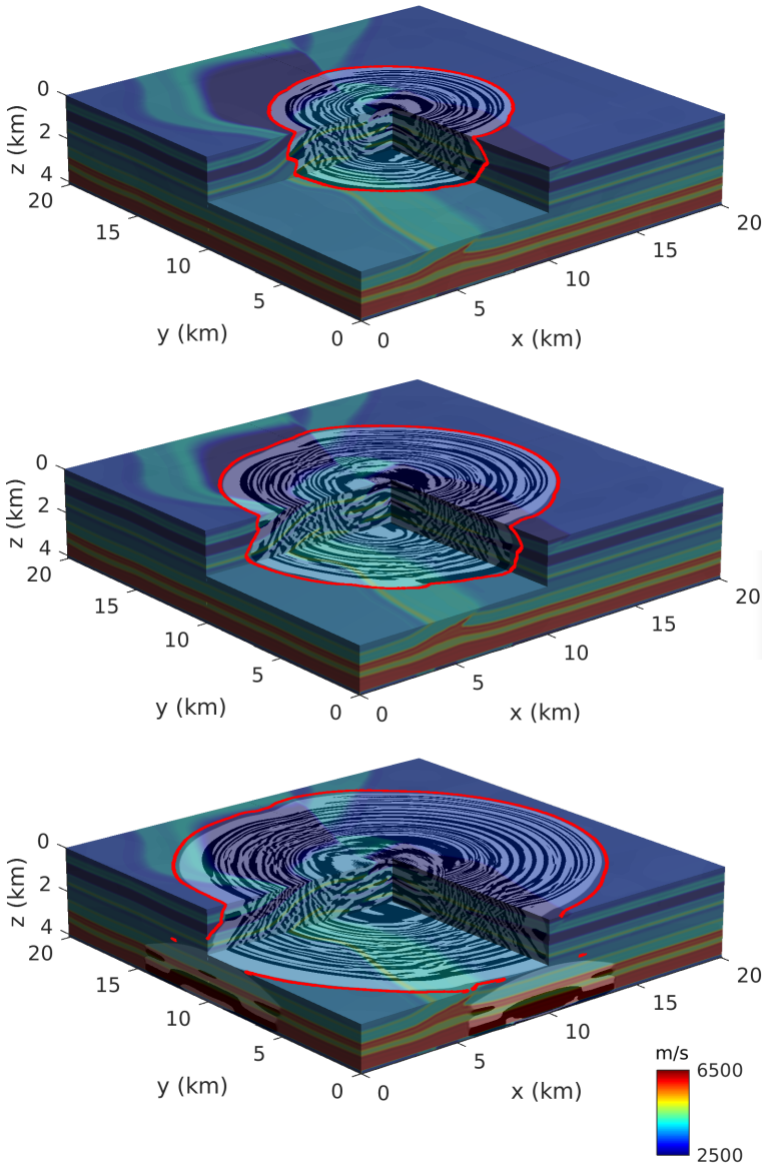

Figure 12: Superposition of the solution to the elastic wave equation (in black and white) and the solution to the eikonal equation (in red), with the background representing the velocity of the wave. The different snapshots are obtained at time: $t = 1.5s$ (top), $t = 2s$ (middle) and $t = 2.5s$ (bottom), for a propagation starting from the middle of the domain. Details on the numerical simulation can be found in Section 6.

where ϵ is the strain tensor, related to the amplitude displacement vector by

$$\epsilon = \frac{1}{2}(\nabla \mathcal{U} + \nabla \mathcal{U}^\top). \tag{10}$$

This relation between the stress tensor and the strain tensor (9) is a behavioral law, in the context of linear elastodynamic.

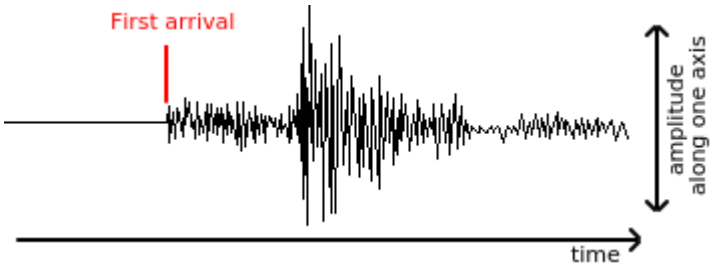

Figure 13: Example of seismogram, with the indication of the first arrival traveltime.

From these equations, we deduce the *elastic wave equation*, which is an equation for the displacement vector $\mathcal{U}$, of the form

$$\rho\partial_{tt}\mathcal{U}_i - \sum_{j,k,l}\partial_j(c_{ijkl}\partial_l\mathcal{U}_k) = F_i, \tag{11}$$

The eikonal equation can be obtained from a high-frequency approximation of the elastic wave equation (following [Sla03]). First, we consider the time Fourier transform, for an angular frequency ω, as

$$\mathcal{U}(x,t) := \widetilde{\mathcal{U}}(x,\omega),$$

which gives the elastic wave equation in the frequency domain

$$\rho\omega^2\widetilde{\mathcal{U}}_i + \sum_{j,k,l}\partial_j(c_{ijkl}\partial_l\widetilde{\mathcal{U}}_k) = 0,$$

considered away from any external source field, i.e. $F = 0$.

We then consider the ray ansatz:

$$\widetilde{\mathcal{U}}(x,\omega) = e^{i\omega u(x)}\sum_{n=0}^{+\infty}A^{(n)}(x)(i\omega)^{-n},$$

where $A^{(n)}(x)$ denotes the vectorial amplitude coefficients and $u(x)$ is the phase function. In the context of high-frequency approximation, we consider only the first term, which means

$$\widetilde{\mathcal{U}}(x,\omega) = A(x)e^{i\omega u(x)},$$

for a vectorial amplitude coefficient $A := A^{(0)}$.

Note that $u(x) = t$ represents the moving wavefront at time t, and therefore $u(x)$ is the traveltime of the propagating wave.

When inserting the ray ansatz into the elastic wave equation, we get

$$\begin{aligned}&\Big(\sum_{j,k,l}c_{ijkl}\partial_j u\partial_l u A_k - \rho A_i\Big)\\ +(\mathrm{i}\omega)^{-1}&\Big(\sum_{j,k,l}c_{ijkl}\partial_j u\partial_l A_k + \sum_{j,k,l}\partial_j(c_{ijkl}\partial_l u A_k)\Big)\\ +(\mathrm{i}\omega)^{-2}&\Big(\sum_{j,k,l}\partial_j(c_{ijkl}\partial_l A_k)\Big) = 0,\end{aligned} \tag{12}$$

which must be satisfied for any frequency ω. In the high-frequency approximation, we can consider only the first term in (12), which gives

$$\sum_{j,k,l} c_{ijkl}\partial_j u \partial_l u A_k - \rho A_i = 0,$$

which can also be written as an equation on the vectorial amplitude coefficient A as

$$\sum_k \big(\sum_{j,l} c_{ijkl}\partial_j u \partial_l u - \rho \,\mathrm{Id}\,\big) A_k = 0. \tag{13}$$

A necessary condition for (13) is the *Christoffel equation*: the first arrival traveltime of the seismic wave, still denoted by $u(x)$, is solution to

$$\det(\rho \,\mathrm{Id} - m_c(\nabla u)) = 0, \quad \text{where } m_c(p)_{ik} := \sum_{j,l} c_{ijkl} p_j p_l \tag{14}$$

However, not only the first arrival traveltime is solution of (14), but also later arrival traveltimes corresponding to arrival traveltimes of qS-waves.

Note that by setting the second term of (12) to zero as well, we recover a transport equation for the amplitude. The third term in (12) can also be taken into account for a better consideration of the effect of a finite frequency for the traveltime in the high-frequency approximation.

We present an alternative writing of the Christoffel equation. This particular formulation will not be used afterwards, but is still interesting to understand its relation to the first arrival traveltime. The Christoffel equation can be factored as (see [Sla03]):

$$(\|p\|_2^2 - \frac{1}{v_1^2(x, \frac{p}{\|p\|_2})})(\|p\|_2^2 - \frac{1}{v_2^2(x, \frac{p}{\|p\|_2})})(\|p\|_2^2 - \frac{1}{v_3^2(x, \frac{p}{\|p\|_2})}) = 0, \tag{15}$$

where $p := \nabla u$ is the slowness vector, and v_1, v_2, v_3 correspond to the velocities of the three types of waves propagating in anisotropic media.

This factoring requires a *positivity* assumption on the Hooke tensor: v_1, v_2, v_3 must be real and positive. The *separability* assumption is also required for our numerical schemes, to distinguish the fastest velocity from the others: $\max\{v_1, v_2\} < v_3$, where by convention $v_1 \leq v_2 \leq v_3$. See Section 8 for more details on positivity and separability. All three velocities correspond to eikonal equations, of the form:

$$\|p\|_2 = \frac{1}{v_i(x, \frac{p}{\|p\|_2})}.$$

The fastest velocity corresponds to the propagation of qP-wave (quasi-pure pressure wave) and consequently the first arrival traveltime, which is the equation of interest in our case. The other two velocities corresponds to qSV and qSH-waves (quasi-pure vertical and horizontal shear waves), with later arrival traveltimes. It is usually possible to factorize

the part corresponding to the qSH-wave, see (19) and Figure 14 for a presentation in the TTI case.

Following the framework of Section 1, we may rephrase this equation using the dual metric $\mathcal{F}_x^*(p;i) := v_i(x, \frac{p}{\|p\|_2})$. It is known that the dual metric $\mathcal{F}_x^*(p;3)$, corresponding to the fastest velocity, is a convex function of p. However, $\mathcal{F}_x^*(p;1)$ and $\mathcal{F}_x^*(p;2)$ are in general non-convex with regard to p, in such a way that the framework of viscosity solutions to eikonal equations does not apply (the viscosity solution then corresponds to a non-physical convexified metric; in constrast, ray-tracing methods can still apply).

We can select only the first arrival traveltime (of the qP-wave) by considering a spectral norm instead of the determinant in the Christoffel equation (14): the first arrival traveltime $u(x)$ is the unique viscosity solution of the *eikonal equation* of the form

$$\mathcal{F}_x^*(\nabla u(x)) = 1, \tag{16}$$

with $u(x) = 0$ for $x \in \partial\Omega$, where

$$\mathcal{F}_x^*(v) := \sqrt{\|m_c(v)\|}, \tag{17}$$

with $\|.\|$ denoting the spectral norm of the matrix $m_c(v)$ (i.e. its highest eigenvalue) defined in (14). Note that this eikonal equation could be written without the square root, but the square root is required to define a 1-homogenous metric. Also, $\mathcal{F}_x^*$ is the dual of the metric $\mathcal{F}_x$, and $\mathcal{F}_x$ can be deduced from $\mathcal{F}_x^*$ by the duality relation: $\mathcal{F}_x(v) = \sup\{\langle v, w\rangle,\ \mathcal{F}_x^*(w) \leq 1\}$. The metric $\mathcal{F}_x(v)$ does not admit a closed form expression in general. The efficient numerical computation of the metric is in itself a non-trivial problem, addressed in Section 6.

2.2 Origin of the anisotropy in geophysics

The velocity of a wave can depend not only on its position, but also on its orientation: the medium of propagation is then called *anisotropic*. For seismic waves propagating in the Earth interior, anisotropy must be taken into account for a proper modeling [BC91]. However, implementing numerical solvers for the eikonal equation with anisotropic metrics is a technical challenge.

Different models of the Earth exist, with varying complexity. The simplest model corresponds to an isotropic metric defined with Lamé parameters. More complex models exist, with anisotropic metrics such as elliptic anisotropy (or Riemannian anisotropy), transverse isotropy, orthorhombic anisotropy and triclinic anisotropy. Details on the different models can be found in Section 2.3.

We identify two main origins for the anisotropy in the Earth interior:

- *Intrinsic anisotropy:* anisotropy can naturally occur from the shape of minerals, and in particular from the layout of crystals at the atomic scale. For example, the olivine crystal can be found in the uppermost mantle under oceans and leads to a preferred direction up to 25% faster than other directions, with a speed profile corresponding to orthorhombic anisotropy [Hes64].

- *Extrinsic anisotropy:* thin structures of isotropic materials, such as sedimentary layers, can also affect the wavefront in the same way that an anisotropic medium would affect them: if the wavelength of the seismic wave is larger than the typical length of the heterogeneities, they are seen as part of an equivalent anisotropic medium by the wavefield. Note that this happens at all scales, as the typical size of the wavelength depends on the application at hand: it can vary from a few meters when studying near-surface propagations, to hundreds of kilometers when considering the totality of the Earth. For sedimentary layers, this homogenization process usually leads to a medium with a rotational symmetry along the layer axis, called "transverse isotropy" (TI). The layers are usually horizontal, leading to a medium with vertical rotational symmetry axis, called "vertical transverse isotropy" (VTI). Some shifts of the symmetry axis can also occur with tectonic movements, leading to "tilted transverse isotropy" (TTI). In case of fractures or in other complex geophysical locations, the equivalent anisotropy derived from the homogenization process can be even more complex, leading to orthorhombic or even triclinic media [CC18].

Before going further, we want to precise some vocabulary to describe the anisotropy of a metric, with a distinction between two concepts:

- *Strength of the anisotropy:* at a given position, the strength refers to the ratio between the highest and the lowest achievable velocity depending on the orientation.

- *Complexity of the anisotropy:* it refers to the number of parameters needed to characterize the metric, i.e. the velocity profile: 1 arameter is needed for isotropic metrics, 3 parameters for Riemannian metrics also referred to as elliptic isotropy (6 parameters in case of a tilt to define the rotation), 5 parameters for TI metrics (7 if tilted), 9 parameters for orthorhombic metrics (12 if tilted). For triclinic media, corresponding to the most general setting for elastic parameters, 21 parameters are required. Note that in the context of geophysics, the metric for the eikonal equation is usually deduced from the Hooke tensor, but sometimes it does not require the use of all the elastic parameters. Indeed, the Hooke tensor describes not only the qP-wave but also the qS-wave propagation, which can require additional parameters. For example, an isotropic Hooke tensor is described by 2 parameters (called Lamé parameters), while 1 parameter is enough for an isotropic metric.

These two concepts both give specific challenges for the design of numerical solvers for the eikonal equation: most numerical solvers are limited to a particular complexity (up to VTI for most existing numerical schemes), and could also fail if the strength of the anisotropy is too pronounced. Besides, these two concepts of strength and complexity are independent, as can be seen in Figure 14: in this example, the metric with elliptic anisotropy has a much stronger anisotropy than the VTI metric even though it has a less complex form of anisotropy, as can be seen through a visual representation of metrics (called slowness surface, see Section 2.3).

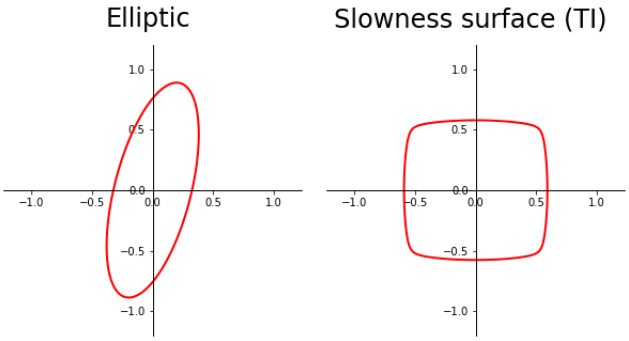

Figure 14: Examples of 2D cross-sections of metrics (represented by their slowness surface, i.e. the unit ball of the dual gauge, see Section 2.3) : for each orientation from the origin, the red curve indicates the slowness (inverse of the speed) of the wave. *Left:* elliptic anisotropy (low complexity), with high strength. *Right:* transversely isotropy (higher complexity), with low strength.

2.3 Properties of the Hooke tensor

A 3D geological medium is described by a fourth-order elasticity tensor, referred to as the Hooke tensor and denoted by $c = (c_{ijkl})$, where $i, j, k, l \in \{1, 2, 3\}$, and by the density ρ of the medium. When studying the first-arrival traveltime, only the ratio $\frac{c}{\rho}$ actually matters, and so it is possible to set $\rho = 1$ without loss of generality, by taking it into account in the Hooke tensor.

The Hooke tensor is subject to the symmetry relations $c_{ijkl} = c_{jikl} = c_{klij}$. Therefore, it only has 21 independent components, allowing it to be represented as a 6×6 symmetric matrix $\mathfrak{C}$ using Voigt notation: $\mathfrak{C}_{v_{ij}, v_{kl}} := c_{ijkl}$, with v defined as: $v := \begin{bmatrix} 1 & 6 & 5 \\ 6 & 2 & 4 \\ 5 & 4 & 3 \end{bmatrix}$

Some additional symmetries are often considered for a geological medium.

- First, an *isotropic* Hooke tensor, for which slowness surfaces are circles, is of the form:
$$c_{ijkl} = \lambda \delta_{ij} \delta_{kl} + \mu(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk})$$
with Lamé parameters λ and μ. With the Voigt notation, it leads to a matrix of the form: $\mathfrak{C} = \begin{bmatrix} \lambda + 2\mu & \lambda & \lambda & 0 & 0 & 0 \\ \lambda & \lambda + 2\mu & \lambda & 0 & 0 & 0 \\ \lambda & \lambda & \lambda + 2\mu & 0 & 0 & 0 \\ 0 & 0 & 0 & \mu & 0 & 0 \\ 0 & 0 & 0 & 0 & \mu & 0 \\ 0 & 0 & 0 & 0 & 0 & \mu \end{bmatrix}$.
- A *transversely isotropic* medium is a geological medium whose local elasticity properties are invariant by rotation around a specific axis. It is called *vertically transversely isotropic* (VTI) in case of invariance around the vertical axis, and *tilted*

transversely isotropic (TTI) otherwise. In the case of VTI symmetry, the Hooke tensor (in Voigt notation) only has 5 independent elastic parameters and can be written as [Tho86]:

$$\mathfrak{C}^{VTI} = \begin{pmatrix} c_{11} & c_{12} & c_{13} & 0 & 0 & 0 \\ c_{12} & c_{11} & c_{13} & 0 & 0 & 0 \\ c_{13} & c_{13} & c_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & c_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & c_{44} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{c_{11}-c_{12}}{2} \end{pmatrix}.$$

- An *orthorhombic* medium is a medium with three mutually orthogonal planes of symmetry, which leads to a Hooke tensor with 9 independent elastic parameters. In Voigt notation, the corresponding elasticity matrix is:

$$\mathfrak{C}^{orthorhombic} = \begin{pmatrix} c_{11} & c_{12} & c_{13} & 0 & 0 & 0 \\ c_{12} & c_{22} & c_{23} & 0 & 0 & 0 \\ c_{13} & c_{23} & c_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & c_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & c_{55} & 0 \\ 0 & 0 & 0 & 0 & 0 & c_{66} \end{pmatrix}.$$

- A *triclinic* medium is a medium with no plane of symmetry, which leads to a Hooke tensor with 21 independent elastic parameters, which is the most general form of anisotropy. In Voigt notation, the corresponding elasticity matrix is:

$$\mathfrak{C}^{triclinic} = \begin{pmatrix} c_{11} & c_{12} & c_{13} & c_{14} & c_{15} & c_{16} \\ c_{21} & c_{22} & c_{23} & c_{24} & c_{25} & c_{26} \\ c_{31} & c_{32} & c_{33} & c_{34} & c_{35} & c_{36} \\ c_{41} & c_{42} & c_{43} & c_{44} & c_{45} & c_{46} \\ c_{51} & c_{52} & c_{53} & c_{54} & c_{55} & c_{56} \\ c_{61} & c_{62} & c_{63} & c_{64} & c_{65} & c_{66} \end{pmatrix}.$$

In the following, we focus on the TTI symmetry. A Hooke tensor with TTI symmetry can be obtained from a Hooke tensor with VTI symmetry C^{VTI} and a 3×3 rotation matrix R defining the axis of rotation, through the change of variables formula:

$$c^{TTI}_{i'j'k'l'} = \sum_{i,j,k,l} c^{VTI}_{ijkl} R_{ii'} R_{jj'} R_{kk'} R_{ll'}. \tag{18}$$

Besides, since the VTI Hooke tensor has one rotational axis of symmetry, only two angles are enough to define the rotation for a TTI Hooke tensor.

Conversely, a (non-convex) projection procedure allows to reconstruct a VTI tensor and a rotation R from a given Hooke tensor [CMA+20], up to some accuracy loss if the latter only has approximate TTI symmetry.

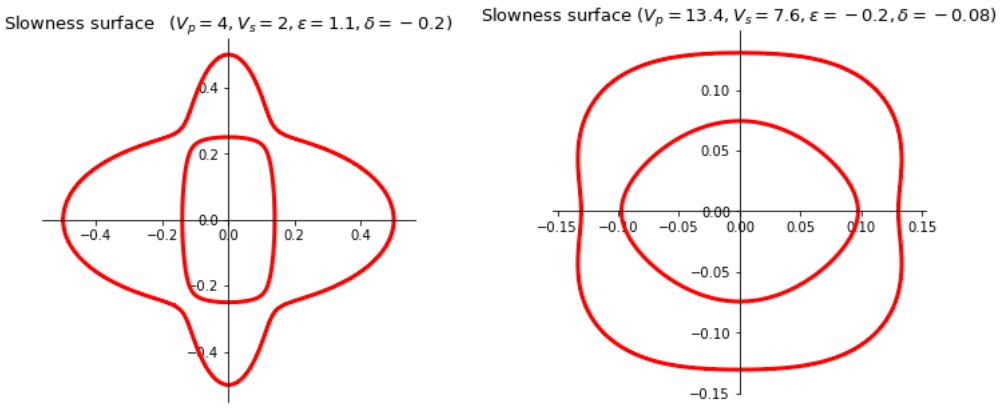

Figure 15: Example of slowness surfaces for the TTI anisotropy, with the inner surfaces corresponding to the qP-wave, and the outer surfaces corresponding to the qSV-wave.

We now consider the Christoffel equation, presented in (14), in the context of a TTI metric. Define the *slowness vector* as $(p_x, p_y, p_z) := R\nabla u$, and let $p_r^2 := p_x^2 + p_y^2$. Then the Christoffel equation for a TTI symmetry can be algebraically factored as follows:

$$(\frac{c_{11} - c_{12}}{2} p_r^2 + c_{44} p_z^2 - 1) \times \tag{19}$$

$$(c_{11}c_{44}p_r^4 + c_{33}c_{44}p_z^4 - (2c_{13}c_{44} + c_{13}^2 - c_{11}c_{33})p_r^2 p_z^2 - (c_{11} + c_{44})p_r^2 - (c_{33} + c_{44})p_z^2 + 1) = 0. \tag{20}$$

The first factor of this equation characterizes the arrival traveltime of the SH (horizontal shear wave) propagation. This factor defines a Riemannian eikonal equation, which can be solved numerically [Mir19], but is of no interest for the computation of the first arrival traveltime. The second factor corresponds to the coupling P-SV, between the qP (quasi-pure pressure wave) and the qSV (quasi-pure vertical shear wave), and is the factor we need to consider for the first arrival traveltime. Obtaining the fully factored Christoffel equation such as in (15) is however not straightforward.

The P-SV equation for a TTI symmetry, henceforth referred to as the TTI eikonal equation, is a non-Riemannian anisotropic eikonal equation of degree four, mathematically more complex than the SH equation. It can be summarized as

$$ap_r^4 + bp_z^4 + cp_r^2 p_z^2 + dp_r^2 + ep_z^2 = 1, \quad \text{where } (p_x, p_x, p_z) = R\nabla u \text{ and } p_r^2 := p_x^2 + p_y^2, \tag{21}$$

with coefficients (a, b, c, d, e) derived from the Hooke tensor as above. For the Hooke tensors considered in geophysics, the TTI equation has two distinct surfaces that are solutions. In the (p_x, p_y, p_z) coordinate system, these solutions are called *slowness surfaces*, and are invariant by the rotation R. The inner surface corresponds to the slowness of the P wave (that is, the inverse of its velocity), whereas the outer surface corresponds to the slowness of the S wave, see Figure 15.

Thomsen's elastic parameters $(V_p, V_s, \epsilon, \delta)$ correspond to another approach to obtain

the TTI eikonal equation (21), with the conversion formula [Tho86]:

$$V_p = \sqrt{\frac{c_{33}}{\rho}}, \qquad V_s = \sqrt{\frac{c_{44}}{\rho}}, \qquad \varepsilon = \frac{c_{11} - c_{33}}{2c_{33}}, \qquad \delta = \frac{(c_{13} + c_{44})^2 - (c_{33} - c_{44})^2}{2c_{33}(c_{33} - c_{44})}.$$

Thomsen's parameters have physical interpretations in a weakly anisotropic setting: in particular, V_p approximates the speed of the qP-wave, and V_s of the qS-wave. Nevertheless this is only an approximation in a special asymptotic setting, and in general both the P and S slowness surfaces depend on the four Thomsen's parameters. For this reason we do not use here the convention $V_s = 0$, which has sometimes been considered to simplify the PDE (21) such as in [LBMV18], when one is only interested in the first arrival traveltime computation corresponding to the qP-wave.

3 Contributions: Fast Marching solvers for anisotropic media

3.1 Fast Marching method

Among the numerical methods to compute the shortest paths, we make a distinction between causal schemes and non-causal schemes. In the discrete formulation, causality (defined in Section 3.2.1) represents the deterministic nature of the underlying optimal control problem. It is equivalent, for a problem on graphs, to the positivity of the length of the edges, which enables the fast computation of shortest paths by the Dijkstra algorithm. Causal numerical schemes can be solved by the Fast Marching algorithm.

The Dijkstra algorithm has been published in 1959 by E. Dijkstra [D$^+$59]. It can calculate the shortest paths in a discrete oriented graph, with positive weights on the edges. For a graph with n vertices, the complexity of the Dijkstra algorithm is in $n \log(n)$: the computation can be done in a single-pass in the graph, by updating the propagation front from the source point.

Graph based methods cannot have a high degree of consistency, since the front is constrained to travel along the edges, and so the Fast Marching algorithm has been developped to overcome this constraint [Tsi95, Set96]. This algorithm is similar to the Dijkstra algorithm and uses the fact that the information propagates from the source point, but it considers that the front can travel with any orientation locally, and not only along the edges.

The discretized eikonal equation, whose solution is an approximate distance map, usually takes the form of a fixed point problem

$$\Lambda U = U, \tag{22}$$

for all $U \in X \setminus \partial X$, where the unknown is $U : X \to \mathbb{R}$ with $U = 0$ on ∂X, with X a discretization of the domain Ω, ∂X a discretization of the source S, and $\Lambda : \mathbb{R}^X \to \mathbb{R}^X$

Algorithm 1 Fast Marching

Require: X, ∂X, Λ, V
  Label ∂X as $Front$, and other points as Far.
  Initialize U with value $+\infty$ on X, and boundary conditions on ∂X.
  While $Front$ is not empty **Do**
    Find a point q of $Front$ which minimizes U.
    Label q as *Accepted*.
    For all neighbour $p \in V(q)$ labelled as $Front$ or Far **Do**
      If p is Far, label p as $Front$.
      Modify $U(p)$ as $\Lambda U(p)$.
    End For
  End While
  Return U

an operator. Outflow boundary conditions are natural in the discretized eikonal equation, and require no special treatment. For this reason, and at the price of possibly some inconsistency of notation with the continuous case, we denote by ∂X the set of source points for the propagation in the discrete case, where the boundary condition $U = 0$ is applied.

If Λ is monotonous and causal (see Section 3.2.1 for definitions), then (22) can be solved in a single pass using the FMM. If Λ is monotonous but (possibly) not causal, then iterative methods such as FSM can be used to solve (22) We also define the stencil V_x at each point $x \in X$, as the set of all points in the neighbourhood of x used in the numerical evaluation of $\Lambda U(x)$.

The algorithm for the FMM is presented in Algorithm 3.1. We use the labels Far, $Front$ and *Accepted*: each point of the domain is labelled as *Accepted* only once, and the value of the numerical solution U does not change any longer afterwards, which is why this method is a single-pass method. During the algorithm, the points labelled as $Front$ can be seen as a discretization of the propagation front of the wave.

It is also possible to use pre-processing and post-processings when considering each point to enhance the performance of the algorithm, in particular by using a source factorization, or for higher-order numerical methods.

The use of FMM was first limited to isotropic metrics, but has been extended to deal with strong Riemannian anisotropy with work from [Mir14a]. In the next sections, we show generalizations of this work to obtain operators Λ and stencils V for numerical schemes which can be solved by the FMM. Two methods are presented: the "semi-Lagrangian" method and the "Eulerian" method. Even though both methods give numerical solutions to the eikonal equation, they are based on different discretized formulations, with various advantages and drawbacks depending on the use-case. The generalizations of these methods in the context of geophysics are the main contributions of this PhD, leading to associated publications presented in Section 6 and Section 8.

3.2 Semi-Lagrangian scheme for the eikonal equation

3.2.1 Semi-Lagrangian scheme

First, we define two properties for operators $\Lambda : \mathbb{R}^X \to \mathbb{R}^X$:

- An operator is called "monotonous" if

$$\forall U_1, U_2 \in \mathbb{R}^X,\ U_1 \leq U_2 \implies \Lambda U_1 \leq \Lambda U_2.$$

- An operator is called "causal" if

$$\forall U_1, U_2 \in \mathbb{R}^X, t \in \mathbb{R},\ U_1^{<t} = U_2^{<t} \implies (\Lambda U_1)^{\leq t} = (\Lambda U_2)^{\leq t},$$

  where $U^{<t}(p) = (U(p) \text{ if } U(p) < t,\ +\infty \text{ otherwise})$.

The FMM computes the exact solution of (22) if the operator Λ is both monotonous and causal. Iterative schemes can solve (22) using monotony only, but are generally slower than the FMM.

Besides, by considering the convergence of the viscosity solution for optimal control problems, we can show that the solution of $\Lambda_h U_h = U_h$ (on a grid with scale h) converges towards the solution of the eikonal equation when $h \to 0$.

Semi-Lagrangian methods for the computation of the first arrival traveltime are based on an identity satisfied by the solution and known as Bellman's optimality principle, presented in Equation 6 and that we recall here:

$$u(x) = \inf_{y \in \partial \mathcal{V}(x)} (d_{\mathcal{F}}(y, x) + u(y)),$$

for $\mathcal{V}(x)$ a neighbourhood of x which does not intersect the starting area $\partial\Omega$.

Semi-Lagrangian numerical schemes translate this principle in a discrete space: denote by X and ∂X discrete domains representing Ω and $\partial\Omega$, and for all $p \in X$, $V(p)$ a polytope containing p and whose vertices belong to X, called the *stencil*. For the discretization of Bellman's optimality principle, we define $\Lambda : \mathbb{R}^X \to \mathbb{R}^X$ as, for $U : X \to [0, \infty]$ and $p \in X \setminus \partial X$:

$$\Lambda U(p) = \inf_{q \in \partial V(p)} \mathcal{F}_p(p - q) + I_{V(p)} U(q). \tag{23}$$

The distance $d_{\mathcal{F}}$ is approximated by the local metric $\mathcal{F}_p(p - q)$, and the values U at positions $q \in V(p)$ that are not vertices are approximated by the linear interpolation $I_{V(p)}U(q)$ on the faces of the stencil $V(p)$. The discrete counterpart of Bellman's optimality principle takes the form of the fixed point equation:

$$\Lambda U = U, \tag{24}$$

with $U(p) = 0$ for all $p \in \partial X$.

The operator Λ defined from (23) is a monotonous operator, but it is not always causal. For Λ to be causal, the stencils have to verify a geometric property of *acute angles* with regard to the gauges of the metric.

- *(F-acute angle)* We introduce a generalized measure of angle, associated with a gauge. Let F be a gauge, which is differentiable except at the origin, and let $p, q \in \mathbb{R}^2 \setminus \{0\}$. We say that p, q form an F-acute angle if $\langle \nabla F(p), q \rangle \geq 0$.
- *(F-acute stencil)* A stencil V is called *acute* with respect to a gauge F if for all p, q in a common facet of V, p and q form an F-acute angle.

If all stencils are acute with regard to the gauges of the metric, the operator Λ from (23) is causal and the FMM can solve the fixed point problem of (22) in a single pass. For an isotropic metric, it simply corresponds to the stencil having acute Euclidian angles, when considering each angle between two neighbouring vertices measured from the center of the stencil. Therefore, even the simplest octahedron stencil is always acute in an isotropic setting.

However, for an anisotropic metric, the stencil needs to be more refined in directions in which the anisotropy is stronger, and a precise geometric study of the relation between the anisotropy and the shape of the stencils is required. On the other hand, we also want to keep the stencils with a minimal size, to improve accuracy and reduce computation time. If the stencils grow too large, the advantages of the FMM are negated and the use of iterative schemes with simpler stencils could lead to a better result. Examples of stencils with increasing complexity are presented in Figure 16.

In the case of Riemannian metrics, very efficient acute stencils can be constructed [Mir14a], by considering a Minkowski-reduced basis of vectors for the symmetric definite matrix defining the Riemannian metric. However, for more general types of anisotropy, building acute stencils is a challenging problem. In the next section, we consider the use of fixed stencils to solve the eikonal equation in the context of geophysics with the semi-Lagrangian method.

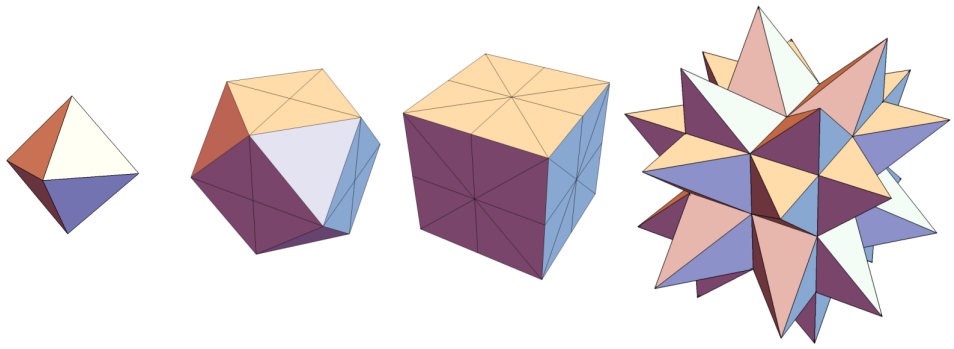

Figure 16: 3D stencils of increasing complexity, used in the finite-difference scheme for the eikonal equation using the semi-Lagrangian method (octahedron, cut-cube, cube and spiky-cube).

3.2.2 Semi-Lagrangian scheme for geophysics

In [DCC+21] (shown in Section 6), the eikonal equation is considered with a geological medium defined with the most complex form of anisotropy considered in geophysics, i.e. a triclinic medium, with a fully general Hooke tensor (21 independent parameters). The numerical scheme can handle this complex anisotropy, but it has a (known) limitation based on the strength of the anisotropy, which can be directly linked to the 3D stencils that we choose to use in the numerical scheme. Using more refined 3D stencils than the spiky-cube from Figure 16 could allow to handle stronger anisotropy, but the numerical cost of the FMM would become prohibitive as it is proportional to the number of points of the stencil, and would destroy its competitive advantage over the FSM. Besides, building adaptative stencils $V(p)$, i.e. locally adapted to the metric $\mathcal{F}_p$ to be as small as possible, is a difficult problem for non-Riemannian metrics, and we only consider fixed stencils chosen from Figure 16.

The stencils are chosen before solving the eikonal equation, and we can check that they verify the acuteness property with regard to the gauge, as defined in Section 3.2.1. Once the stencils are chosen, they do not need to be updated anymore during the Fast Marching algorithm or in case of a change of source point with the same medium.

The strength of the anisotropy is usually not too pronounced in geophysics, and we show that the cut-cube stencil (see Figure 16) is good enough to be acute for most geological media. The cube or spiky-cube stencils can handle anisotropy of even greater intensities, and are required for crystallographic media. Alternatively, if the stencils are not acute, an iterative method such as the FSM may be used instead of the FMM.

The limitations of this algorithm can be expressed from the anisotropy ratio of a gauge F, defined as

$$\mu(F) := \max_{|u|=|v|=1} \frac{F(u)}{F(v)}. \tag{25}$$

In Table 2, we present the condition on $\mu(F)$ under which each stencil from Figure 16 is guaranteed to be F-acute, and so provide a causal numerical scheme that can be solved from the FMM.

In Figure 17, we illustrate these limitations in the case of TTI media, accounting for the worst possible rotation of the metric. We consider the TI examples presented in [Tho86], and observe that almost all of them can be solved with the cut-cube stencil, and only a few require the cube or spiky-cube. Note that the numerical method can also handle anisotropy more complex than TTI.

In the rest of this section, we present the technical difficulty that is encountered when trying to apply the semi-Lagrangian scheme to metrics from geophysics. Indeed, the FMM requires the computation of the update operator Λ of (23). This operator is used to compute the first arrival traveltime at any position as a minimization problem over first arrival traveltimes estimated on the facets of the stencil at this position. However,

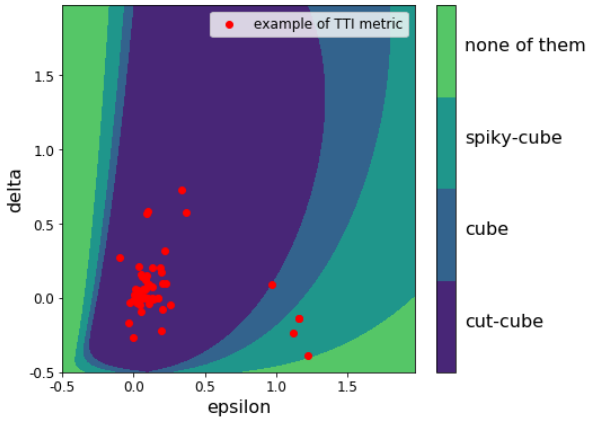

Figure 17: Acuteness property for TTI metrics. The Figure shows whether a given stencil is acute with respect to a TTI metric determined by Thomsen parameters (ϵ, δ), with $V_p = 1$ and $V_s = 0$, while taking into account the worst possible rotation of the metric. The domain labelled for the spiky-cube also contains the domain labelled for the cube, which also contains the domain labelled for the cut-cube. Red dots show all examples of TI metrics presented in [Tho86].

	Octahedron	Cut-cube	Cube	Spiky-cube
Elliptic	iff $\mu = 1$	iff $\mu \leq \sqrt{3}$	iff $\mu \leq (1+\sqrt{3})/\sqrt{2}$	iff $\mu \leq 1+\sqrt{2}$
Anelliptic	iff $\mu = 1$	if $\mu \leq 2/\sqrt{3}$	if $\mu \leq \sqrt{3/2}$	if $\mu \leq \sqrt{2}$

Table 2: Condition under which a stencil V satisfy the acuteness property for a gauge. Note that the bound is sharp for elliptic norms (if and only if), but only sufficient for anelliptic gauges (except for the octahedron). Stencils illustrated in Figure 16. Numerical values for the first two lines: $(1, 1.73, 1.93, 2.41)$ and $(1, 1.15, 1.22, 1.41)$.

the computation of Λ involves the computation of the primal metric $\mathcal{F}$. In the context of geophysics, the primal gauge F can only be expressed from the dual gauge F^* by the duality relation, which is a maximization of a linear form subject to a non-linear convex constraint:

$$F(p) = \sup\{\langle p, q\rangle, q \in \mathbb{E}, F^*(q) \leq 1\} \tag{26}$$

with

$$F^*(q) := \sqrt{\|m_c(p)\|},$$

for a Hooke tensor c_{ijkl}, with $\|.\|$ denoting the spectral norm of the matrix $m_c(p)$ (i.e its highest eigenvalue) defined by

$$m_c(p)_{ik} := \sum_{j,l} c_{ijkl} p_j p_l.$$

We rely on sequential quadratically-constrained quadratic programming to address this problem numerically. For this method, the constraint needs to take the form "$f \leq 1$" where the function f is both:

- Strongly convex.
- Efficiently evaluated numerically, as well as its gradient and Hessian.

The constraint $F^* \leq 1$ in (26) does not satisfy any of these properties. We consider an alternative expression of the constraint, for $\alpha \geq 0$, of the form

$$\exp[-\alpha f_c] \leq 1. \tag{27}$$

where $f_c(p) := \det(\mathrm{Id} - m_c(p))$.

We show that the function $\exp[-\alpha f_c]$ is smooth, defined over the whole of $\mathbb{R}^d$, strongly convex in the domain of interest, and easy to evaluate, thus complies with all the requirements. We can then replace the highly non-linear constraint $F^* \leq 1$ with the constraint (27).

3.2.3 Stencil construction in 2D

In the case of dimension two, the construction of acute stencils is possible for all forms of anisotropy, with an algorithm presented in [Mir14b] and based on the Stern-Brocot tree. A 2D acute stencil is obtained by starting from a simple stencil and successively refining it in the directions in which the causality is not verified due to the strength of the anisotropy, see Figure 18. With that, we obtain an algorithm to provide 2D stencils locally adapted to the metric at all positions, and the resulting size of the stencils can be compared with the strength of the anisotropy. Such a construction cannot be easily generalized to dimension three, which is why we have to restrict ourselves to fixed 3D stencils in the numerical scheme presented in Section 6.

In [MD20] (presented in Section 7), we study a stronger requirement of acuteness, called "strict acuteness". Strict acuteness leads to a stricter version of causality in the

numerical scheme, and it is helpful to ensure that the numerical scheme remains causal even after small perturbations, with such perturbations potentially arising from source factorization or methods for high-order accuracy. We study the size and number of vertices for 2D stencils built with this algorithm, in the worst case and in average case over rotations of the anisotropic gauge (referred to as asymmetric norm in Section 7).

We present a few definitions to finally arrive at the concept of (F, α)-acute stencil, for a gauge F and $\alpha \in]0, \pi/2]$.

- *(unoriented Euclidean angle)* We denote by $\measuredangle(u, v) \in [0, \pi]$ the unoriented Euclidean angle between two vectors $u, v \in \mathbb{R}^2 \setminus \{0\}$, which is characterized by the identity

$$\cos \measuredangle(u, v) = \frac{\langle u, v \rangle}{\|u\| \, \|v\|}.$$

- *(F-angle)* We define the F-angle $\measuredangle_F(u, v) \in [0, \pi/2] \cup \{\infty\}$ by

$$\cos \measuredangle_F(u, v) := \langle \nabla F(u), v \rangle / F(v) \tag{28}$$

  if u, v form an F-acute angle. Otherwise we let $\measuredangle_F(u, v) := +\infty$. If F is isotropic, i.e. if $F(x) = c \, \|x\|$ with $c > 0$, then the F-angle is simply the Euclidian angle.

- *(2D stencil)* A 2D stencil is a finite sequence of pairwise distinct vectors $u_1, \cdots, u_n \in \mathbb{Z}^2$, $n \geq 4$, such that

$$\det(u, v) = 1, \qquad \langle u, v \rangle \geq 0,$$

  for all $u = u_i$, $v = u_{i+1}$, $1 \leq i \leq n$, with the convention $u_{n+1} := u_n$.

- *((F, α)-acute stencil)* A stencil is said (F, α)-acute, where F is a gauge and $\alpha \in]0, \pi/2]$, iff with the same notations one has

$$\measuredangle_F(u, v) \leq \alpha, \qquad \measuredangle_F(v, u) \leq \alpha. \tag{29}$$

  We let $N(F, \alpha)$ denote the minimal cardinality of an (F, α)-acute stencil.

The cardinality of (F, α)-acute stencils is directly proportional to the algorithmic complexity of our eikonal PDE solver, hence it is important to choose them as small as possible. When $\alpha = \pi/2$, we recover the typical acute stencils considered before in Section 3.2.1.

In pratice, we choose $\alpha = \pi/3$, which allows to keep the acute angle property under reasonably small perturbations, while not increasing too much the number of points of the stencil.

The main result is the following estimate of the cardinality $N(F, \alpha)$ of an (F, α)-acute stencil, both in the worst case and in the average case over random rotations of the gauge F. For any gauge F and any $\alpha \in]0, \pi/2]$, one has

$$N(F, \alpha) \leq C \frac{\mu}{\alpha^2} \ln\big(\frac{\ln \mu}{\alpha^2}\big), \qquad \int_0^{2\pi} N(F \circ R_\theta, \alpha) \, \mathrm{d}\theta \leq C \frac{\ln(\mu)}{\alpha^2} \ln\big(\frac{\mu}{\alpha^2}\big). \tag{30}$$

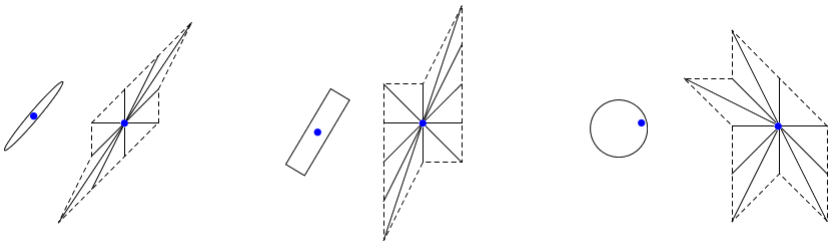

Figure 18: Examples of unit balls for 2D anisotropic gauges, with the corresponding acute stencils by the algorithm from [Mir14b] based on the Stern-Brocot tree.

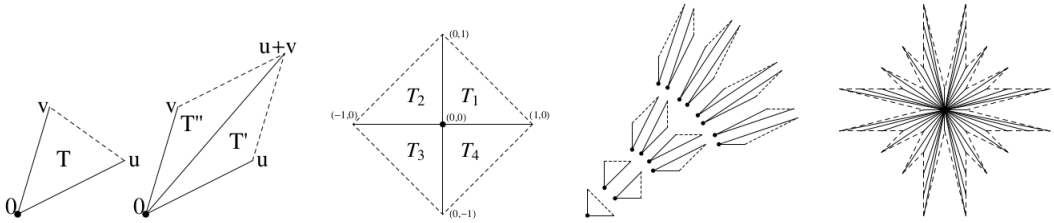

Figure 19: Refinement of a triangle (left). Mesh T_0 (center left). First levels of the binary tree (center right) defined by the recursive refinements of T_1. Mesh defined by a refinement of depth 5 (right).

where $\mu = \max\{\mu(F), 12\}$, R_θ denotes the rotation of angle $\theta \in \mathbb{R}$, and C is an absolute constant.

With $\alpha = \pi/3$ considered in practice, we get $O(\mu \ln(\ln \mu))$ in the worst case, which is essentially linear in anisotropy, and only $O((\ln \mu)^2)$ in the average case.

We explain the refinement process for a stencil. An elementary triangle T is defined as a triangle satisfying the following properties:

- One of the vertices of T is the origin $(0, 0)$, and the other two belong to $\mathbb{Z}^2$.
- Denoting by u, v the non-zero vertices of T, one has $|\det(u, v)| = 1$ and $s(T) := \langle u, v\rangle \geq 0$.

A 2D stencil (with center set to the origin) can be seen as an union of such triangles. The refinement of an elementary triangle T of non-zero vertices u, v consists of the two elementary triangles T' and T'' of non-zero vertices $(u, u+v)$, and $(u+v, v)$, respectively, which are referred to as its children.

Based on this refinement principle, we establish an algorithm to successively refine triangles of a stencil, and we can set a stopping criterion based on the angular size of the triangle defined from $s(T)$. To count the elements of the resulting stencil, we make a one to one correspondance between stencils and *finite sub-forests* of the Stern-Brocot tree, an infinite binary tree labeled with rationals, which we interpret as a subdivision of

the interval $[0, 2\pi]$ into unequal parts whose endpoints have rational tangents. With this angular partition, this choice of subdivision yields an efficient construction of (F, α)-acute stencils with minimal cardinality.

3.3 Eulerian scheme for the eikonal equation

After the "semi-Lagrangian" scheme, we now present another numerical scheme, referred to as the "Eulerian" scheme, which also computes the solution to the eikonal equation. However, while the semi-Lagrangian scheme is based on a discretization of Bellman's optimality principle, the Eulerian scheme is a direct discretization of the eikonal equation, which can lead to totally different numerical methods for anisotropic metrics.

3.3.1 Eulerian scheme

A scheme on a finite space X is an application $\mathfrak{F}$ of the form:

$$\mathfrak{F}(p, U(p), (U(p) - U(q))_{q \in X \setminus \{p\}}),$$

for $p \in X$ and $u : \mathbb{R}^X \to \mathbb{R}^+$. We also use the notation

$$\mathfrak{F}U(p) := \mathfrak{F}(p, U(p), (U(p) - U(q))_{q \in X \setminus \{p\}}).$$

- A scheme $\mathfrak{F}$ is called "monotonous" if $\mathfrak{F}$ is increasing in its second and third variables.
- A scheme $\mathfrak{F}$ is called "causal" if $\mathfrak{F}$ only depends on the positive part of its third variable.

In the Eulerian scheme, the eikonal equation is discretized as

$$\mathfrak{F}U(p) = 0, \tag{31}$$

for all $p \in X$, with boundary conditions $U = 0$ on ∂X. Again, X and ∂X are discretized sets for the approximation of the domain Ω with boundary $\partial\Omega$.

The properties of a scheme $\mathfrak{F}$ can be directly linked with the properties of the operator Λ considered in the semi-Lagrangian scheme as in Section 3.2.1: for $\mathfrak{F}$ a scheme, we can define Λ by:

$$\Lambda U(p) = \lambda \text{ such that } \mathfrak{F}(p, \lambda, (\lambda - U(q))_{q \in X \setminus \{p\}}) = 0.$$

If $\mathfrak{F}$ is monotonous (resp. causal), then Λ is also monotonous (resp. causal). Similarly to the semi-Lagrangian scheme, the computation of the solution to (31) requires $\mathfrak{F}$ to be a monotonous scheme, and the FSM can be used. Besides, if $\mathfrak{F}$ is a causal scheme, then (31) can be solved in a single pass with the FMM.

The Eulerian scheme uses a direct discretization of the eikonal equation, which usually gives a different numerical scheme than the semi-Lagrangian scheme based on a discretization of Bellman's optimality principle.

Examples of monotonous and causal schemes $\mathfrak{F}$ includes all schemes which can be expressed as:

$$\mathfrak{F}U(p) = \underset{\alpha\in A}{\text{mix}} \sum_{\beta\in B} \rho_{\alpha\beta}(p) \max\{0, \frac{U(p) - U(p \pm he_{\alpha\beta})}{h}\} \tag{32}$$

where $\text{mix} \in \{\max, \min\}$, A and B are finite sets, $\rho_{\alpha\beta}(p) \geq 0$ and $e_{\alpha\beta}(p) \in \mathbb{Z}^d$. In that case, the numerical solution to the eikonal equation can be computed from (31) using the FMM.

In the specific case of Riemannian metric, we show how the corresponding scheme can be written in the form respecting the formulation of (32):

- A Riemannian metric in dimension n is locally characterized by a continuous field of symmetric positive definite matrix $M : \Omega \to S_n^{++}$, with the corresponding eikonal equation: $\|\nabla u(x)\|_{D(x)} = 1$, where $D(x) := M^{-1}(x)$.

- Assume that we have the decomposition: $D = \sum_{i=1}^{d} \rho_i e_i e_i^\top$, with $\rho_i > 0$ and $e_i \in \mathbb{Z}^n$. Such a decomposition can be obtained in dimension 3 with the Selling decomposition.

- Then we can consider the numerical scheme on the Cartesian grid with grid size h:

$$\|\nabla u(x)\|_{D(x)}^2 = \sum_{i=1}^{d} \rho_i \max\{0, \frac{u(x) - u(x \pm he_i)}{h}\}^2 + \mathcal{O}(h), \tag{33}$$

  which is of the form required for the Eulerian scheme. The scheme involves stencils with vertices $x \pm he_i$.

However, not all numerical schemes for the eikonal equation can be easily written in the way required for the Eulerian scheme. In the next section, we show how to obtain this formulation in the case of TTI anisotropy.

3.3.2 Eulerian scheme for geophysics

In [DMM22] (shown in Section 8), we use the framework from the Eulerian scheme to solve the eikonal equation in a geophysical setting. In this case, only anisotropy up to TTI complexity can be tackled, but there is no longer any limit based on the strength of anisotropy. In comparison, the previous algorithm presented in Section 3.2.2 and based on the semi-Lagrangian scheme can handle even the most complex form of anisotropy in the context of geophysics, but suffers from a limitation based on the strength on the anisotropy, even for the TTI case. Besides, thanks to the simplicity of the Eulerian structure, a GPU implementation of the numerical scheme has been made, and is presented in more details in Section 3.3.3. The GPU acceleration leads to computation time up to 50 times faster compared with the algorithm based on the semi-Lagrangian method, using a single GPU node.

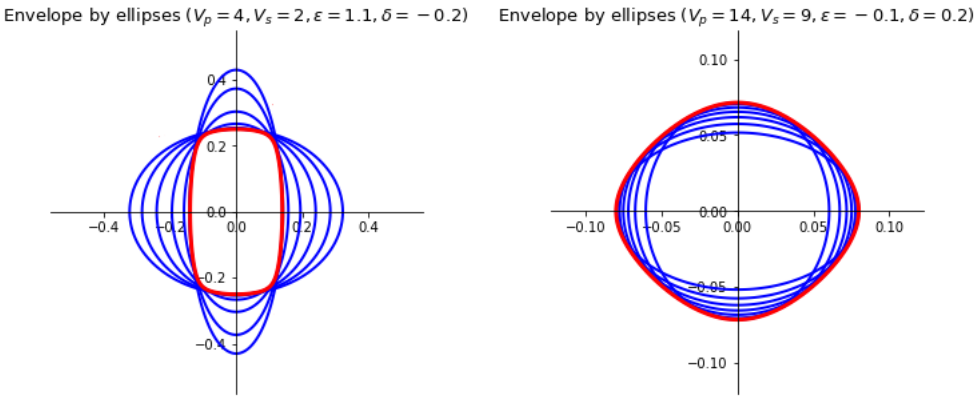

Figure 20: Slowness surfaces (red) defined by (21), in the (p_r, p_z) plane. The coefficients (a, b, c, d, e) are derived from the supplied Hooke parameters $(V_p, V_s, \varepsilon, \delta)$. Only the inner slowness surface is considered, and our numerical method involves its approximation by an intersection of ellipses (left) or a union of ellipses (right), shown blue.

In this algorithm based on the Eulerian scheme, the slowness surface of a TTI metric is represented as an envelope of ellipses, as can be seen in Figure 20, with each ellipse corresponding to a Riemaniann metric. From that, we can write the update operator as a maximum or minimum of update operators from the Riemannian case, which can each be computed efficiently with the framework from the Eulerian numerical scheme presented in Section 3.3.1.

The first arrival traveltime is solution to the P-SV equation presented in (21). We first need to disambiguate the PDE by distinguishing the role of the inner slowness surface, relative to the qP-wave (the fastest). For that purpose, we introduce given coefficients $\sigma = (a, b, c, d, e) \in \mathbb{R}^5$ the quadratic function $\mathcal{Q}_\sigma$ and the set $\mathcal{B}_\sigma$ defined as follows

$$\mathcal{Q}_\sigma(r, z) := ar^2 + bz^2 + crz + dx + ez, \tag{34}$$

$$\mathcal{B}_\sigma := \mathrm{CC}_0\{(p_x, p_y, p_z) \in \mathbb{R}^3; \mathcal{Q}_\sigma(p_x^2 + p_y^2, p_z^2) \leq 1\}. \tag{35}$$

By considering only the connected component of the origin, denoted by CC_0 in (35), we obtain that ∂B_σ is the inner slowness surface defined by $\mathcal{Q}_\sigma$.

The objective is now to write ∂B_σ as an envelope of ellipses. For that, we consider the change of variables $h_r = p_r^2 = p_x^2 + p_y^2$ and $h_z = p_z^2$, and we define the following set, illustrated in Figure 21

$$\mathcal{A}_\sigma := \mathrm{CC}_0\{(h_r, h_z) \in \mathbb{R}^2_+;\ \mathcal{Q}_\sigma(h_r, h_z) \leq 1\}, \tag{36}$$

We show that $\mathcal{A}_\sigma$ is either a union or an intersection of triangular regions: for $\sigma \in \mathbb{R}^5$ admissible, there exists $0 < \alpha_* \leq \alpha^* < 1$ and $\mu \in C^\infty([\alpha_*, \alpha^*],]0, \infty[)$ such that one of the following "max" and "min" cases holds:

(max) μ is convex, and $\mathcal{A}_\sigma = \{(h_r, h_z) \in \mathbb{R}^2_+;\ \forall \alpha \in [\alpha_*, \alpha^*],\ (1-\alpha)h_r + \alpha h_z \leq \mu(\alpha)\}$.

(min) μ is concave, and $\mathcal{A}_\sigma = \{(h_r, h_z) \in \mathbb{R}^2_+;\, \exists \alpha \in [\alpha_*, \alpha^*],\, (1-\alpha)h_r + \alpha h_z \leq \mu(\alpha)\}$.

We deduce that the TTI unit ball (35) can be obtained as a union or an intersection of ellipses, depending on the alternative above and as illustrated in Figure 21: denoting $E(\alpha) := \{(1-\alpha)(p_x^2 + p_y^2) + \alpha p_z^2 \leq \mu\}$

$$(max) : \mathcal{B}_\sigma = \bigcap_{\alpha \in [\alpha_*, \alpha^*]} E(\alpha), \qquad (min) : \mathcal{B}_\sigma = \bigcup_{\alpha \in [\alpha_*, \alpha^*]} E(\alpha) \tag{37}$$

As a consequence, we obtain a new expression of the operator $\mathcal{F}^*_\sigma(R\,\cdot)$ for the TTI eikonal equation, as an extremum of Riemannian norms: let $\sigma \in \mathbb{R}^5$ be admissible, and let $R \in \mathrm{GL}_3(\mathbb{R})$, then for any $p \in \mathbb{R}^3$

$$\mathcal{F}^*_\sigma(Rp) = \underset{\alpha \in [\alpha_*, \alpha^*]}{\mathrm{mix}}\, \mu(\alpha)^{-\frac{1}{2}} \|p\|_{D(\alpha)}, \qquad \text{where } D(\alpha) := R^\top \begin{pmatrix} 1-\alpha & & \\ & 1-\alpha & \\ & & \alpha \end{pmatrix} R, \tag{38}$$

denoting by $\mathrm{mix} \in \{\max, \min\}$ the corresponding case at hand.

In (38), for a fixed $\alpha \in [\alpha_*, \alpha^*]$, we recognize a Riemannian eikonal equation

$$\mathcal{F}^*_\sigma(Rp) = \mu(\alpha)^{-\frac{1}{2}} \|p\|_{D(\alpha)},$$

for which we define a numerical scheme $\mathfrak{F}_\alpha$ from the Eulerian method presented in Section 3.3.

We then have the following scheme $\mathfrak{F}$ corresponding to (38):

$$\mathfrak{F} := \underset{\alpha \in [\alpha_*, \alpha^*]}{\mathrm{mix}}\, \mathfrak{F}_\alpha, \tag{39}$$

where $\mathfrak{F}_\alpha$ denotes the Riemannian scheme shown in (33, right), applied to the matrix $\mu(\alpha)^{-1} D(\alpha)$.

However, solving the maximum or minimum in (39) is non-trivial: it corresponds to a 1D-optimization problem that is neither globally convex nor globally concave. We consider two methods to solve it:

- *Newton-like algorithm:* From studying the stencils required in the numerical scheme based on the Eulerian scheme, we prove that the optimization problem can be divided into a finite number of intervals with at most one optimum in each interval. Therefore, a Newton-like search algorithm is possible in each of these intervals. An illustration is presented in Figure 20, with the red vertical lines indicating the different sections.

- *Exhaustive grid-search:* We can also consider the optimum over k ellipses only. This method is much less precise compared with the previous algorithm, but can be done much faster. Besides, it is more suited to GPU implementation. This is an advantage of the Eulerian framework, as a GPU implementation of the semi-Lagrangian framework would be more difficult due to the sequential nature of the algorithm.

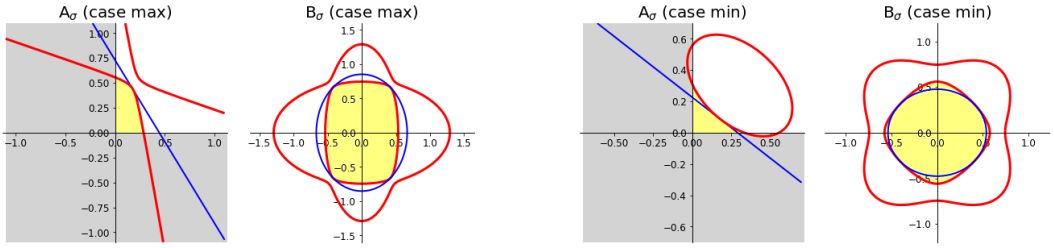

Figure 21: The set $\mathcal{A}_\sigma \subset \mathbb{R}^2_+$ (resp. $\mathcal{B}_\sigma \subset \mathbb{R}^2$ as defined from (p_r, p_z)), in yellow, is bounded by a conic curve (resp. a quartic curve), in red. Tangent lines to $\partial\mathcal{A}_\sigma$ correspond to tangent ellipses to $\partial\mathcal{B}_\sigma$, in blue. If the conic curve defines a convex (resp. concave) boundary, in case (max) see left (resp. case (min) see right) then the ellipses are exterior (resp. interior) tangent.

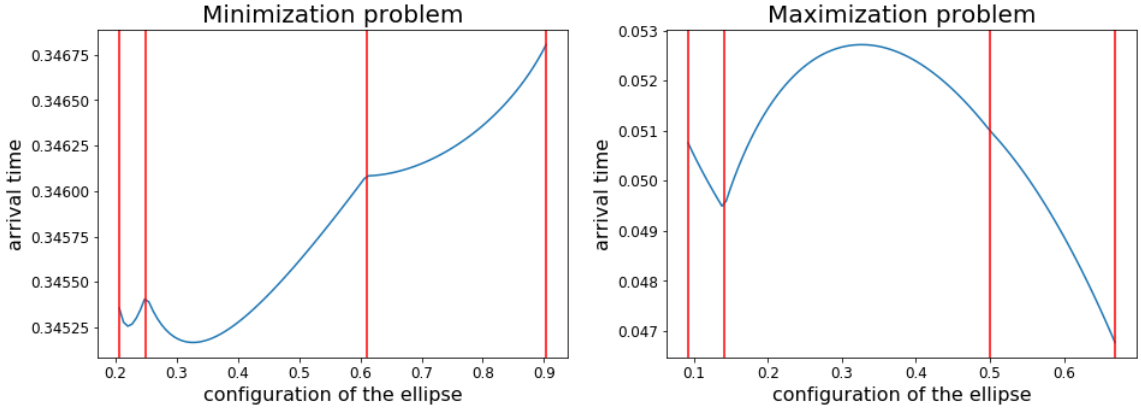

Figure 22: Mapping $\alpha \in [\alpha_*, \alpha^*] \mapsto f(\alpha) := \Lambda^\alpha u(q)$ obtained for some TTI parameters σ, R, a point $q \in h\mathbb{Z}^d$, and an arbitrary mapping $u : \Omega_h \to \mathbb{R}$. The vertical red lines correspond to the abscissas $\alpha_0 \leq \cdots \leq \alpha_K$ at which the stencils in the numerical scheme require a modification of their offsets in the Selling decomposition, here with $K = 3$. Left (resp. right) Figure illustrates case (max) (resp. case (min)), where the function f is quasi-convex (resp. quasi-concave) on each sub-interval $[\alpha_k, \alpha_{k+1}]$, $0 \leq k \leq K$, and must be minimized (resp. maximized).

3.3.3 GPU implementation

In [MGB+21] (shown in Section 9), we present how a numerical solver for the eikonal equation, also based on the Eulerian scheme, can be used with a GPU implementation. Due to the intrinsically sequential nature of the FMM, it is not possible to adapt the FMM for GPU. Instead, we consider iterative algorithms to solve the discretized problem.

We consider a looser version of the propagation front, which does not increase point by point anymore such as the FMM, but block by block instead. An iterative method is used to compute the numerical solution in each block. This algorithm is more suited for parallelization and use of GPU.

In this work, we consider settings that can be solved with the Eulerian method, with in particular the computation of paths globally minimizing an energy involving their curvature. Several models for the curvature penalization have been considered, see Figure 23. The efficiency and robustness of the method is illustrated in various contexts, ranging from motion planning to vessel segmentation and radar configuration, presented in details in Section 3.3.4 corresponding to the article of Section 10. Accelerations by a factor 30 to 120 are obtained when comparing it with a sequential implementation.

The method is presented in Algorithm 2, Algorithm 3. The assignment of a value *val* to a scalar (resp. array) variable *var* is denoted by $var \leftarrow val$ (resp. $var \Leftarrow val$). Illustrations of the procedure are presented in Figure 24. Algorithm 2 is very similar to the FMM, with the difference that the propagation front is made of *blocks* of points, instead of points. The update of each block is presented in Algorithm 3. It uses the Eulerian scheme, for which stencils can be precomputed, as well as the weights and offsets used for the stencils at each point. The update is then done at the block in an iterative way (R iterations), and uses the update operator Λ coming from the Eulerian scheme.

Algorithm 2 Parallel iterative solver (Python)

Variables:

$u : X_h \to [0, \infty]$ (The problem unknown)

$active, next : B_h \to \{0, 1\}$. (Blocks marked for current and next update)

Initialization:

$u \Leftarrow \infty$; $active, next \Leftarrow 0$.

$u[p_*] \leftarrow 0$; $active[b_*] \leftarrow 1$. (Set seed point value, and mark its block for update)

While an *active* block remains:

  For all *active* blocks b in parallel: (CUDA kernel lauch)

    For all $p \in X_h^b$ in parallel: (Block of threads)

      BlockUpdate$(u, next, b, p)$

  $active \Leftarrow next$; $next \Leftarrow 0$.

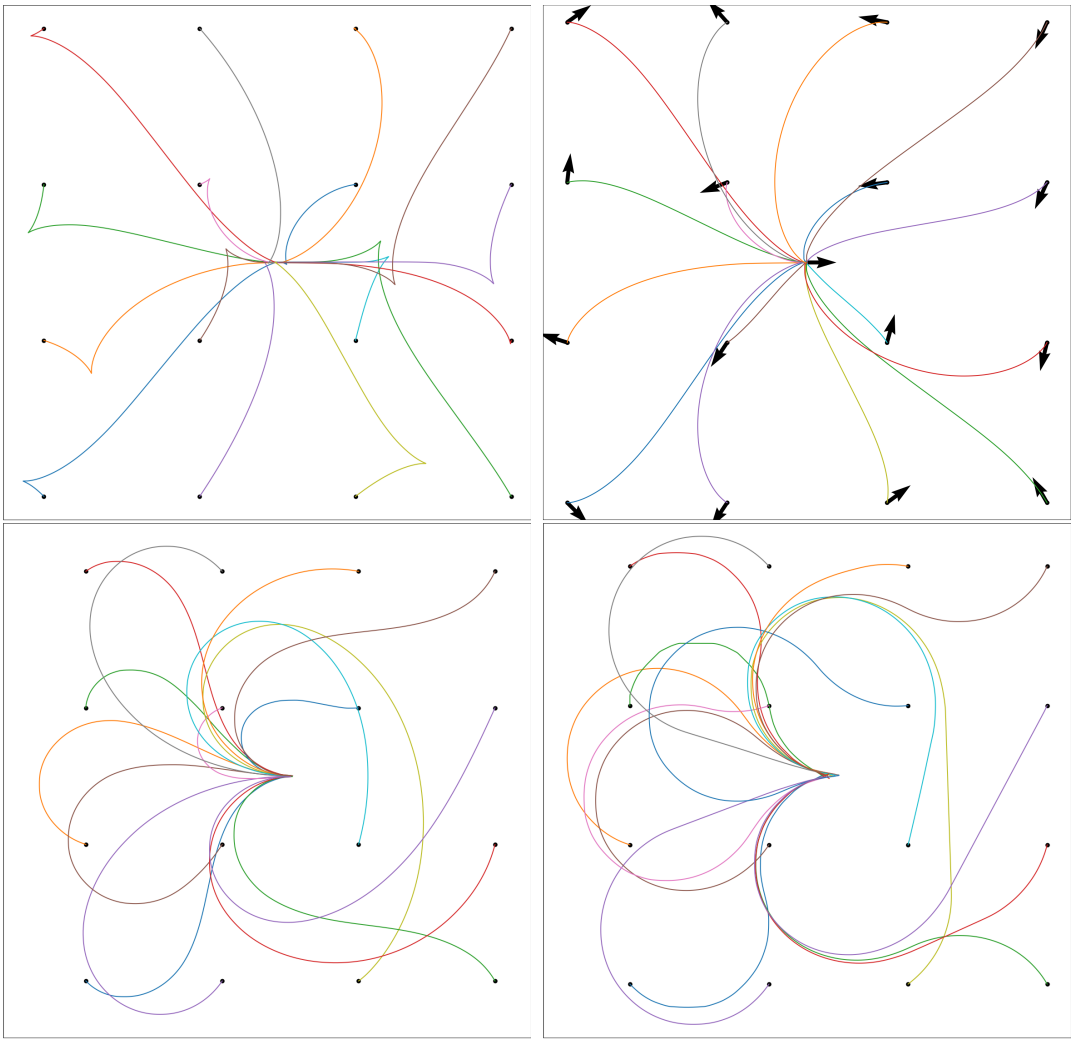

Figure 23: Planar projections of minimal geodesics for the Reeds-Shepp, Reeds-Shepp forward, Elastica and Dubins models (left to right). Seed point $(0, 0)$ with horizontal tangent, regularly spaced tip point with random tangent (but identical for all models).

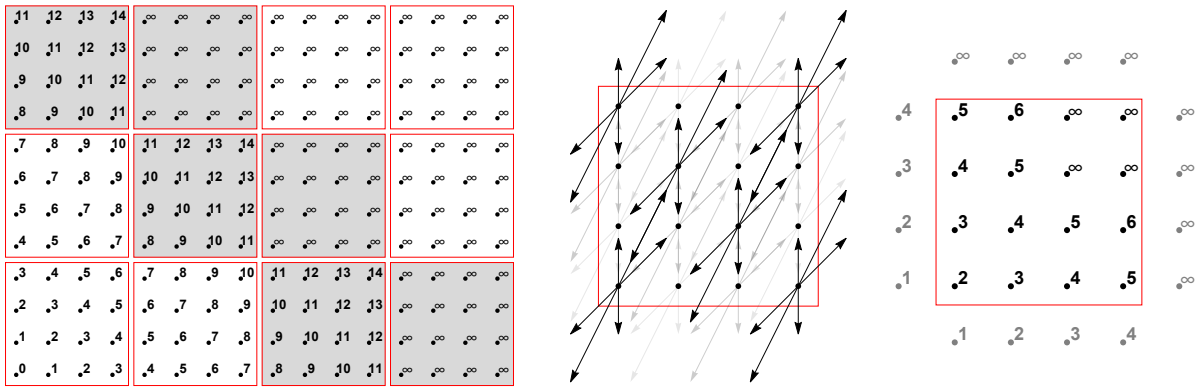

Figure 24: *Left:* Decomposition of the Cartesian grid X_h into tiles X_h^b, with block index $b \in B_h$. Grayed blocks are tagged *active*. *Center:* Updating a block $b \in B_h$ requires loading the unknown values $u : X_h \to \mathbb{R}$, both within X_h^b and at some neighbor points. *Right:* Several local updates are performed within a block (two here).

Algorithm 3 BlockUpdate$(u, next, b, p)$, where $p \in X_h^b$ (CUDA)

Global variables: $u : X_h \to [0, \infty]$, $next : B_h \to \{0, 1\}$, $\rho : X_h \to \mathbb{R}$ (the r.h.s).
Block shared variable: $u_b : X_h^b \to [0, \infty]$.
Thread variables: $\alpha_i \geq 0$, $e_i \in \mathbb{Z}^d$, $u_i \in \mathbb{R}$, for all $1 \leq i \leq I$.

$u_b(p) \leftarrow u(p)$; __syncthreads() (Load main memory values into shared array)
Load or compute the stencil weights $(\alpha_i)_{i=1}^I$ and offsets $(e_i)_{i=1}^I$.
$u_i \leftarrow u(p + he_i)$, for all $1 \leq i \leq I$ such that $p + he_i \notin X_h^b$. (Load the neighbor values)

For r from 1 to R:
    $u_i \leftarrow u_b(p + he_i)$, for all $1 \leq i \leq I$ such that $p + he_i \in X_h^b$. (Load shared values)
    $u_b(p) \leftarrow \Lambda(\rho(p), \alpha_i, u_i, 1 \leq i \leq I)$ (Update $u_b(p)$, unless p is the seed point)
    __syncthreads() (Sync shared values)

$u(p) \leftarrow u_b(p)$ (Export shared array values to main memory)
If appropriate, $next[b] \leftarrow 1$ and/or $next[b'] \leftarrow 1$ for each neighbor block b' of b. (Thread 0 only)

3.3.4 Optimization of the metric in the context of a two-players game

In [DDBM19], presented in Section 10, we provide another application of the FMM, in the context of motion planning: we consider the trajectories of vehicles inside a domain protected by a radar network, where the goal is to optimize the configuration of the radar network. The optimization of the radars is an inverse problem, with the forward problem being the computation of the threatening trajectories with the eikonal solver.

More precisely, the goal is to maximize the probability of detection of the most dangerous trajectory integrated along its path between a given origin and a place to protect. We see it as a non-cooperative zero-sum game: a first player chooses a setting ξ for the radar detection network Ξ, and the other player chooses a trajectory γ from the admissible class Γ assuming full information over the network. The players' objective is respectively to maximize and minimize the path cost:

$$C(\Xi, \Gamma) = \sup_{\xi \in \Xi} \inf_{\gamma \in \Gamma} \Theta_\xi(\gamma) \tag{40}$$

where $\Theta_\xi(\gamma)$ is the detection probability (integrated along the path) of the trajectory γ in the network ξ. Minimization over $\gamma \in \Gamma$ (given $\xi \in \Xi$) is performed using a variant of the FMM. In this work, we rely on the CMA-ES algorithm [VAB+18] for the subsequent optimization over $\xi \in \Xi$, which is rather difficult (non-convex, non-differentiable).

We also consider that the trajectories have a lower bound in their turning radius, due to the vehicle high speed. This leads to a model with curvature penalization, presented in [MD17] and similar to the one considered in Section 3.3.3, and which can be solved by the FMM with Eulerian scheme from [MP19].

For the modeling of the radar, we consider the *ambiguity* map, which accounts for the probability of detection of a generic target by a radar, depending on the distance and the radial speed of the target relatively to the radar, see Figure 25. There are blind speed areas, due to sampling repetition interval and pulse duration causing blind radial distances and blind radial speeds. The positions of the blind areas are periodical and depend on internal parameters of the radar that can be optimized: signal wavelength, and pulse repetition interval.

One interesting behaviour is that in a non-optimized network, when considering all the most threatening trajectories starting from several points in a circle around the target as in Figure (26, left), we observe that all the trajectories concentrate along a single path, a "blind spot" for the network. However, for an optimized network as in Figure (26, right), we see that there are several instances of most threatening trajectories: they all correspond to "local blind spots" as well, but the probability of being detected along these paths is much greater than the most dangerous trajectories from the previous case, which is the goal of this optimization.

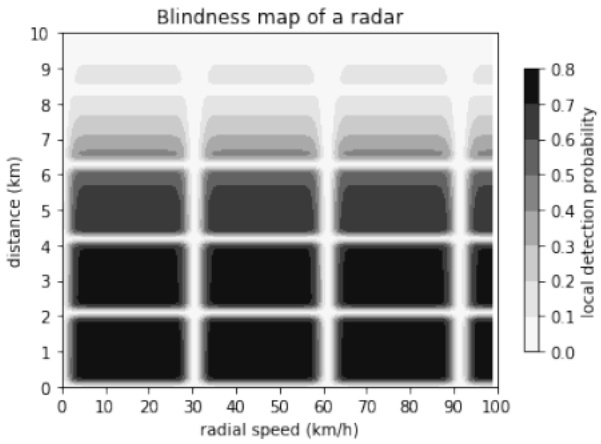

Figure 25: Ambiguity map for a selected waveform waveform

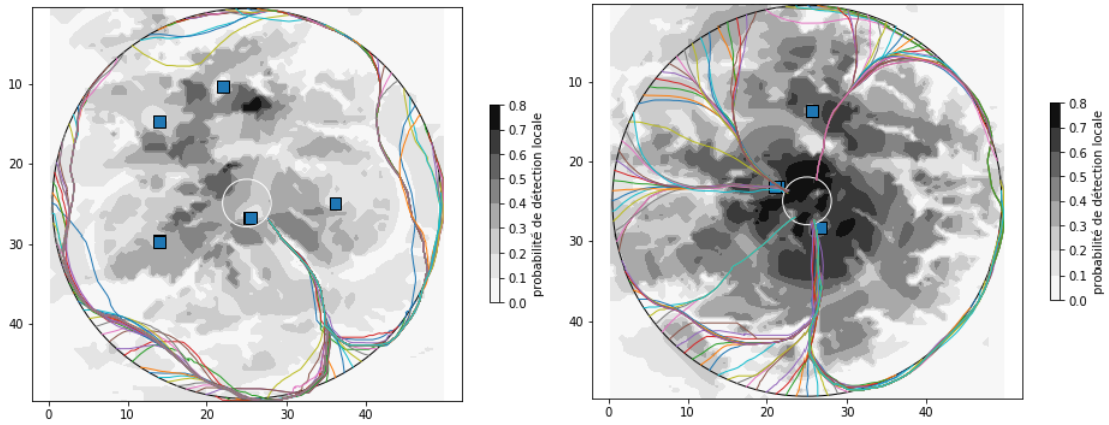

Figure 26: Threatening trajectories in a non-optimized network (left) and an optimized network (right).

4 Conclusion and perspectives

4.1 Conclusion

Taking anisotropy into account is necessary to produce realistic models of the Earth interior, and we present efficient schemes using the FMM to solve the eikonal equation in that context for the first time. Strong and complex anisotropy, however, brings technical challenges for the conception of numerical schemes, and a careful design is needed to satisfy the monotony and causality properties, which are prerequisite of the FMM and ensure the convergence of the method.

We focus on complex forms of anisotropy, with a "semi-Lagrangian scheme" able to tackle triclinic media, the most complex form of anisotropy in the context of geophysics, and an "Eulerian scheme", specifically adapted to TTI media.

The first scheme presented is based on the semi-Lagrangian method, and can tackle

anisotropy coming from a fully general Hooke tensor in 3D media. However, it suffers from a limitation on the strength of the anisotropy, with a known limit depending on the stencil used in the discretization scheme. In the event where this condition fails, the correct solution to our scheme can nevertheless be computed using an iterative method such as the FSM, although we lose advantages of the FMM.

In 2D, this limitation can be removed, and we provide an algorithm to compute strictly acute stencils in the most efficient way. With this strict acuteness, we ensure that the numerical scheme is monotonous and causal, even with small perturbations potentially caused by source factorization or modifications for high-order accuracy. However, this is difficult to generalize such stencils to 3D problems.

We present another numerical scheme, based on the Eulerian scheme, to solve the eikonal equation in the context of geophysics. In this case, only anisotropy up to TTI complexity can be tackled, which is more specific compared with the previous semi-Lagrangian scheme. However, the limitation on the strength of anisotropy is no longer present. Besides, thanks to the simple Eulerian structure, the scheme can also be implemented with GPU acceleration, which makes it more than fifty times faster compared with the semi-Lagrangian scheme, using a single GPU node.

We present the GPU implementation for the Eulerian scheme in details. The FMM is not adequate due to its sequential nature, and we use an iterative solver on blocks, still mimicking the evolution of the propagation front in a looser way.

Last, we consider an application to motion planning, with an algorithm to compute the optimal configuration of a radar network against threatening trajectories, seen as a two-player game between the attacker and the radar network. It provides an example of an inverse problem, as the choice of the optimal radar network corresponds to the optimization of parameters for the metric defining the eikonal equation.

4.2 Perspectives

Several extensions of this work are under consideration.

First, an extension of the Eulerian method to orthorhombic anisotropy seems possible, whereas it can only handle anisotropy up to TTI complexity for now. Indeed, every 2D cross-section of an orthorhombic tensor results in a TTI Hooke tensor. This generalization would involve a two-dimensional minimization, maximization, or min-max saddle point optimization problem at each grid point, instead of the current one-dimensional optimization problem.

It is also possible to use our eikonal solver in applications to inverse problems in seismic imaging. Anisotropy plays a crucial role in the crust and shallow crust, and anisotropic eikonal solvers can be used in the frame of first arrival tomography and stereotomography algorithms [Nol08], to account for anisotropy not only in the modeling but also in the inversion. Besides, the GPU implementation greatly reduces the computation time of the eikonal solver, which can allow to solve several inverse problems simultaneously, and study statistical results such as uncertainty estimations for the inverse problem [TBM19a].

Last, a completely original finite-difference numerical solver for the elastic wave equation is in consideration. It is based on the Selling decomposition of the Hooke tensor,

which breaks down the Hooke tensor into a sum of terms for which a finite-difference approximation can be done, similarly to how the Selling decomposition has been used for matrices defining Riemannian metrics in the Eulerian scheme for the eikonal equation. Hooke tensor with fully general anisotropy can be tackled, with no limitation. Algorithms for the elastic wave equation are already studied extensively, but this numerical solver would be of interest as the first finite-difference scheme to handle general anisotropy for the elastic wave equation.

5 Outline of the PhD thesis

The remainder of the thesis is now divided into two main Parts, related to the "semi-Lagrangian" and "Eulerian" schemes, which are two different methods of computing solutions to the eikonal equation. Each Section within these two Parts corresponds to a publication which has been written and at least published during my PhD training.

- Part II is dedicated to the semi-Lagrangian scheme and its extension to the eikonal equation in geophysics. Within this framework, a numerical solver has been developed to solve the eikonal equation in geophysics with a Hooke's tensor of general anisotropy, and presented in Section 6 [DCC+21]. However, there is a (known) limit based on the strength of the anisotropy, due to the difficulty of obtaining suitable 3D stencils. This limit on the strength of anisotropy does not exist in dimension 2, and a precise description of 2D stencils able to handle any kind of anisotropy is also presented in Section 7 [MD20].

- Part III is dedicated to the Eulerian scheme and its extension to the eikonal equation in geophysics. With that framework, a numerical solver has been developed to solve the eikonal equation in geophysics with a Hooke's tensor, and presented in Section 8 [DMM22]. Only anisotropy up to TTI complexity can be tackled, but there is no limit based on the strength of anisotropy such as in the previous scheme. Besides, a GPU implementation has been done, which leads to a much faster computation time, and we present it in Section 9 [MGB+21]. Finally, this algorithm can be used to solve inverse problems, i.e. the optimization of parameters in the metric, which has been done in a particular use-case related to the configuration of a radar network for the detection of threatening trajectories, and presented in Section 10 [DDBM19].

Part II
Semi-Lagrangian scheme for the eikonal equation

Contents

6 Single pass computation of first seismic wave travel time in three dimensional heterogeneous media with general anisotropy [DCC+21]

This section corresponds to the paper (with minor modifications):

- François Desquilbet, Jian Cao, Paul Cupillard, Ludovic Métivier, and Jean-Marie Mirebeau. Single pass computation of first seismic wave travel time in three dimensional heterogeneous media with general anisotropy. *Journal of Scientific Computing*, 89(1):1–37, 2021

Abstract

We present a numerical method for computing the first arrival travel-times of seismic waves in media defined by a general Hooke tensor, in contrast with previous methods which are limited to a specific subclass of anisotropic media, such as "tilted transversally isotropic" (TTI) media or "tilted orthorhombic" (TOR) media. Our method proceeds in a single pass over the discretized domain, similar to the fast marching method, whereas existing methods for these types of anisotropy require multiple iterations, similar to the fast sweeping method. We introduce a new source factorization model, making it possible to achieve third-order accuracy in smooth media. We also validate our solver by comparing it with the solution of the elastic wave equation in a 3D medium with general anisotropy.

6.1 Introduction

The eikonal equation characterizes the first arrival travel-time of a front, propagating inside a domain at a speed dictated by a given metric. In geophysics, an eikonal equation can be obtained as the high frequency approximation of a wave equation, with the underlying metric defined by the properties of the geological medium.

Computing the solution of the wave equation in three dimensional complex media can be expensive. Indeed, the scale of the discretization grid needs to be substantially smaller than the oscillation wavelength, while the time step is itself bounded by the Courant-Friedrichs-Levy stability condition. In contrast, the eikonal equation is a static (no time dependency) partial differential equation, with a non-oscillatory solution. For this reason, efficient schemes for the eikonal equation have been developed along the years, with several applications in mind: earthquake hypocenter relocation through backpropagation of the data recorded at the surface by seismic stations [MvN92], asymptotic approximation of Green's functions for Kirchhoff migration to build high resolution images in seismic exploration [Bey87, Ble87, LOP+03], or tomographic inversions to determine seismic wave velocities from global and regional scale [Nol08] to exploration and near surface scale targets [BL98, TNCC09] .

However, the metrics from geophysics are often anisotropic, which has been a technical challenge for the numerical solvers designed for the eikonal equation. Anisotropy

can occur from the shape of minerals, with for example the olivine that can be found in the uppermost mantle under oceans [Hes64] and can lead to a preferred direction up to 25% faster than other directions. Besides, thin sedimentary layers of isotropic materials can also be treated as an anisotropic medium in order to smooth strong heterogeneities. It usually leads to a homogenized medium with a faster horizontal speed compared with vertical speed, which is called "vertical transverse isotropy" (VTI) in terms of symmetry in the Hooke tensor. Some shifts can also occur with tectonic movements, leading to the "tilted transverse isotropy" (TTI). More complex types of anisotropy have been considered in the case of fractures, leading to "tilted orthorhombic" (TOR) symmetry or a fully general Hooke tensor.

One option to compute first arrival travel-time is the well-established ray-tracing method [Cer05]. However, several drawbacks have been identified: one ray does not necessarily correspond to the first arrival travel-time, the computation cost increases strongly when many travel paths to many points are needed, and calculations can be difficult in shadow zones which can occur even in a smooth medium. These issues no longer occur when solving the eikonal equation with finite-difference schemes. Note that, conversely, computing second or later arrival travel-times with an eikonal solver is a non-trivial problem [RS04], which is not further discussed in this paper.

The first finite-difference scheme for the eikonal equation has been developed by Vidale [Vid88], in the isotropic case only and with first-order accuracy. It has later been extended to anisotropy [Lec93]. This solver works by induction on the boundary of a square expanding from the source point, but it has no guarantee of success in the case of strong heterogeneities or anisotropy: causality cannot be guaranteed as soon as a ray goes back into the expanding square. This method lacks the robustness and guarantees of an approach based on strong theoretical foundations.

In [OS91], the isotropic eikonal equation is solved by treating it as a dynamic (time-dependent) Hamilton-Jacobi equation, with an "essentially non-oscillatory" (ENO) scheme. This approach has been extended to VTI anisotropy and high order accuracy in [DS97], with the "down & out" (DNO) strategy. A post-treatment (PS) is added in [KC99], with second-order accuracy, resulting in the ENO-DNO-PS scheme, which was extended to TTI anisotropy in [Kim99]. However, the method is computationally expensive. Some other algorithms have been considered to solve the dynamic eikonal equation, but algorithms for the static (time-independent) eikonal equation have been found to be more efficient [LBMV18].

More efficient algorithms have then been developed thanks to the level-set framework [Set96], and the numerical solution of the static eikonal equation as considered in this paper. These numerical methods can be divided into two classes: *iterative* methods and *single pass* methods, which respectively generalize the algorithms of Bellman-Ford and of Dijkstra for graph distance computation. The best known iterative method is presumably the fast sweeping method. Originally introduced in isotropic settings [Zha05], the fast sweeping method has been extended to 2D elliptic anisotropy [TCOZ03], 2D TTI

symmetry [LCZ14], 3D TTI symmetry [PWZ17] with a third-order Lax-Friedrich fast sweeping scheme. It also got extended to 3D TOR symmetry [WYF15] treated as an iterative problem on elliptic anisotropy, and more recently [LBLM] for the 3D TOR symmetry with high order accuracy. Other iterative methods include the adaptive Gauss-Seidel iteration [BR06], or the buffered fast marching method [Cri09], which can both handle some amount of anisotropy. Recently, iterative methods which can take advantage of massively parallel computational architecture, graphics processing units (GPU) in particular, have been proposed in the isotropic settings [JW08], and for elliptic anisotropy [GHZ18].

On the other hand, the best single pass method is presumably the fast marching method (FMM) [Tsi95, Set96], but the extension of the FMM to anisotropic geometries proved more difficult. Early studies [KS98, SV01, AM12] involve wide stencil numerical schemes, leading to increased computation times and reduced accuracy, and therefore negating many of the advantages of the FMM. More recently [Wah20], an algorithm using the FMM has been developed for the 3D TTI anisotropy: it works by solving a fixed point problem on VTI elliptic anisotropy. While the authors illustrate numerically that the algorithm can converge when the considered anisotropy is close from an elliptic VTI anisotropy, there is no formal proof of convergence of the fixed point iteration they implement.

In the past years, one of the authors has proposed extensions of the FMM to 2D anelliptic anisotropy [Mir14b], and 3D elliptic anisotropy [Mir14a, Mir19], as well as various types of degenerate anelliptic anisotropy related with curvature penalization [Mir18]. In these studies, techniques from lattice geometry make it possible to keep the size of the discretization stencil under tight control, thus preventing any loss in computation time and accuracy, even for a very strong anisotropy (with propagation speed ten times faster in the fast direction than in the slow directions). In this paper, we propose a numerical solver using the FMM to solve the eikonal equation with an anelliptic anisotropy defined by a general Hooke tensor. Such a general anisotropy is usually mildly pronounced in absolute terms (with propagation speed at most twice faster in the fast direction compared with the slow directions), but it nevertheless raises a number of specific computational challenges. The method we develop here can be implemented up to third-order accuracy, as illustrated in the numerical experiments in §6.4.

When discussing about the anisotropy of a metric, we make a distinction between two concepts: its "strength" and its "complexity". The strength of the anisotropy refers to the ratio between the highest and lowest achievable speed, depending on the orientation at a given position. The complexity of the anisotropy refers to the number of parameters needed to characterize the metric: for three-dimensional media, 1 parameter is needed for isotropic metrics, 6 parameters for Riemannian metrics (elliptic anisotropy), 8 parameters for TTI metrics, 12 parameters for orthorhombic metrics, and finally 21 parameters for metrics defined by a fully general Hooke tensor. For two-dimensional media, 1 parameter is needed for isotropic metrics, 3 for Riemanian metrics, 5 for TTI metrics and 6 for a fully general Hooke tensor. Our numerical scheme can handle the most complex metrics with all 21 parameters from the Hooke tensor. Such fully general Hooke tensors can arise from

homogenization procedures, see [CC18] and §6.4.2. However, we still have a limitation on the strength of the anisotropy that we can handle: the fast marching method is applicable to our scheme as long as the strength of anisotropy is lower than a given bound, depending on the discretization stencils, see §6.2.3 (in the event where this condition fails, the correct solution to our scheme can nevertheless be computed using an iterative method such as fast sweeping, see Appendix 6.D). Yet, we have verified that we could tackle the strength of anisotropy from most cases found in seismical media, see Table 3 and Figure 28.

Throughout this paper, $\mathbb{R}^d$ denotes the usual Euclidean space, where $d \in \{2, 3\}$ is the ambient dimension. A closed, bounded and connected subset $\Omega \subset \mathbb{R}^d$ is fixed, representing the physical domain. It is equipped with a positive density field $\rho : \Omega \to \mathbb{R}$, as well as a field of Hooke 4th-order tensors $c(x) = (c_{ijkl}(x))$, where $i, j, k, l \in \{1, \cdots, d\}$, describing the elastic properties of the medium, with the usual symmetry assumptions (minor and major symmetries). For any point $x \in \Omega$ and any $p \in \mathbb{R}^d$, regarded as a co-vector, we define

$$m_x(p)_{ik} := \frac{1}{\rho(x)} \sum_{j,l} c_{ijkl}(x) p_j p_l, \qquad N_x^*(p) = \sqrt{\|m_x(p)\|}. \tag{41}$$

Thus $m_x(p)$ is a $d \times d$ symmetric matrix, and $N_x^*(p)$ is the square root of its spectral norm. Note the homogeneity relations $m_x(\lambda p) = \lambda^2 m_x(p)$ and $N_x^*(\lambda p) = |\lambda| N_x^*(p)$ for any $\lambda \in \mathbb{R}$. Unless stated otherwise, summation as in (41) over the indices i, j, k, or l is from 1 to d. We assume that the Hooke tensor $c(x)$ is strictly elliptic, ensuring that $m_x(p)$ is positive definite for all $p \neq 0$ and that N_x^* is a norm on $\mathbb{R}^d$, see Definition 6.8 and Remark 6.12 in §6.3.1.

In this paper, we present an efficient numerical method for computing the unique viscosity solution $u : \Omega \to \mathbb{R}$, see [BCD08], of the generalized eikonal equation

$$N_x^*(\nabla u(x)) = 1, \tag{42}$$

for all $x \in \mathrm{int}(\Omega) \setminus \{x_0\}$, where x_0 is a prescribed source point. This equation is complemented with the boundary condition $u(x_0) = 0$ at the source, and outflow boundary conditions on $\partial\Omega$. One can rewrite (42) under the following classical form [Sla03] which stems from the high frequency analysis of elastic waves

$$\det\Big(\sum_{j,l} c_{ijkl}(x) \partial_j u(x) \partial_l u(x) - \rho(x) \delta_{ik} \Big) = 0, \tag{43}$$

where δ_{ik} denotes Kronecker's symbol. Equation (42) contains the additional information that only the fastest propagation speed is considered. Note that lower propagation speeds formally yield an eikonal equation similar to (42) but involving a non-convex Hamiltonian in general instead of (41, right). Therefore their viscosity solution does not correspond to a travel time of the P-SV modes in the elastic wave equation, but yields non-physical values corresponding to the convex envelope of the Hamiltonian.

We introduce a discretization of the PDE (42), which is solved in a single pass over the domain, using a variant of the FMM. As the algorithm progresses, the successive values $u(x_0) \leq u(x_1) \leq u(x_2) \leq \cdots$ of the numerical solution on Ω_h are computed and then frozen, one by one and in increasing order. The algorithm is *strictly causal*, in the sense that the numerical value $u(x_n)$ computed at a given point only depends on already frozen and strictly smaller values of the solution $u(x_m) < u(x_n)$, $m < n$. This property of the discretized system reflects the deterministic nature of the wave front motion: a present arrival travel-time only depends on the earlier ones, and not on the future ones.

In comparison with iterative methods, such as fast sweeping [TCOZ03], adaptive Gauss-Seidel iterations [BR06], or buffer based methods [Cri09], the FMM used here has a number of appealing properties:

- **Robustness.** The FMM does not require setting any stopping criterion, and is deterministically guaranteed to terminate in a finite number of steps. In addition it is able to tackle general anisotropy associated with a general Hooke tensor $c_{ijkl}(x)$.

- **Speed.** The computational complexity of the FMM is in $\mathcal{O}(N \ln N)$ (where N is the total number of degrees of freedom), *independently of the problem instance*, in contrast with sweeping methods which require a variable number of sweeps depending on the medium complexity (dozens in a complex seismic medium, hundreds in some applications to medical imaging [Mir14a]). A variant of the FMM [Tsi95, RS09] achieves $\mathcal{O}(N)$ complexity, and some level of parallelism, but due to the large hidden constant in the complexity estimate it is rarely used.

- **Accuracy.** Simple enhancements to the FMM allow to formally achieve second and third-order accuracy [Set99], which is confirmed in the numerical experiments presented in §6.4. Such high order schemes are required to estimate the elastic wave amplitudes, or the curvature of the front, as their computations involve second-order spatial derivatives of the arrival travel-times.

- **Differentiability.** The Jacobian matrix of the FMM has a sparse and upper triangular structure, allowing for efficient inversion by direct substitution [MD17].

- **Extensibility.** *Dynamic fast marching* methods modify the numerical scheme online, as the front propagation proceeds, depending on various properties of the minimal paths such as their curvature [LRW13]. In the context of seismic imaging, this flexibility could be used to take into account non-linear effects due to amplitude [VN18].

On the negative side, setting up the FMM for non-isotropic metrics requires substantial work, depending on the geometrical properties of the equation solved [KS98, SV01, AM12, Mir14b, Mir14a, Mir18, Mir19]. In the present state, our numerical method complies to the following restrictions:

- **Parallelism.** The FMM is intrinsically sequential, and thus cannot take advantage of parallel or massively parallel architectures such as [JW08, GHZ18].

- **Stencil construction and size.** The discretization stencils need to obey specific angular properties, depending on the nature and the strength of the anisotropy. Since the overwhelming majority of materials encountered in seismology exhibit rather mild anisotropy in absolute terms, our generalized fast marching method can be usually instantiated with a compact stencil known as the cut-cube, see Figures 27 and 28. However, for crystals such as mica, which are some of the most anisotropic materials encountered in seismology, more extended stencils must be used, see Figure 27 (right), at the possible expense of speed and accuracy. In addition, extending our approach from Cartesian grids to unstructured grids would require substantial effort, in the spirit of [KS98, LFH11].

Outline: An overview of the proposed numerical scheme is presented in §6.2. Implementation details for the critical routines are detailed in §6.3. Numerical experiments presented in §6.4 illustrate the method accuracy and computational efficiency. Finally, we present a conclusion with future perspectives in §6.5.

6.2 The fast marching method

This section describes (a generalization of) the fast marching method [Tsi95], which is used in this paper to solve the generalized eikonal equation (42). The discussion in this section applies to general anelliptic metrics, see Definition 6.6, and the implementation details related with the specific algebraic form (43) of the equation encountered in seismic imaging are postponed to §6.3. The first two subsections §6.2.1 and §6.2.2 introduce classical mathematical tools, that are at the foundation of our approach. The main contributions of this section lie in the angular distortion estimates of §6.2.3, and the choice of source factorization (60) in §6.2.4. A summary of the method is presented in §6.2.5.

The physical domain Ω is discretized on a Cartesian grid of scale $h > 0$,

$$\Omega_h := \Omega \cap h\mathbb{Z}^d, \tag{44}$$

and we assume for simplicity that the source $x_0 \in \Omega_h$.

6.2.1 Geometrical formulation of the eikonal equation.

The generalized eikonal equation (42) is written in terms of a norm N_x^*, at each point $x \in \Omega$, on the space of co-vectors[3]. Following [BCD08] we characterize its unique solution in geometrical terms, involving a norm N_x on vectors, and a distance map between points. For any $v \in \mathbb{R}^d$, regarded as a vector, define

$$N_x(v) := \max\{\langle p, v\rangle;\ p \in \mathbb{R}^d,\ N_x^*(p) \leq 1\}. \tag{45}$$

In the context of seismic imaging, the norm N_x has no closed form expression, but is defined by the above optimization problem in terms of the dual norm N_x^* which is itself

[3]In this paper, the distinction between vectors and co-vectors is kept at an informal level, and we do not distinguish between the spaces $\mathbb{R}^d$ and $(\mathbb{R}^d)^*$

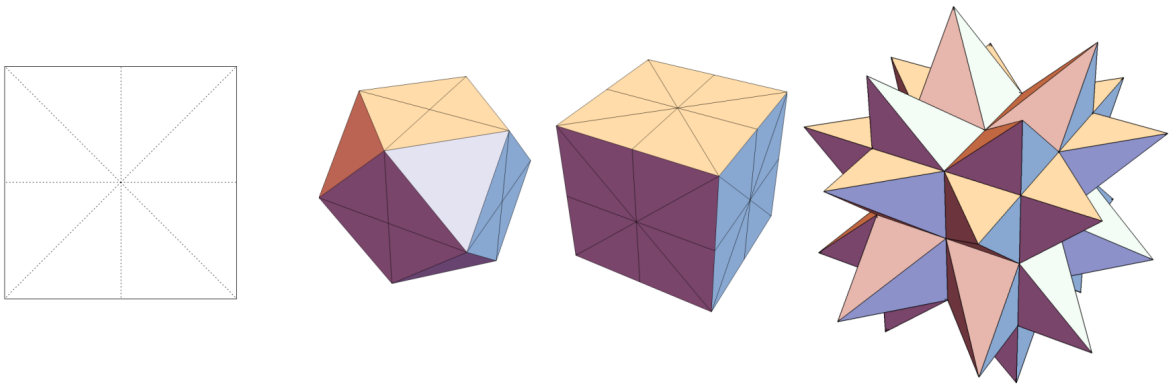

Figure 27: Stencils used, in dimension 2 (square) and 3 (cut-cube, cube and spiky-cube)

the root of a third-order polynomial (41, right). Our numerical scheme depends on the efficient numerical computation of $N_x(v)$ and of its gradient, which is discussed in §6.3.

Denote by $\Gamma := \mathrm{Lip}([0,1], \overline{\Omega})$ the set of all paths within the domain closure, with locally Lipschitz regularity. The length of a path $\gamma \in \Gamma$, and the distance between two points $x, y \in \overline{\Omega}$, are defined as

$$\mathcal{L}(\gamma) := \int_0^1 N_{\gamma(t)}(\gamma'(t))\,\mathrm{d}t, \qquad d(x,y) := \min\{\mathcal{L}(\gamma);\ \gamma \in \Gamma, \gamma(0) = x, \gamma(1) = y\}.$$

The unique viscosity solution to the eikonal equation (42) is the distance from the source point [BCD08]

$$u(x) = d(x_0, x). \tag{46}$$

From this characterization, one can derive Bellman's optimality principle, stating that for any $x \in \Omega \setminus \{x_0\}$, and any neighborhood $\mathcal{V} \subset \Omega$ of x not containing the source point x_0, one has

$$u(x) = \min_{y \in \partial\mathcal{V}} u(y) + d(y, x). \tag{47}$$

6.2.2 Discretization of the eikonal equation

Following [KS98, SV01, BR06, Mir14a, Mir14b], our discretization of the eikonal equation (42) mimics and discretizes Bellman's optimality principle (47). For that purpose, we introduce for each $x \in \Omega_h$ a polygonal neighborhood $\mathcal{V}_h^x$, with vertices in $h\mathbb{Z}^d$ (recall that $\Omega_h := \Omega \cap h\mathbb{Z}^d$), referred to as the stencil, along with a piecewise linear interpolation operator I_h^x on its facets. Given $u : \Omega_h \to \mathbb{R}$, we approximate the right-hand side of (47) by interpolating the arrival travel-times between the vertices of the stencil, and approximating the distance function with the local norm as

$$\Lambda_h u(x) = \min_{y \in \partial\mathcal{V}_h^x} \mathrm{I}_h^x\, u(y) + N_x(x - y). \tag{48}$$

The function u is extended by $+\infty$ on $h\mathbb{Z}^d \setminus \Omega_h$ in (48), which naturally implements outflow boundary conditions on $\partial\Omega$. The numerical computation of (48), which accounts for the bulk of the computation time of our numerical method, is detailed in §6.3.3.

In this paper, the stencil $\mathcal{V}_h^x$ is obtained by rescaling and translating one of the instances shown in Figure 27. This is adequate because the types of anisotropy encountered in seismology are rather mild in absolute terms, even for crystal materials; in contrast, a data adaptive construction is preferred for applications involving much stronger types of anisotropy [Mir14b, Mir14a]. The discretization principle (48) is often referred to as semi-Lagrangian, in contrast with purely Eulerian finite-difference approximations of the eikonal equation such as [Set96, Mir18, Mir19].

The numerical approximation of the arrival travel-time (46) is defined as the unique solution $u : \Omega_h \to \mathbb{R}$ to the discrete system

$$u(x) = \Lambda_h u(x) \tag{49}$$

for all $x \in \Omega_h \setminus \{x_0\}$, and $u(x_0) = 0$ at the source point. Both the eikonal equation and its discretization (48) benefit from comparison principles, see Proposition 6.31 for the latter case. From these, and under mild technical assumptions, one proves that there exists a unique discrete solution u_h to (49) on Ω_h, for each $h > 0$, which converges uniformly as $h \to 0$ to the unique viscosity solution of (42). This approach is standard and not detailed here, see [BR06].

The solution to the discrete non-linear system (49) may be computed using fixed point iterations, see Proposition 6.32, or using any of the iterative methods considered in the introduction, such as the fast sweeping method [TCOZ03]. However, in this study, we are interested in using the fast marching method [Tsi95] (see Algorithm 4), which benefits from a number of advantages listed in the introduction. This requires a careful choice of the stencil $\mathcal{V}_h^x$, as described in the next subsection.

Algorithm 4 The fast marching method (FMM)

Initialize: $u(x_0) = 0$, and $u(x) = +\infty$ for all $x \in \Omega_h \setminus \{x_0\}$. Tag all points as non-accepted.

While a non-accepted point remains:
  Denote by y the non-accepted point minimizing $u(y)$.
  Tag y as accepted. (Optionally, for e.g. higher order methods: PostProcess(y)).
  For each non-accepted point x such that $y \in \mathcal{V}_h^x$:
    $u(x) \leftarrow \tilde{\Lambda} u(x)$ (modified operator using only the values from accepted points).

6.2.3 Acuteness and causality

In this subsection, we establish a property of the numerical scheme, known as *causality*. It can be informally rephrased as follows: the arrival travel-times (i.e. the values of the solution to (49)) smaller than some value τ dictate the arrival travel-times smaller than $\tau + \delta_1$, where δ_1 is a positive constant. An abstract formulation of causality is presented in Proposition 6.30, following [SV01, AM12, Mir14a, Mir14b, Mir19]. We show in Proposition 6.32 that with causality, the system (49) can be solved in finitely many fixed point iterations; see [Mir19, Proposition A.2] for the proof that this system is correctly solved by the FMM (Algorithm 4). In the literature related to fast sweeping methods, causality

is often given a different (non-equivalent) meaning, related to upwindness, stability, and to the monotony property in Proposition 6.30.

Following [SV01], the causality principle we conder here is derived from a geometrical *acuteness* property of the norms and discretization stencils, see Proposition 6.3. Finally, we discuss whether this property holds for various choices of norms and stencils. For that purpose, we introduce the central object of this section, which could be described as the angular width of the facets of a stencil measured with respect to a norm. The unoriented angle $\sphericalangle(u,v) \in [0,\pi]$ between two vectors $u, v \in \mathbb{R}^d \setminus \{0\}$ is defined as

$$\sphericalangle(u,v) := \arccos\big(\frac{\langle u,v\rangle}{\|u\|\|v\|}\big). \tag{50}$$

Definition 6.1. *Let N be a norm on $\mathbb{R}^d$, differentiable on $\mathbb{R}^d \setminus \{0\}$, and let $\mathcal{V}$ be a polygonal neighborhood of the origin. We let*

$$\Theta(N,\mathcal{V}) := \max\{\sphericalangle(\nabla N(v), w);\ v, w \text{ in a common facet of } \partial\mathcal{V}\}.$$

The differentiability assumption in Definition 6.1 is not essential, and could be removed as in [Mir14b]. It is however not restrictive for the application considered in this paper, which does involve differentiable norms, see Lemma 6.15.

The next definition accounts for the distortion of lengths by an anisotropic norm, thus completing Definition 6.1 which is related to the distortion of angles. The length distortion of a norm is also referred to as the strength of its anisotropy, and corresponds to the ratio between the highest and lowest value of the norm with respect to the orientation.

Definition 6.2 (Length distortion)**.** *For any norm N on $\mathbb{R}^d$, define*

$$\mu_*(N) := \min_{v\neq 0} \frac{N(v)}{\|v\|}, \qquad \mu^*(N) := \max_{v\neq 0} \frac{N(v)}{\|v\|}, \qquad \mu(N) := \frac{\mu^*(N)}{\mu_*(N)}.$$

The following proposition, which has numerous variants to be found in [Tsi95, KS98, SV01], governs the applicability of the fast marching method.

Proposition 6.3 (Acuteness implies causality)**.** *Let N and $\mathcal{V}$ be as in Definition 6.1. Let $\mathrm{I}_{\mathcal{V}}$ be the linear interpolation operator on $\partial\mathcal{V}$, and let u be a map defined at the vertices of $\mathcal{V}$. Define*

$$\lambda = \min_{y\in\partial\mathcal{V}} \mathrm{I}_{\mathcal{V}}\, u(y) + N(x-y) \tag{51}$$

and assume that the minimum is attained at a point $y = \alpha_1 y_1 + \cdots + \alpha_d y_d \in \partial\mathcal{V}$, where $y_1, \cdots, y_d$ are the vertices of a common facet of $\mathcal{V}$, and $\alpha_1, \cdots, \alpha_d$ are the barycentric coordinates. If $\Theta(N,\mathcal{V}) \le \pi/2$, then for any $1 \le i \le d$ such that $\alpha_i > 0$, one has

$$\lambda \ge u(y_i) + \|x - y_i\| \mu_*(N) \cos\Theta(N,\mathcal{V}). \tag{52}$$

A proof of Proposition 6.3 is presented in §6.3.3. If $\Theta(N,\mathcal{V}) < \pi/2$, then the update value λ is strictly larger than the neighbor values $u(y_i)$ which play an active role in its computation (in the sense that $\alpha_i > 0$). Adopting the notation of (48), assume that

$$\Theta(N_x, \mathcal{V}_h^x) < \pi/2, \tag{53}$$

for each $x \in \Omega_h$. Then the system (49) is strictly causal, a property that is reformulated in an abstract manner in Proposition 6.30. This makes it possible to apply the FMM. Note that multi-pass iterative methods such as fast sweeping remain applicable even if (53) fails, see Section 6.D for more discussions.

The next definition and proposition bound the angle $\Theta(N, \mathcal{V})$ in terms of the norm and stencil separately. When the strength of the anisotropy in a medium is sufficiently mild, as in the case of geological media, it is possible to select an adequate discretization stencil for the fast marching method, see Tables 3 and 4. In contrast, more pronounced types of anisotropy as considered in [Mir14b, Mir14a, MD20] call for a data-adaptive and anisotropic stencil construction.

Definition 6.4. *Let N and $\mathcal{V}$ be as in Definition 6.1. Define the angular distortion of the norm as*

$$\Theta(N) := \max_{v \neq 0} \measuredangle(\nabla N(v), v), \tag{54}$$

and the angular width of the stencil as

$$\Theta(\mathcal{V}) := \max\{\measuredangle(v, w);\ v, w \text{ in a common facet of } \partial\mathcal{V}\}.$$

In the next proposition we denote by O_d the group of $d \times d$ orthogonal matrices, which are characterized by the identity $R^{-1} = R^{\mathrm{T}}$. Given a norm N and $R \in O_d$, we define a rotated norm by $N \circ R(x) := N(Rx)$ for all $x \in \mathbb{R}^d$.

Proposition 6.5. *Let N and $\mathcal{V}$ be as in Definition 6.1. Then*

$$\Theta(N, \mathcal{V}) \leq \Theta(N) + \Theta(\mathcal{V}). \tag{55}$$

Besides, there exists $R \in O_d$ such that $\Theta(N \circ R, \mathcal{V}) = \Theta(N) + \Theta(\mathcal{V})$.

Proof. Given $u, v \in \mathbb{R}^d \setminus \{0\}$, one has $\measuredangle(\nabla N(u), v) \leq \measuredangle(\nabla N(u), u) + \measuredangle(u, v)$. Thus $\Theta(N, \mathcal{V}) \leq \Theta(N) + \Theta(\mathcal{V})$ by definition. Besides, observing that $\Theta(N \circ R) = \Theta(N)$ for any $R \in O_d$, one obtains the relation: $\Theta(N \circ R, \mathcal{V}) \leq \Theta(N) + \Theta(\mathcal{V})$.

Then, let $u \in \mathbb{R}^d \setminus \{0\}$ be such that $\Theta(N) = \measuredangle(\nabla N(u), u)$, and let v, w in a common facet of $\partial\mathcal{V}$ be such that $\Theta(v) = \measuredangle(v, w)$. Up to replacing N with $N \circ R$, for some $R \in O_d$, we may assume that $u = v$. Up to replacing $\mathcal{V}$ with its image $R'(\mathcal{V})$ by a rotation $R' \in R^d$, we may assume that w lies in the plane generated by u and $\nabla N(u)$, such that $\measuredangle(\nabla N(u), w) = \measuredangle(\nabla N(u), u) + \measuredangle(u, w)$. This shows that $\Theta(N \circ R \circ R'^T, \mathcal{V}) = \Theta(N \circ R, R'(\mathcal{V})) = \Theta(N) + \Theta(\mathcal{V})$, and concludes the proof. □

The angular width $\Theta(\mathcal{V})$ of several stencils is given in Table 3, as well as the angular distortion $\Theta(N)$ of the norm associated with some geological media, numerically computed from their Hooke elasticity tensor as given in [BC91] and a fine sampling of (unit) vectors v in (54). In two dimensions, the square stencil is suitable for all geological media of interest, since $\Theta(N) + \Theta(\mathcal{V}) < \pi/2$. In three dimensions, the cut-cube stencil can be used with olivine and stishovite media, while the more anisotropic mica medium requires the refined spiky-cube stencil, see Figure 27 and Table 4.

The angular distortion $\Theta(N)$ can also be estimated in terms of the length distortion $\mu(N)$ of a norm, as shown in the next proposition. Two estimates are presented: a worst

	Square	Cut-cube	Cube	Spiky-cube
$\Theta(\mathcal{V})$	$\pi/4 = 0.785\ldots$	$\pi/3 = 1.047\ldots$	$\mathrm{acos}(1/\sqrt{3}) = 0.955\ldots$	$\pi/4 = 0.785\ldots$

	Olivine	Stishovite	Mica
$\Theta(N)$	$0.265\ldots$	$0.341\ldots$	$0.753\ldots$

Table 3: Angular width of the stencils illustrated in Figure 27, and angular distortion of the norm associated with several geological media. For completeness, the corresponding Hooke tensors (in GPa units, using Voigt notation) and densities are reproduced below from [BC91].

$$\begin{pmatrix} 323.7 & 66.4 & 71.6 & 0 & 0 & 0 \\ 66.4 & 197.6 & 75.6 & 0 & 0 & 0 \\ 71.6 & 75.6 & 235.1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 64.6 & 0 & 0 \\ 0 & 0 & 0 & 0 & 78.7 & 0 \\ 0 & 0 & 0 & 0 & 0 & 79.0 \end{pmatrix}, \qquad \rho = 3.311\mathrm{g/cm}^3 \text{ (Olivine)}$$

$$\begin{pmatrix} 453 & 211 & 203 & 0 & 0 & 0 \\ 211 & 453 & 203 & 0 & 0 & 0 \\ 203 & 203 & 776 & 0 & 0 & 0 \\ 0 & 0 & 0 & 252 & 0 & 0 \\ 0 & 0 & 0 & 0 & 252 & 0 \\ 0 & 0 & 0 & 0 & 0 & 302 \end{pmatrix}, \qquad \rho = 4.29\mathrm{g/cm}^3 \text{(Stishovite)}$$

$$\begin{pmatrix} 178 & 42.4 & 14.5 & 0 & 0 & 0 \\ 42.4 & 178 & 14.5 & 0 & 0 & 0 \\ 14.5 & 14.5 & 54.9 & 0 & 0 & 0 \\ 0 & 0 & 0 & 12.2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 12.2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 12.2 \end{pmatrix}, \qquad \rho = 2.79\mathrm{g/cm}^3 \text{ (Mica)}$$

	Square	Cut-cube	Cube	Spiky-cube
Elliptic	iff $\mu \leq 1+\sqrt{2}$	iff $\mu \leq \sqrt{3}$	iff $\mu \leq (1+\sqrt{3})/\sqrt{2}$	iff $\mu \leq 1+\sqrt{2}$
Anelliptic	if $\mu \leq \sqrt{2}$	if $\mu \leq 2/\sqrt{3}$	if $\mu \leq \sqrt{3/2}$	if $\mu \leq \sqrt{2}$
Minerals*	✓✓✓	✓✓×	✓✓×	✓✓✓

Table 4: Condition under which a norm N and stencil $\mathcal{V}$ satisfy $\Theta(N)+\Theta(\mathcal{V}) \leq \pi/2$, in terms of the length distortion $\mu = \mu(N)$. Note that the bound is sharp for elliptic norms (if and only if), but only sufficient for anelliptic norms, see Proposition 6.7. Stencils illustrated in Figure 27. Numerical values for the first two lines : $(2.41, 1.73, 1.93, 2.41)$ and $(1.41, 1.15, 1.22, 1.41)$. Minerals*: anisotropy corresponding to the olivine, stishovite and mica minerals respectively.

case estimate in the anelliptic case in §6.B.1, and a sharp estimate in the elliptic case. Note that, strictly speaking, norms defined by a Hooke tensor are anelliptic, but their anellipticity is not always very pronounced.

Definition 6.6 (Norms and elliptic norms)**.** *A norm is a function $N : \mathbb{R}^d \to \mathbb{R}$ such that for all $v, w \in \mathbb{R}^d$, (i) $N(v+w) \leq N(v)+N(w)$, (ii) $N(\lambda v) = \lambda N(v)$ for all $\lambda \geq 0$, (iii) $N(-v) = N(v)$, and (iv) $N(v) \geq 0$ with equality iff $v = 0$. It is said* elliptic *if $N(v) = \sqrt{\langle v, Mv\rangle}$ for all $v \in \mathbb{R}^d$, where $M \in S_d^{++}$ is a positive definite matrix, and* anelliptic *otherwise.*

Proposition 6.7. *For any elliptic (resp. anelliptic) norm N one has*

$$(\mu(N)+\mu(N)^{-1})\cos\Theta(N) = 2 \qquad (\textit{resp.} \quad \mu(N)\cos\Theta(N) \geq 1).$$

The results presented in this subsection also apply if the symmetry assumption (iii) is removed in Definition 6.6. Norms lacking symmetry, often referred to as *quasi*-norms, define quasi-distances which can also be characterized by an eikonal equation and numerically computed using the fast marching method or another iterative method, with straightforward applications of the formalism presented in this paper, see [Mir14b] for two-dimensions applications.

Causality for TTI metrics with the use of fixed stencils. We illustrate the causality property for our stencils on the specific case of TTI anisotropy. However, we keep in mind that our method can similarly handle anisotropy associated with a general Hooke tensor.

The Thomsen parameters [Tho86] are V_p (pressure wave velocity), V_s (shear wave velocity), ϵ and δ, complemented with a rotation R. We define the TTI eikonal equation as

$$ap_r^4 + bp_z^4 + cp_r^2p_z^2 + dp_r^2 + ep_z^2 = 1, \tag{56}$$

where $p_r^2 = p_x^2 + p_y^2$ and $(p_x, p_y, p_z) = R\nabla u$, with parameters (a, b, c, d, e) defined from the

Thomsen parameters as

$$\begin{cases} a &= -(1+2\varepsilon)V_p^2V_s^2, \\ b &= -V_p^2V_s^2, \\ c &= -(1+2\varepsilon)V_p^4 - V_s^4 + (V_p^2 - V_s^2)(V_p^2(1+2\delta) - V_s^2), \\ d &= V_s^2 + (1+2\varepsilon)V_p^2, \\ e &= V_p^2 + V_s^2. \end{cases}$$

The PDE (56) suffers from an ambiguity, similar to (43), in the sense that two propagation speeds are solution, in each direction. In this paper, we only consider the fastest propagation speed, corresponding to pressure waves. From Thomsen parameters and the rotation, one can define a Hooke tensor such that (43) is equivalent to (56).

We use the criterion of Proposition 6.3 to determine whether our fixed 3D stencils provide a causal scheme depending on the Thomsen parameters. With Proposition 6.5, we make our criterion independent from the rotation chosen for the TTI metric. Therefore, we only need to determine the length distortion for a set of Thomsen parameters $(V_p, V_s, \epsilon, \delta)$.

We set $V_s = 0$ for an easier visualization of the results[4]. Besides, we can set $V_p = 1$ with no loss of generality. Results are shown in Figure 28, depending on parameters (ϵ, δ). We also represent the (ϵ, δ) values from 58 examples of media presented in [Tho86]. Out of the 58 media, only 4 of them are such that the cut-cube stencil does not guarantee a causal scheme: these four media are the "Muscovite crystal", "Biotite crystal", "Gypsum-weathered material" and "Aluminum-lucite composite". From these 4 media, only the "Biotite crystal" is such that the spiky-cube stencil does not guarantee a causal scheme. We conclude that the cut-cube stencil suffices to enable the FMM with most practical cases of seismic anisotropy.

6.2.4 Source factorization, and high order finite-differences

We describe enhancements of the numerical scheme (48), aimed at achieving higher accuracy, through an additive factorization of the source singularity, and the use of higher order upwind finite-differences, in the spirit of [LQ12, TH16] and [Set99] respectively. For that purpose, we rely on an infinitesimal variant of Bellman's optimality principle (47):

$$0 = \min_{y \in \partial\mathcal{V}} \langle \nabla u(x), y - x \rangle + N_x(x - y), \tag{57}$$

where $\mathcal{V}$ is a neighborhood of a point $x \in \Omega$. This property can be derived from (47), or alternatively from the eikonal equation (42) and the relation $N_x^*(p) = \max\{\langle p, v\rangle;\ N_x(v) \leq 1\}$ which follows from (45) and Legendre-Fenchel duality. In general, (57) should be understood in the sense of viscosity solutions [BCD08], but for the sake of simplicity we assume in this discussion that u is differentiable at x.

[4]Note that the Thomsen parameter V_s does not exactly correspond to the physical speed of the S-wave, and so V_s does have an impact on the value of the first arrival travel-time, as well as the geometry and anisotropy of the equivalent Hooke tensor. However, on usual values for geophysical media, the impact of V_s is very small so the visualisation of Figure 28 is not significantly altered.

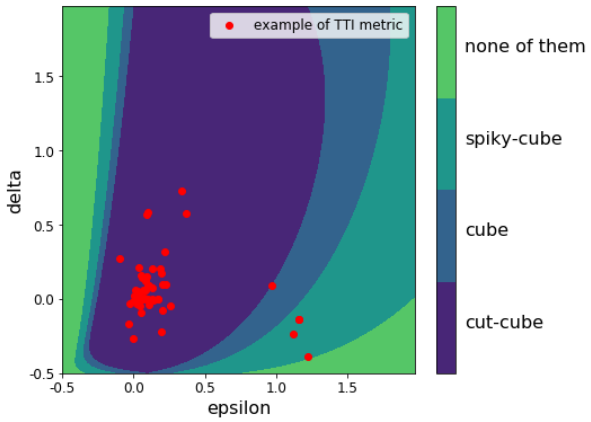

Figure 28: Acuteness property for TTI metrics. The figure shows whether a given stencil is acute with respect to a TTI metric determined by Thomsen parameters (ϵ, δ), with $V_p = 1$ and $V_s = 0$, while taking into account the worst possible rotation of the metric. The domain labelled for the spiky-cube also contains the domain labelled for the cube, which also contains the domain labelled for the cut-cube.

Equation (57) is discretized in a similar fashion as (48) and (49):

$$0 = \min_{y \in \partial \mathcal{V}_h^x} \mathrm{I}_h^x \, D(u, x, y - x) + N_x(x - y), \tag{58}$$

where $D(u, x, y - x)$ is a finite-difference approximation of $\langle \nabla u(x), y - x \rangle$, defined for $x, y \in \Omega_h$. Recall that I_h^x denotes, as in (48), the linear interpolation operator on the polygonal stencil boundary $\partial \mathcal{V}_h^x$, with vertices in Ω_h. For convenience, we also let $D_h^x u(v) := D(u, x, hv)/h$, where $v = (y - x)/h$, which approximates the directional gradient $\langle \nabla u(x), v \rangle$. The original scheme (48) is recovered by letting

$$D_h^x u(v) = \frac{u(x + hv) - u(x)}{h} \qquad \Big(\text{equivalently } D(u, x, y - x) := u(y) - u(x)\Big). \tag{59}$$

Substantial improvements in accuracy can however be obtained using more complex finite-difference schemes which (i) factor the solution singularity at the source point x_0, and (ii) increase the finite-difference order. Note that increased accuracy is suggested by a consistency analysis, and observed numerically §6.4. However, from a theoretical standpoint, we cannot guarantee that these modifications improve numerical accuracy, but only that they do not significantly degrade it, see Section 6.D.

Additive source factorization. The solution to the eikonal equation (42) is non-differentiable at the source point x_0. As a result, the finite-difference approximation of its derivatives, as in (59), is inaccurate for x close to x_0, which degrades the accuracy of the numerical results. Source factorization methods [LQ12] subtract to the unknown u a known function u_* which has same singularity as the solution, but with values and derivatives that can be numerically evaluated to machine precision at a modest cost. The following choices are considered: assuming without loss of generality that $x_0 = 0$

$$u_1(x) := N_0(x), \qquad\qquad u_2(x) := \frac{1}{2}(N_0(x) + N_x(x)). \tag{60}$$

The factor u_2 is considered for the first time in this paper. As illustrated in the numerical experiments §6.4, it is more accurate than u_1, and is required to achieve third-order accuracy.

Following the additive source factorization method [LQ12], and denoting by u_* the chosen factor (u_1 or u_2), we introduce the corrected finite-difference operator

$$D_h^x u(v) = \frac{u(x+hv) - u(x)}{h} + \left(\langle \nabla u_*(x), v \rangle - \frac{u_*(x+hv) - u_*(x)}{h} \right). \tag{61}$$

The resulting modified scheme (58) is only a small perturbation of the original one (49), featuring corrective terms (61, right) of magnitude $\mathcal{O}(h^2/\|x-x_0\|)$. If the original scheme is strictly causal (53), then this perturbation also is, except possibly on a neighborhood of radius $\mathcal{O}(h)$ of the source x_0 where (60) is in any case an excellent approximation of the solution u. As a result, the modified scheme can be solved in a single pass, using the fast marching algorithm, see Proposition 6.33 and the discussion below.

High order finite-differences. High order finite-differences advantageously replace first order ones (59, left) in places where the solution is smooth. These differences should be upwind, i.e. one sided, so as to respect the causality principle underlying the eikonal equation. Second-order upwind finite-differences for instance read

$$D_h^x u(v) = \frac{u(x+hv) - u(x)}{h} - \frac{u(x+2hv) - 2u(x+hv) + u(x)}{2h}, \tag{62}$$

and third-order differences incorporate the additional term

$$\frac{u(x+3hv) - 3u(x+2hv) + 3u(x+hv) - u(x)}{3h}. \tag{63}$$

Note that this approach requires a Cartesian grid discretization, such that $x+2hv$ and $x+3hv$ are points of the discretization grid $h\mathbb{Z}^d$, in addition to x and $x+hv$. Alternative approaches to high order finite-differences also exist, see e.g. [BLZ10]. Source factorization can easily be combined with high order finite-differences: similarly to (61), symbolically compute the directional gradient $\langle \nabla u_*(x), v \rangle$ of the factor (60), and remove its finite-difference approximation $D_h^x u_*(v)$.

In contrast with source factorization, the use of high order finite-differences breaks the fundamental result of Proposition 6.3 on causality, as well as discrete comparison principles (the modified scheme is not monotone). For this reason, they should be used with caution. In practice, following [Set99], we introduce them in the post processing step of the fast marching method, see Line 3. of Algorithm 4. At this stage the numerical scheme is re-evaluated at the accepted point y using high order finite-differences, provided they (i) only involve accepted values, and (ii) are close enough to the standard first-order differences. This avoids introducing instabilities in the FMM, and guarantees that the accuracy is not worse than the first-order scheme, see Proposition 6.34 and the discussion above.

6.2.5 Summary of the numerical method

This paper defines a numerical method, designed to solve the eikonal equation in a geological medium defined with a fully general Hooke tensor (21 parameters). In this subsection, we present a summary of each of its steps. The computer code which implements the method in this study can also be found at: `https://github.com/Mirebeau`, with illustrative Python Notebooks.

Choice of discretization stencils. In order to make it possible to use the FMM, the stencils need to be acute with respect to the metric, see (53). The results presented in this paper make it possible to infer the choice of stencils from possibly known geometric properties of the metric: the angular distortion in Proposition 6.5, and the length distortion in Proposition 6.7. More specific examples are also considered: materials described in terms of Thomsen parameters in Figure 28, and a selection of minerals in Table 4. Typically, the cut-cube stencil is acute for most metrics from geological media.

Numerical solver. Once acute stencils are set, the discrete system (49) can be solved in a single pass by the FMM. The implementation of the FMM is detailed in Algorithm 4 (alternatively, if the stencils are not acute, an iterative method such as fast sweeping may be used, see Section 6.D). For better accuracy, we also use a source factorization, as well as high order finite-differences, see Section 6.2.4.

Computation of the update operator. The FMM requires the computation of the update operator Λ from Definition 48, which is used to compute the arrival travel-time at any position as a minimization problem over arrival travel-times estimated on the facets of the stencil at this position. The computation of the update operator is the most expensive aspect of our numerical method. The minimum is computed on the vertices, edges and faces of the stencil: details of the implementation can be found in §6.3.4.

6.3 Numerical computation of the norm and update operator

We provide in this section the implementation details for the numerical computation of the norm associated with a given Hooke tensor, which is a basic ingredient of our numerical solver of the eikonal equation (42). We also establish Proposition 6.3 in §6.3.3, on causality property of the update operator, and discuss its numerical evaluation in §6.3.4. Throughout this section, we fix a Hooke tensor c, and define $m_c(p) \in S_d$ and $N^*(p)$ for all $p \in \mathbb{R}^d$

$$m_c(p)_{ik} := \sum_{j,l} c_{ijkl}\, p_j\, p_l, \qquad\qquad N_c^*(p) := \sqrt{\|m_c(p)\|}, \tag{64}$$

where the spectral norm (largest eigenvalue) is used in (64, right). This definition is similar to (41), except for the density which is ignored for simplicity, without loss of generality up to rescaling the Hooke tensor. The primal norm N_c is defined as maximization of a linear form subject to a non-linear convex constraint, similar to (45): for all $v \in \mathbb{R}^d$

$$N_c(v) := \max\{\langle p, v\rangle;\ N_c^*(p) \leq 1\}. \tag{65}$$

We rely on sequential quadratically-constrained quadratic programming (SQCQP) to address this problem numerically, see Section 6.C. A variant of the SQCQP also arises in the evaluation of the update operator (48), see §6.3.4. For this method, the constraint needs to take the form "$f \leq 0$" (resp. "$f \geq 0$") (or another constant bound) where the function f is both:

(a) Strongly convex (resp. strongly concave), see Definition 6.11.

(b) Efficiently evaluated numerically, as well as its gradient and Hessian.

The constraint $N_c^* \leq 1$ in (65) does not satisfy any of these properties. Considering $(N_c^*)^2 \leq 1$ fixes (a), see Theorem 6.10, but not (b) since $N_c^*(p)$ involves the spectral norm $\|m_c(p)\|$ which is itself the root of a third-order polynomial. We thus consider alternative expressions of the constraint, and denote

$$B_c^* := \{p \in \mathbb{R}^d; N_c(p) \leq 1\}, \qquad f_c(p) := \det(\mathrm{Id} - m_c(p)). \tag{66}$$

We prove in Proposition 6.14 below, under suitable assumptions, that B_c^* is the connected component of the origin in the set $\{f_c \geq 0\}$.

We can thus replace the highly non-linear constraint $N_c^*(p) \leq 1$ with the constraint $f_c(p) \geq 0$. Since f_c is a polynomial (of $2d$ order in d variables, inhomogeneous), it complies with (b). However, f_c is in general not concave, thus fails (a), see nevertheless Remark 6.18. For this reason we consider yet other alternative expressions of the constraint

$$f_c^{\frac{1}{d}} \geq 0, \tag{67}$$

and, for $\alpha \geq 0$,

$$\exp[-\alpha f_c] \leq 1. \tag{68}$$

The function $f_c^{\frac{1}{d}}$ used in (67) is a barrier function for the set B_c^*: it is strictly positive in its interior, cancels on its boundary, and is strongly concave, see Theorem 6.10. However, it is not defined outside of B_c^*, hence its use is restricted to optimization procedures using only interior points, which is not the case of SQCQP. Even so, it is natural to consider this expression of the constraint before moving to (68), see proof of Theorem 6.10.

On the other hand, the function $\exp[-\alpha f_c]$ is smooth, defined over the whole of $\mathbb{R}^d$, and easy to evaluate, thus complies with (b). The main result of this section, Theorem 6.10, establishes that it is strongly convex on a neighborhood of B_c^*, thus also complies with (a), when α is sufficiently large. See Remark 6.18 on the effective choice of this constant α. This reformulation of the constraint is thus suitable for applying the SQCQP optimization routine to compute $N_c(v)$, see Section 6.C. For numerical stability, the exponential is taken into account in the optimization via Remark 6.29.

In order to state our results, we need to introduce some definitions. Recall that a Hooke tensor is a 4th-order tensor $c = (c_{ijkl})$, $i, j, k, l \in \{1, \cdots, d\}$, obeying the symmetry relations $c_{ijkl} = c_{klij} = c_{jikl}$. Given two symmetric matrices m_1, m_2, we write $m_1 \succeq m_2$ (resp. $m_1 \succ m_2$) if $m_1 - m_2$ is positive semi-definite (resp. positive definite); this is known as the Loewner order.

Definition 6.8. *A Hooke tensor c is said* elliptic *(resp.* strictly elliptic*) iff $m_c(p)$ is positive semi-definite (resp. positive definite) for each $p \in \mathbb{R}^d \setminus \{0\}$.*

We let c_{ell} be the largest constant such that $m_c(p) \succeq c_{\text{ell}} \operatorname{Id} \|p\|^2$ for all $p \in \mathbb{R}^d$, and note that $c_{\text{ell}} \geq 0$ (resp. $c_{\text{ell}} > 0$) if c is elliptic (resp. strictly elliptic).

Definition 6.9. *A Hooke tensor c is said* separable *iff the largest eigenvalue of $m_c(p)$ has multiplicity one for all $p \neq 0$.*

A Hooke tensor is separable, in the sense of Definition 6.9, iff the pressure waves are strictly faster than the other modes of propagation (e.g. P-SV-waves), which is typical of the materials encountered in seismology. The notion of Hooke tensor ellipticity is further discussed in Remark 6.12, and is unrelated with elliptic anisotropy, see Definition 6.6.

Theorem 6.10. *Let c be a strictly elliptic Hooke tensor. Then (i) N_c^* is a norm, and $(N_c^*)^2$ is $2c_{\text{ell}}$-convex, (ii) $f_c^{\frac{1}{d}}$ is $2c_{\text{ell}}$-concave in B_c^*, (iii) if in addition c is separable, and α is large enough, then $\exp[-\alpha f_c]$ is strongly convex in a neighborhood of B_c^**

A property closely related to point (i) is established in [Mus03], with a different proof.

Definition 6.11. *A function f, defined on a convex domain $\Omega \subset \mathbb{R}^d$, is said δ-convex iff for all $p, q \in \Omega$ and all $t \in [0,1]$ one has*

$$f((1-t)p + tq) \leq (1-t)f(p) + tf(q) - \frac{\delta}{2} t(1-t)\|p-q\|^2. \tag{69}$$

A 0-convex function is simply said convex, and a δ-convex function for some $\delta > 0$ is said strongly convex. A function f is said δ-concave iff $-f$ is δ-convex. If f is twice continuously differentiable, then δ-convexity is equivalent to the property $\nabla^2 f \succeq \delta \operatorname{Id}$. If f_1 and f_2 are δ-convex, then $f = \max\{f_1, f_2\}$ also is.

The following variant of the parallelogram identity, which has obvious ties to (69), is used twice in the proof of Theorem 6.10: for any quadratic form A, any points p, q, and any $t \in \mathbb{R}$

$$A((1-t)p + tq) = (1-t)A(p) + tA(q) - t(1-t)A(p-q). \tag{70}$$

Remark 6.12 (Hooke tensor positivity)**.** *Following [BST83], a Hooke tensor is said elliptic (resp. positive) if for all $p, q \in \mathbb{R}^d$ (resp. $m \in S_d$) one has*

$$\sum_{i,j,k,l} c_{ijkl}\, p_i\, q_j\, p_k\, q_l \geq 0 \qquad \Big(\text{resp. } \sum_{i,j,k,l} c_{ijkl}\, m_{ij}\, m_{kl} \geq 0\Big).$$

This notion of ellipticity is clearly equivalent with Definition 6.8. Note also that positivity implies ellipticity, as already observed in [BST83], by choosing $m_{ij} = \frac{1}{2}(p_i q_j + p_j q_i)$.

Remark 6.13 (Gradient of $N_c(v)$)**.** *Let c be a strictly elliptic Hooke tensor, let $v \in \mathbb{R}^d \setminus \{0\}$, and let p be optimal in* (65)*. Then $\nabla N_c(v) = p$ by the envelope theorem [Car01]. This observation allows to numerically implement source factorization, see* (60) *and* (61)*.*

We conclude the introduction of this section with a description of the constraint set B_c^* in terms of the level sets of f_c. Denote by $\mathrm{CC}_x(X)$ the connected component of a point x in a set X.

Proposition 6.14. *If c is elliptic and separable, then $B_c^* = \mathrm{CC}_0\{f_c \geq 0\}$.*

Proof. For all $p \in \mathbb{R}^d$, denote by $\lambda_1(p) \geq \cdots \geq \lambda_d(p)$ the eigenvalues of $m_c(p)$, which depend continuously on p. Note that $\lambda_d(p) \geq 0$ since c is elliptic, and that $\lambda_1(p) > \lambda_2(p)$ for all $p \neq 0$ since c is separable. Note also that $N_c(p) = \sqrt{\lambda_1(p)}$ and $f_c(p) = (1 - \lambda_1(p)) \cdots (1 - \lambda_d(p))$.

Proof of the direct inclusion: if $p \in B_c^*$, then $1 \geq \lambda_1(p)$. Therefore $f_c(\sigma p) = (1 - \sigma^2\lambda_1(p)) \cdots (1 - \sigma^2\lambda_d(p)) \geq 0$ for all $\sigma \in [0, 1]$, thus $p \in \mathrm{CC}_0\{f_c \geq 0\}$ as announced.

Proof of the reverse inclusion: the sets $B_c^* = \{\lambda_1 \leq 1\}$ and $E := \{\lambda_2 \geq 1\}$ are closed, and disjoint by separability. Since $\{f_c \geq 0\} \subset B_c^* \sqcup E$, any connected component of $\{f_c \geq 0\}$ is entirely contained in either B_c^* or E. It follows that $\mathrm{CC}_0\{f_c \geq 0\} \subset B_c^*$ which concludes. □

6.3.1 Convexity and smoothness of the dual norm N_c^*

We establish point (i) of Theorem 6.10, and also discuss the smoothness and uniform convexity properties of N_c^*. Let c be an elliptic Hooke tensor. Then for any $p, q \in \mathbb{R}^d$ one has,

$$\|q\|_{m_c(p)}^2 = \sum_{i,j,k,l} c_{ijkl}\, q_i\, p_j\, q_k\, p_l = \|p\|_{m_c(q)}^2,$$

where $\|q\|_m := \sqrt{\langle q, mq\rangle}$. Therefore

$$N_c^*(p) := \sqrt{\|m_c(p)\|} = \max_{|q|=1} \|q\|_{m_c(p)} = \max_{|q|=1} \|p\|_{m_c(q)}. \tag{71}$$

Proof of point (i) of Theorem 6.10. The function $p \in \mathbb{R}^d \mapsto \|p\|_m$ is convex if m is a symmetric positive semi-definite matrix, and is a norm if m is positive definite. In the latter case, $p \mapsto \|p\|_m^2$ is δ-convex with parameter $\delta = 2\lambda_{\min}(m)$, as follows from the parallelogram identity (70). The announced result follows from (71), the stability of these properties under the max operation, and the observation that $\lambda_{\min}(m_c(q)) \geq c_{\mathrm{ell}}$ if $\|q\| = 1$. □

Lemma 6.15. *Let c be an elliptic and separable Hooke tensor. Then N_c^* is C^∞ smooth on $\mathbb{R}^d \setminus \{0\}$.*

Proof. By construction, $N_c^*(p)^2$ is the largest root of the polynomial $\lambda \mapsto \det(\lambda\,\mathrm{Id} - m_c(p))$, which by assumption is positive and simple (i.e. of multiplicity one). The result immediately follows from the regularity of a polynomial's simple roots with respect to variations in the coefficients. □

The strong convexity of $(N_c^*)^2$ and its C^∞ smoothness on $\mathbb{R}^d \setminus \{0\}$, see Theorem 6.10 and Lemma 6.15, imply the same properties of the norm N_c by Legendre-Fenchel duality.

6.3.2 Convexity of the alternative barriers for the dual unit ball

We conclude in this subsection the proof of Theorem 6.10.

Proof of point (ii) of Theorem 6.10. Each component of $p \mapsto m_c(p)$ is a quadratic form by (64, left), hence by the parallelogram identity (70) one has for any $p, q \in \mathbb{R}^d$ and any $t \in [0,1]$

$$m_c((1-t)p + tq) = (1-t)m_c(p) + tm_c(q) - t(1-t)m_c(p-q). \tag{72}$$

Recall that $\det^{\frac{1}{d}}$ is concave[5] over the cone S_d^+. By homogeneity, this implies super-additivity: $\det(A+B)^{\frac{1}{d}} \geq \det(A)^{\frac{1}{d}} + \det(B)^{\frac{1}{d}}$ for all $A, B \in S_d^+$. Using successively (72), the super-additivity and the concavity of $\det^{\frac{1}{d}}$ on S_d^+, we obtain for any $p, q \in B_c^*$, $t \in [0,1]$, and denoting $M[p] := \mathrm{Id} - m_c(p)$

$$\begin{aligned}
\det(M[(1-t)p+tq])^{\frac{1}{d}} &= \det\big((1-t)M[p] + tM[q] + t(1-t)m_c(p-q)\big)^{\frac{1}{d}} \\
&\geq \det\big((1-t)M[p] + tM[q]\big)^{\frac{1}{d}} + t(1-t)\det\big(m_c(p-q)\big)^{\frac{1}{d}} \\
&\geq (1-t)\det(M[p])^{\frac{1}{d}} + t\det(M[q])^{\frac{1}{d}} + t(1-t)c_{\mathrm{ell}}\|p-q\|^2. \quad \square
\end{aligned}$$

Given a twice continuously differentiable function f, and constants $\alpha, \beta \in \mathbb{R}$, we recall the expression of the composite Hessians,

$$\nabla^2 \exp(-\alpha f) = \big(\alpha \nabla f \nabla f^T - \nabla^2 f\big)\mu_1, \qquad \nabla^2(f^\beta) = \big(\frac{\beta-1}{f}\nabla f \nabla f^T + \nabla^2 f\big)\mu_2, \tag{73}$$

only defined where $f > 0$ for (73, right). We denote $\mu_1 := \alpha \exp(-\alpha f)$ and $\mu_2 := \beta f^{\beta-1}$. Proposition 6.17 below and (73, left) together imply point (iii) of Theorem 6.10. Recall that the comatrix $\mathrm{co}(M)$ has polynomial entries in a matrix M, and satisfies $M^{-1} = \mathrm{co}(M)^{\mathrm{T}}/\det(M)$ when M is invertible.

Lemma 6.16. *Let $M \in S_d$ and $v \in \mathbb{R}^d$. Assume that $\langle w, Mw\rangle \geq c\|w\|^2$ for all $w \in v^\perp$, where $c > 0$. Then $M + \alpha vv^{\mathrm{T}} \succ 0$ iff $\alpha > \alpha_* := -\det(M)/\langle v, \mathrm{co}(M)v\rangle$. (Also, $\langle v, \mathrm{co}(M)v\rangle \geq c^{d-1}$.)*

Proof. Up to a linear change of coordinates, we may assume that $v = (1, 0, \cdots, 0)$. Denote by $\tilde{M} \in S_{d-1}$ the matrix extracted from M by omitting the first line and first column, which by assumption satisfies $\tilde{M} \succeq c\,\mathrm{Id}$. Then $\langle v, \mathrm{co}(M)v\rangle = \det \tilde{M} \geq c^{d-1}$ as announced. Also $\det(M + \alpha vv^T) = \det(M) + \alpha \det(\tilde{M}) = \det(M) + \alpha\langle v, \mathrm{co}(M)v\rangle$ is positive iff $\alpha > \alpha_*$.

Assume for contradiction that there exists a sequence $(w_n)_{n\geq 0}$ such that $\|w_n\| = 1$ and $\langle w_n, (M + nvv^T)w_n\rangle \leq 0$ for all $n \geq 0$. Up to extracting a subsequence, we may assume that $w_n \to w_*$ as $n \to \infty$, where $\|w_*\| = 1$. Noting that $\langle w_n, v\rangle^2 \leq -\langle w_n, Mw_n\rangle/n \to 0$ as $n \to \infty$, we obtain that $\langle w_*, v\rangle = 0$. Then $0 \geq \langle w_n, (M + nvv^T)w_n\rangle \geq \langle w_n, Mw_n\rangle \to \langle w_*, Mw_*\rangle > 0$, as $n \to \infty$, which is a contradiction. We conclude that there exists n_* such that $M + n_* vv^T \succ 0$.

The set $I = \{\alpha \in \mathbb{R};\, M + \alpha vv^T \succ 0\}$ is the connected component of n_* in the set $J = \{\alpha \in \mathbb{R};\, \det(M + \alpha vv^T) > 0\}$. Noting that $J =]\alpha_*, \infty[$, we conclude the proof. $\square$

[5] There are countless proofs of this fact, related to the Brunn-Minkowski theorem. For instance, a reduction shows that one can suppose that one matrix is the identity and the other is diagonal, in which case the inequality follows from the convexity of $t \in \mathbb{R} \mapsto \ln(1 + e^t)$.

Proposition 6.17. *Let c be a strictly elliptic and separable Hooke tensor. Then there exists a constant $\alpha \geq 0$ such that $g(\alpha, p) := \alpha \nabla f_c(p) \nabla f_c(p)^T - \nabla^2 f_c(p)$ is positive definite for all p in a neighborhood of B_c^*.*

Proof. Let $p \in \mathrm{int}(B_c^*)$. By point (ii) of Theorem 6.10 and (73, right) one has $g(\alpha(p), p) \succ 0$ with $\alpha(p) := (1 - 1/d)/f(p)$.

Let $p \in \partial B_c^*$, and let $\lambda_1 \geq \cdots \geq \lambda_d$ be the eigenvalues of $m_c(p)$. One has $\lambda_1 = 1$ since $p \in \partial B_c^*$, and $\lambda_2 < 1$ by separability. Therefore $f_c((1+\varepsilon)p) = \det(\mathrm{Id} - (1+\varepsilon)^2 m_c(p)) = -2\varepsilon(1-\lambda_2)\cdots(1-\lambda_d) + \mathcal{O}(\varepsilon^2)$, which shows that $v := \nabla f_c(p)$ is non-zero. On the other hand, the strong convexity of $(N_c^*)^2$, and the fact that the level sets $N_c^* = 1$ and $f_c = 0$ coincide, implies that $\langle w, \nabla^2 f_c(p) w \rangle < 0$ for all $w \in v^\perp$. From these properties and Lemma 6.16, we obtain that $g(\alpha(p), p) \succ 0$ for sufficiently large $\alpha(p)$.

The announced result follows from the compactness of B_c^*, the continuity of ∇f_c and $\nabla^2 f_c$ and openness of S_d^{++}, and the existence of a suitable $\alpha = \alpha(p)$ at each $p \in B_c^* = \mathrm{int}(B_c^*) \cup \partial B_c^*$ as shown above. □

Remark 6.18 (Effective value of α)**.** *Proposition 6.17 produces the constant α by a compactness argument, which is not quantitative. We numerically approximate this constant, using Lemma 6.16 and a fine sampling of B_c^*, for a variety of materials, and found for example that $\alpha = 40$ is suitable for the Hooke tensor derived from the mica material, and $\alpha = 62$ is suitable for the Hooke tensor derived from the stishovite material (as defined in Table 3).*

Numerical computation of $N_c(v) = \max\{\langle p, v\rangle;\, \exp[-\alpha f_c(p)] \leq 1\}$ is implemented using the SQCQP method, described in Section 6.C and using Remark 6.29 to avoid overflow and underflow error associated with the evaluation of $\exp[-\alpha f_c]$. Our experiments eventually led to the observation that SQCQP is robust and rather insensitive to the parameter α. In particular and to our surprise, no failure of this iterative optimization procedure was observed when letting $\alpha \to 0$, which amounts to applying SQCQP with the constraint $f_c \geq 0$: the point $p = 0$ appears to remain in the basin of attraction of the solution, even though the constraint function is non-concave. Explaining this fortunate behavior is beyond the scope of the current work.

6.3.3 Proof of the causality property

This subsection is devoted to the proof of Proposition 6.3, which relates a geometrical property of the stencils with a causality of the update operator (48) of the fast marching method. For that purpose, let us recall that this operator is defined as a minimization problem over a triangulated surface: the boundary $\partial \mathcal{V}_h^x$ of the discretization stencil, see (48). For each k-dimensional facet of this surface, we thus solve an optimization problem of the following form

$$\lambda = \min_{\xi \in \Xi_k} \langle l, \xi \rangle + N(V\xi), \qquad \text{where} \quad \Xi_k := \{\xi \in [0, \infty[^{k+1};\, \langle \xi, \mathbb{1}_k \rangle = 1\}, \tag{74}$$

where $\mathbb{1}_k := (1, \cdots, 1) \in \mathbb{R}^{k+1}$. We denote by N a norm on $\mathbb{R}^d$, assumed to be differentiable on $\mathbb{R}^d \setminus \{0\}$, and by V a matrix of shape $d \times (k+1)$. Note that the norm N and the set Ξ_k are convex, hence this problem is amenable to numerical optimization, see

§6.3.4. In the context of (48), $N = N_x$, the matrix V collects the vertices of the k-facet of interest of $\partial\mathcal{V}_x^h$, and the vector l collects the values of the unknown u at the vertices of the facet.

Lemma 6.19. *Assume that the minimum (74, left) is attained at a point ξ of the relative interior of Ξ_k. Then denoting $p = \nabla N(V\xi)$, one has*

$$\lambda \mathbb{1}_k = l + V^{\mathrm{T}} p, \qquad\qquad N^*(p) = 1 \tag{75}$$

Proof. Equation (75, right) follows from $p = \nabla N(V\xi)$ and Legendre-Fenchel duality. The Karush-Kuhn-Tucker relations for the optimization problem (74), given that the non-negativity constraints are inactive, yield (75, left) with an arbitrary Lagrange multiplier λ'. The equality of λ' with the value λ of the optimization problem (74) is established as follows:

$$\lambda' = \lambda' \langle \xi, \mathbb{1}_k \rangle = \langle \xi, l + V^{\mathrm{T}} p \rangle = \langle l, \xi \rangle + \langle p, V\xi \rangle = \langle l, \xi \rangle + N(V\xi) = \lambda. \qquad \square$$

The next result links the acuteness of the stencil with the causality of the numerical scheme, following [Tsi95, KS98, SV01]. For that purpose we denote, for each matrix V of shape $d \times (k+1)$

$$\Theta(N, V) := \max_{\xi, \xi' \in \Xi_k} \sphericalangle(\nabla N(V\xi), V\xi'). \tag{76}$$

Proposition 6.20 (Acuteness implies causality, single facet version)**.** *Assume that the minimum (74, left) is attained at a point ξ in the relative interior of Ξ_k, and that $\Theta(N, V) \leq \pi/2$. Denote by $l_0, \cdots, l_k$ the components of l, and by $v_0, \cdots, v_k$ the columns of V. Then for any $0 \leq i \leq k$*

$$\lambda \geq l_i + \|v_i\| \mu_*(N) \cos \Theta(N, V) \tag{77}$$

Proof. Considering (75, left) componentwise, we obtain $\lambda = l_i + \langle v_i, p \rangle$. Since v_i and $V\xi$ belong to the same facet of the stencil, one obtains using the angle condition and Lemma 6.23

$$\begin{aligned} \langle v_i, p \rangle = \langle v_i, \nabla N(V\xi) \rangle &\geq \|v_i\| \|\nabla N(V\xi)\| \cos \Theta(N, V) \\ &\geq \|v_i\| \mu_*(N) \cos \Theta(N, V). \end{aligned} \qquad \square$$

Proof of Proposition 6.3. Up to renumbering the vertices, and eliminating those for which the barycentric coordinate vanishes, we assume that the minimum of (51) is attained at a point $y = \alpha_0 y_0 + \cdots + \alpha_k y_k$, where $\alpha_i > 0$ for all $0 \leq i \leq k$, and $y_0, \cdots, y_k$ are the vertices of minimal facet of $\partial\mathcal{V}$ containing y (the dimension of this facet is k, with $0 \leq k < d$). Denote by V the matrix of columns $y_0 - x, \cdots, y_k - x$, and let $l = (u(y_0), \cdots, u(y_k))$. Then (77) yields (52), since in view of Definition 6.1 one has $\Theta(N, V) \leq \Theta(N, \mathcal{V})$. This concludes the proof. $\square$

6.3.4 Numerical computation of the update operator

The core of our numerical solver of the eikonal equation is devoted to the numerical computation of the update operator (48), defined by the minimization problem

$$\min_{y \in \partial \mathcal{V}_h^x} \mathrm{I}_h^x u(y) + N_x(x-y).$$

Since the stencil boundary $\partial \mathcal{V}_h^x$ is a triangulated surface, see Figure 27, we can minimize over each facet separately. Optimization over a single given facet takes the form (74), which is a mathematically well posed problem: the minimization of a convex functional over a simplex. However, efficient numerical implementation bears importance, as it dominates the computational cost of our method. A first optimization, specific to the FMM, is to consider only the facets of $\mathcal{V}_h^x$ containing the point y that was last accepted, and triggered the update see Algorithm 4. Indeed, the values of u associated with the other vertices of $\mathcal{V}_h^x$ have not changed since the previous update at x.

We discuss here the key ingredients of the implementation of (74), for a norm $N = N_c$ associated with a Hooke tensor c. We focus on the case of a three dimensional stencil ($d = 3$) and distinguish cases depending on the dimensionality k of the sub-facet: a vertex (0-facet), an edge (1-facet), or a face (2-facet) which must be a triangle.

Vertex (k=0). The optimization problem (74) associated with a vertex v_0 is simplified into the trivial expression $\lambda = N_c(v_0) + l_0$. The edge length $N_c(v_0)$ is numerically evaluated as described in the introduction of this section.

Edge (k=1). The optimization problem (74) associated with an edge $[v_0, v_1]$, can be rephrased as the minimization over the interval $[0, 1]$ of the smooth and convex function

$$f(t) := (1-t)l_0 + tl_1 + N_c((1-t)v_0 + tv_1).$$

Our first step is to numerically evaluate $f'(0) = l_1 - l_0 + \langle \nabla N_c(v_0), v_1 - v_0 \rangle$, and likewise $f'(1)$, see Remark 6.13 for the numerical computation of $\nabla N_c(v)$. If $f'(0) \geq 0$ (resp. $f'(1) \leq 0$), then the convex function f reaches its minimum at 0 (resp. at 1), and the problem is solved.

Otherwise, recall that the value to be computed reads

$$\min_{\xi \in \Xi_1} \langle l, \xi \rangle + N_c(V\xi) = \min_{\xi \in \Xi_1} \max_{N_c^*(p) \leq 1} \langle l, \xi \rangle + \langle p, V\xi \rangle.$$

Exchanging the min and max, and using the optimality relation (75, left), we rephrase (74) as

$$\max\{\lambda; (\lambda, p) \in \mathbb{R} \times \mathbb{R}^d,\ \lambda \mathbb{1}_1 = l + V^{\mathrm{T}} p,\ N_c^*(p) \leq 1\}. \tag{78}$$

This problem has the same structure as the primal norm $N_c(v)$ computation, see (65), up to the additional linear equality constraint which raises no particular issue. It is solved using the same approach, namely a reformulation of the contraint as (67, right), and SQCQP, see Appendix 6.C (i.e. we repeatedly solve, in closed form, the approximate problem where the non-linear constraint is replaced with a second-order expansion). For best efficiency, an initial guess for (λ, p) is constructed from the norm gradients at v_0 and v_1, and a quadratic model.

Face (k=2). We turn to the optimization problem (74), posed on a triangle of vertices (v_0, v_1, v_2). The first step is to minimize (74) over the edges $[v_0, v_1]$, $[v_1, v_2]$, and $[v_2, v_0]$ as described in the previous paragraph. Examining the norm of the gradients at these minimizers, one can decide whether the minimum of the convex optimization problem (74) is attained on the boundary of Ξ, in which case the problem is solved.

Otherwise, since V is a square invertible matrix, we can invert (75, left) into $p = V^{-\mathrm{T}}(\lambda 1 - l)$, and turn (75, right) into an univariate polynomial equation of 2d order with respect to $\lambda \in \mathbb{R}$

$$\det(\mathrm{Id} - m_c(V^{-\mathrm{T}}(\lambda 1 - l))) = 0. \tag{79}$$

A Newton method is used to solve this equation, with a suitable initial guess based on the result of the minimization over the three edges $[v_0, v_1]$, $[v_1, v_2]$, $[v_2, v_0]$, and a quadratic model.

6.4 Numerical experiments

In this section, we present numerical experiments to illustrate the properties of the numerical solver introduced in this study. We first perform convergence order and computational complexity analysis. To do so, we make use of particular 3D metrics computed from the conformal transformation of constant metrics. This conformal transformation makes it possible to determine an analytical solution of the eikonal equation in a 3D anisotropic medium presenting heterogeneities (spatial variations of its elastic properties).

In a second experiment, we consider a 3D general anisotropic medium coming from the homogenization of the 3D SEG/EAGE overthrust benchmark model, which is widespread in the seismic exploration community. For this model, no analytical solution to the eikonal equations exists. Therefore, we validate our approach by comparing our computed first-arrival travel-time with the wavefront of the 3D elastic wave equation solution computed in the same medium, using a volumetric method (spectral element strategy).

These two experiments also illustrate the ubiquity and various causes of anisotropy in seismic data. Indeed, the design of the first experiment involves Hooke tensors associated with crystals, with anisotropy originating from the atomic layout at the nanometer scale [BC91]. In contrast, the second experiment illustrates the apparent anisotropy arising from homogeneization at the interfaces of kilometer wide structures [CMA+20]. Let us also acknowledge that a central assumption of homogeneization techniques is that the seismic waves have a limited frequency spectrum, in apparent contradiction with the eikonal equation formalism which is derived from the high frequency approximation; this point deserves investigation in its own right, both theoretical and numerical, but is outside the scope of this paper.

6.4.1 Convergence order and computational complexity

In order to validate the convergence order of the proposed method, we generate a non-trivial test case with an explicit solution, obtained as the conformal transformation of a constant material. We refer to §6.A for details, and simply mention here that the test is parametrized with a (single) Hooke tensor c and a vector $b \in \mathbb{R}^3$. It features a fully

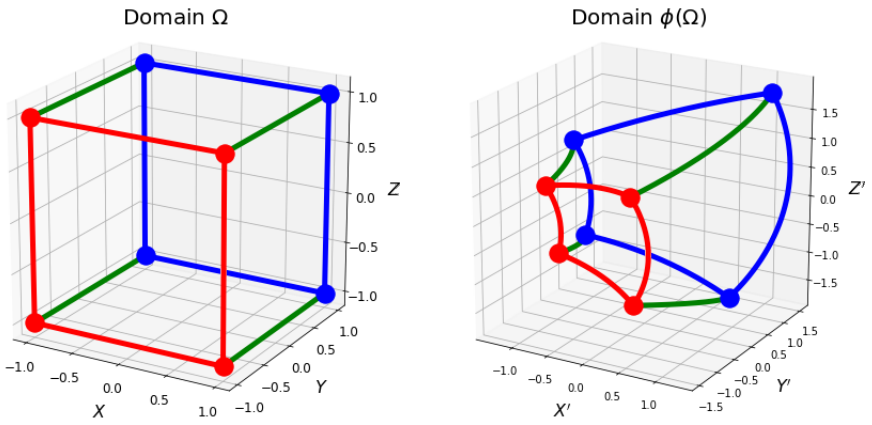

Figure 29: Edges of the domain $\widetilde{\Omega} =]-1,1[^3$ (a cube) and of its image $\Omega = \phi(\widetilde{\Omega})$ by the special conformal transformation (80). See §6.4.1 and §6.A.

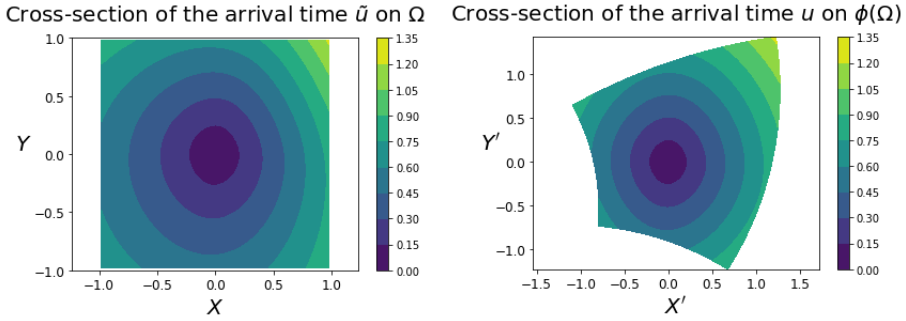

Figure 30: Cross-section at $Z = 0$ of the arrival travel-time for a non-trivial metric on $\widetilde{\Omega}$ (left), which corresponds to a constant metric on the transformed domain $\Omega = \phi(\widetilde{\Omega})$ (right). See §6.4.1 and §6.A.

non-trivial metric on $\widetilde{\Omega} :=]-1,1[^3$, and admits the following explicit solution:

$$\tilde{u}(x) = N_c(\phi(x) - x^*), \qquad \text{with } \phi(x) := \frac{x - b\|x\|^2}{1 - 2\langle b, x\rangle + \|b\|^2\|x\|^2}. \tag{80}$$

For the numerical tests, we consider the Hooke tensors for both the olivine and mica media as defined in Table 3. The olivine medium has orthorhombic symmetry, and an anisotropic length distortion of approximately 0.265. The mica medium has hexagonal symmetry, and an anisotropic length distortion of approximately 0.753. We use our numerical scheme with three different 3D geometrical stencils (cut-cube, cube and spiky-cube), see Figure 27.

We have already shown in §6.2.3 that all three stencils give a causal scheme for the length distortion of the olivine. Indeed we can see in Figure 31 that the L^2-error decreases with the expected order (2 or 3) with the step size of the grid, whereas the computation time is proportional to the total number of grid points.

However, for the mica, only the spiky-cube stencil guarantees a causal scheme. As can be expected, we see in Figure 32 that the cut-cube and cube stencils give poor results

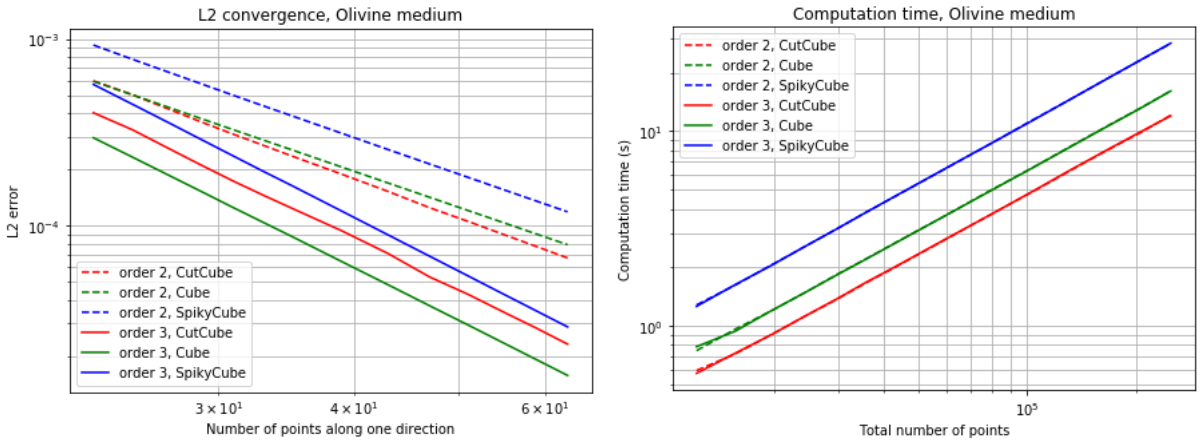

Figure 31: Convergence orders for the olivine, comparison between stencils

here, with a systematic error coming from the scheme being non causal. Conversely, the spiky-cube stencil provides the expected order of convergence. An alternative approach to ensure convergence, not illustrated here, would be to use the non-causal cut-cube or cube stencils in combination with an iterative solver such as the fast sweeping method.

Remark 6.21. *The complexity of the fast marching method is* $\mathcal{O}(C_0 N \ln N + C_1 N)$*, where the first term accounts for the cost of maintaining a priority queue of the non-accepted points, and the second term accounts for the numerical evaluation of the update operator* (48)*, see Lines 2. and 5. respectively in Algorithm 4 in §6.2.2. The structure of the norm involved in the update operator of this study is rather complex since it is defined implicitly, see Equation* (45)*, from an already complex algebraic expression, see Equation* (41)*. As a result, one has* $C_1 \gg C_0$ *and the second contribution* $\mathcal{O}(C_1 N)$ *to the complexity is dominant in our numerical experiments (see Fig 31 and Fig 32), typically accounting for* 90% *of the CPU time. Therefore, the computation times related to the second and third-order schemes appear linear and are very close: most of the CPU time is taken by evaluations of the update operator (similar in both cases) as opposed to e.g. memory accesses (which are slightly more numerous when using third-order finite differences, see Equation* (63)*).*

These numerical experiments on non-trivial 3D metrics confirm that our numerical solver achieves third-order convergence and a quasi-linear computation complexity.

6.4.2 Numerical validation in a 3D fully anisotropic medium

We consider here a 3D model with a fully anisotropic Hooke tensor (21 independent coefficients). This model is obtained through the homogenization (equivalent medium theory) of a fine scale isotropic model, known as the SEG/EAGE overthrust model.

The SEG/EAGE overthrust model is a 3D exploration scale benchmark subsurface model designed in the 1990s to foster the development of wave propagation modeling and inversion tools. It covers an area of 20 km $\times$ 20 km $\times$ 4 km. It represents an onshore structure affected by erosion, in which can be identified faults, a salt layer, and

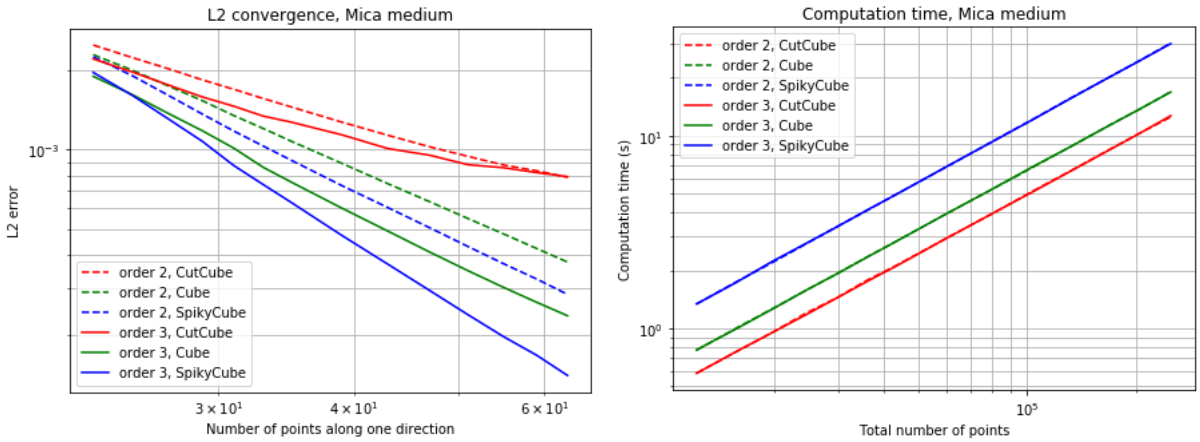

Figure 32: Convergence orders for the mica, comparison between stencils

superficial lateral velocity variations induced by buried topography, channels and lenses. More details can be found in [ABK97].

The initial SEG/EAGE overthrust model is an isotropic model described by pressure and shear wave velocities, and density. Recently, as an illustration of non-periodic two-scale homogenization for elastic media, an upscaled version of the SEG/EAGE overthrust model has been presented [CMA+20]. This branch of homogenization, derived from micro-mechanics [BLP11], aims at computing effective subsurface elastic models for seismic waves propagating at finite-frequency. The leading idea is that subsurface heterogeneities smaller than the propagated wavelengths lead to apparent anisotropy. Effective subsurface media for a given frequency range thus correspond to smooth fully anisotropic media. This theory has now been well established (see [CC18] and references therein). The interest is to reduce the computational cost for volumetric wave propagation method, the computation in a smooth anisotropic medium making it possible to use a coarse Cartesian grid instead of the fine unstructured mesh which would be required in the corresponding isotropic fine scale model. Homogenization also starts to be looked at for better constraining the solution space of seismic imaging problems [CM18].

In this study, we use the homogenized version of the 3D SEG/EAGE overthrust model presented in [CMA+20], therefore a fully anisotropic medium with 21 independent coefficients, and a density model. These models are described on a Cartesian grid containing $534 \times 534 \times 107$ points. This makes it possible for us to access a realistic and physically meaningful fully anisotropic stiffness tensor. To assess and illustrate the accuracy of our fast marching based eikonal solver, we compare the first-arrival travel-times we compute with a 3D wavefront propagating from a source located in the middle of the medium at the surface at $x = 10$ km, $y = 10$ km, $z = 0$ km.

The 3D wave propagation problem is solved using the spectral-element based modeling and inversion code SEM46 [TBM+19b, CBM20]. The simulation is performed using a 10 Hz Ricker vertical force source, on a Cartesian-based mesh using $560 \times 560 \times 110$ elements with P^4 Lagrange polynomial. The final time for simulation is set to 2.5 s leading to 10000

time steps with a time sampling $\Delta t = 0.00025$ s. The computation has been performed on the Univ. Grenoble Alpes HPC Dahu platform on 6 nodes of 32 cores (192 cores in total) benefiting from the domain-decomposition algorithm implemented in SEM46. Each node is equipped with two xeon Skylake Gold Intel processors, each featuring 16 cores clocked at 2.1 GHz, and 192 GO of RAM. The elapsed time for the computation in such settings is approximately 3.5 hours.

For the eikonal solver, the anisotropic length distortion is sufficiently small so that the cut-cube stencil is causal. Compared to the full wave modeling using SEM46, the computation of the eikonal solution on the $534 \times 534 \times 107$ grid on a single core of a laptop (with Intel architecture comparable to the one from the Univ. Grenoble Alpes cluster) took approximately 1600s (less than half an hour).

We present in Figure 33 a 3D view of the superposition of the isochrones for the first-arrival travel time computed through our eikonal solver with the wavefront computed using SEM46, at time $t = 1.5$ s, $t = 2$ s, and $t = 2.5$ s. The P-wave velocity model appears in the background. As can be seen, the isochrone contours (in red) accurately follow the elastic wavefront (in black and white) for the different snapshots. Noticeable irregularities of the wavefront can be identified close to the fastest variations of the P-wave velocity model, which are reproduced accordingly using our eikonal solver. Of course, due to finite-frequency effect of the 3D wave propagation problem, the correspondence is not expected to be perfect, however the qualitative match we observe is a validation of our approach on a realistic 3D example.

We complete this comparison with the presentation of a seismogram in Figure 34. The seismogram is extracted on a receiver line at the surface, in the place of the source, from x=0 to $x = 20$ km. On this seismogram, we superpose the first-arrival travel-time computed through our eikonal solver. Again, we can identify a qualitative match between our eikonal solver solution and the first-break of the synthetic seismogram extracted from the wave propagation simulation.

6.5 Conclusion

We presented a numerical solver for the 3D eikonal equation with anisotropy coming from a general Hooke tensor. It uses a single pass method similar to the fast marching method and features a source factorization, which leads to a quasi-linear complexity and up to third-order accuracy.

The scheme features one parameter, which is the choice of discretization stencil, see Figure 27. For the overwhelming majority of materials encountered in seismology, the compact cut-cube stencil provides best results in terms of both accuracy and computation time. However if strongly anisotropic crystalline materials are considered, such as mica, and if one insists in using the single pass fast marching method (as opposed to e.g. the fast sweeping iterative method) to solve the discretized PDE, then a wider stencil is needed to ensure consistency.

Future research will be devoted to (i) applications to seismic imaging by tomographic inversion, (ii) extensions of the method (to take into account the topography, multiple arrivals, amplitude effects, ...), and (iii) optimizations of the scheme for special classes of

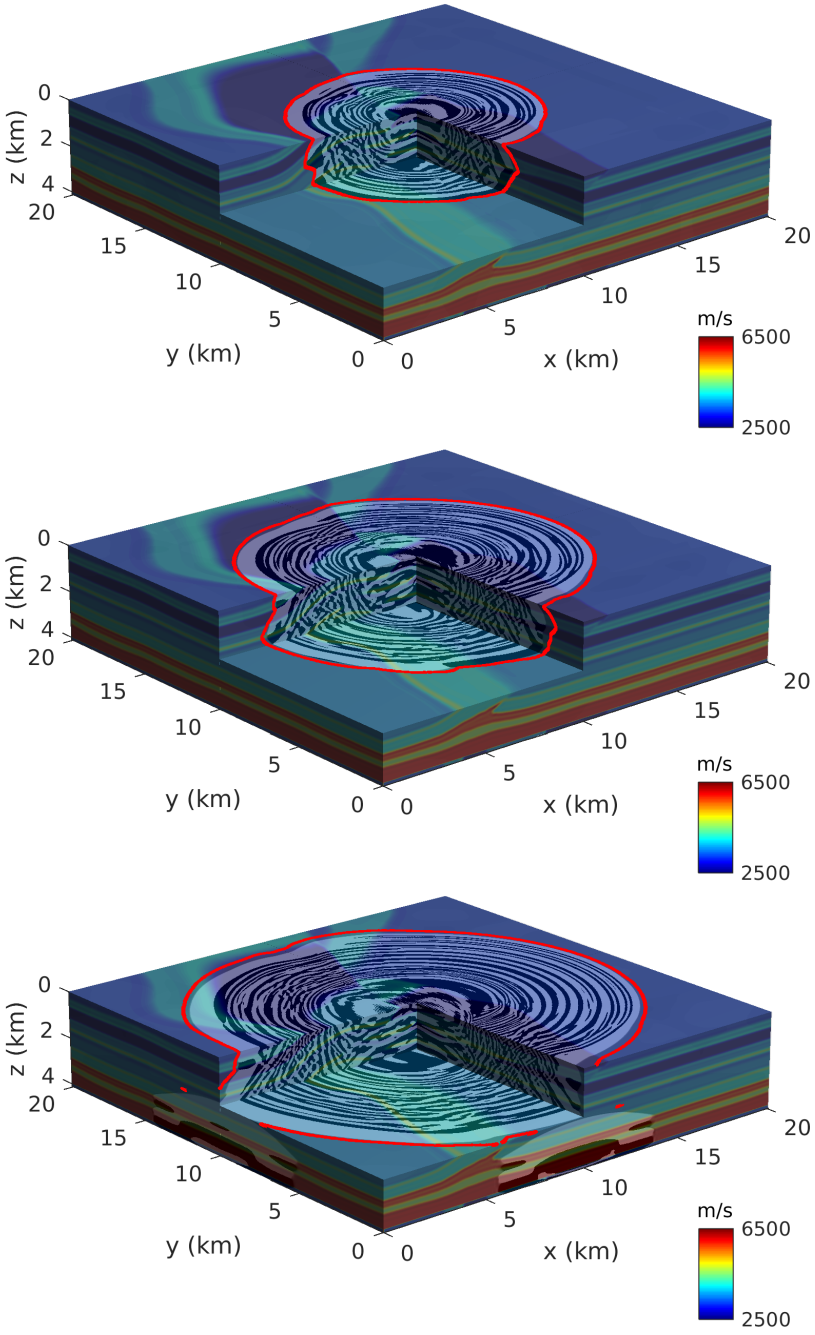

Figure 33: Elastic wavefield (black and white) computed in the 3D fully anisotropic medium coming from the homogenization of the SEG/EAGE overthrust model. The background corresponds to $\sqrt{\frac{C_{33}}{\rho}}$, that is the P-wave velocity of this model if it had a VTI symmetry (which is not the case here, but it still gives a good approximation for illustrative purposes). The red contour corresponds to the isochrone computing through our fast marching eikonal solver. The different snapshots are obtained at $t = 1.5$ s (top), $t = 2$ s (middle) and $t = 2.5$ s (bottom).

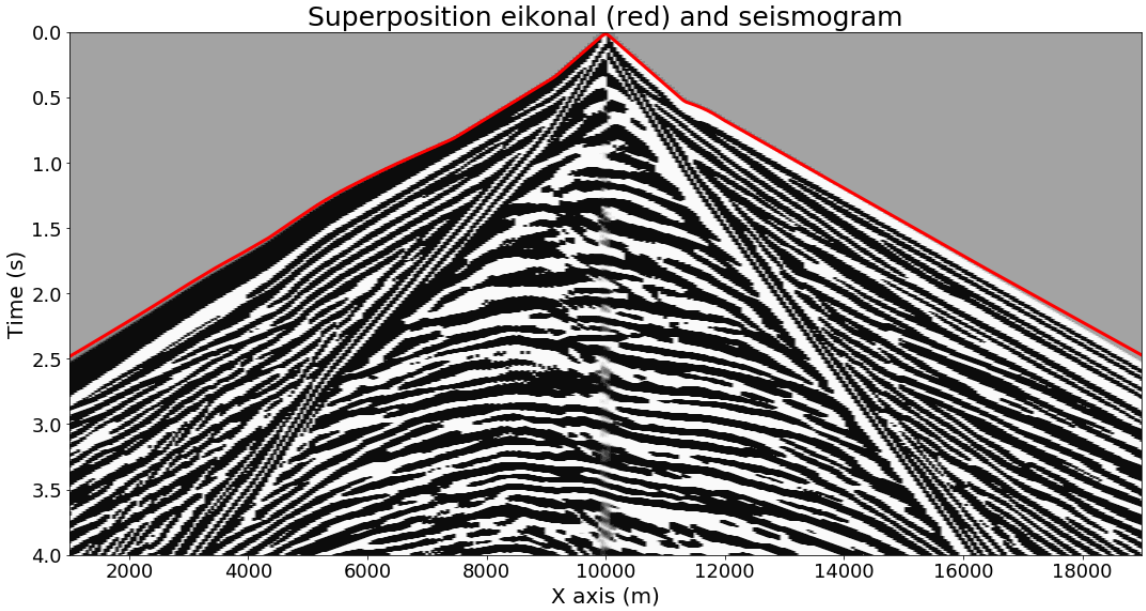

Figure 34: Seismogram recorded along a receiver line located on the surface ($z = 0$ km), in the source plane ($y = 10$ km) along the x-axis (from $x = 0$ km to $x = 20$ km). The vertical displacement is recorded. The vertical displacement intensity is represented in black and white. The first-arrival travel-time computing through our eikonal solver for each receiver position is superposed to the seismogram in red. The resulting red-contour matches the synthetic first-arrival travel-time corresponding to the 3D spectral-element simulation.

Hooke tensors such as tilted transversly isotropic materials.

Acknowledgements This study was partially funded by the SEISCOPE consortium (*http://seiscope2.osug.fr*), sponsored by AKERBP, CGG, CHEVRON, EQUINOR, EXXON-MOBIL, JGI, SHELL, SINOPEC, SISPROBE and TOTAL. This study was granted access to the HPC resources of the Dahu platform of the CIMENT infrastructure (*https://ciment.ujf-grenoble.fr*), which is supported by the Rhône-Alpes region (GRANT CPER07_13 CIRA), the OSUG@2020 labex (reference ANR10 LABX56) and the Equip@Meso project (reference ANR-10-EQPX-29-01) of the programme Investissements d'Avenir supervised by the Agence Nationale pour la Recherche, and the HPC resources of CINES/IDRIS/TGCC under the allocation 046091 made by GENCI.

This research has received funding from the European Union's Horizon 2020 research and innovation programme under the ENERXICO project, grant agreement No. 828947.

Hugo Leclerc helped with the optimization of the numerical implementation of a critical sub-routine, which accounts for the bulk of the computation time of our numerical method, namely the evaluation of $\det(\mathrm{Id} - m_c(v))$ and of its gradient and hessian (67).

Availability of data and material: The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Code availability. A software library which implements the numerical method presented in this study can be found at: `https://github.com/Mirebeau`.

6.A Construction of the synthetic test

We describe the synthetic test used in §6.4.1 to validate the convergence order of the proposed numerical scheme. For that purpose, we need to introduce the following notations.

Definition 6.22. *Let $\Omega \subset \mathbb{R}^d$ be a domain, equipped with a metric $N_x(v)$ (resp. dual-metric $N^*_x(p)$), $x \in \Omega$, $p, v \in \mathbb{R}^d$. Their pull-back by a diffeomorphism $\phi : \widetilde{\Omega} \to \Omega$, with Jacobian matrix denoted by Φ, is defined as*

$$\widetilde{N}_x(v) := N_{\phi(x)}\big(\Phi(x)v\big), \qquad \Big(\textit{resp. } \widetilde{N}^*_x(p) := N^*_{\phi(x)}\big(\Phi^{-T}(x)p\big).\Big)$$

By construction, the geometrical quantities defined in §6.2.1 and associated with the metrics N_x and $\widetilde{N}_x$ are closely related: the path-length $\widetilde{\mathcal{L}}(\gamma) = \mathcal{L}(\gamma \circ \phi)$, and distance $\tilde{d}(x, y) = d(\phi(x), \phi(y))$, where $x, y \in \tilde{\Omega}$ and $\gamma : [0, 1] \to \Omega$. Likewise, if $u : \Omega \to \mathbb{R}$ obeys the eikonal equation (42), then so does $u \circ \phi : \widetilde{\Omega} \to \mathbb{R}$ with respect to the pulled-back dual-metric $\widetilde{N}^*_x$, with the appropriate seed point and boundary conditions. In our experiments, we use for simplicity a metric $N_x = N_c$ defined by a constant field of Hooke tensors c, and a star-shaped domain Ω with respect to the origin, which is chosen as the seed point; the eikonal equation on Ω (resp. $\widetilde{\Omega}$) therefore admits the following explicit solution, as announced in (80):

$$u(x) = N_c(x), \qquad \Big(\text{resp. } \tilde{u}(x) = N_c\big(\phi(x)\big).\Big)$$

In general, the pull-back of a metric defined by a Hooke tensor is *not* defined by a Hooke tensor alone, and one has in addition to keep track of the Jacobian matrix both symbolically and numerically. A special case of interest arises however for *conformal transformations*, for which the Jacobian is a scaled rotation, and which thus preserves the metric structure. More precisely, let $x \in \widetilde{\Omega}$ be fixed, assume that $N^*_{\phi(x)}$ is defined as in (41) by a Hooke tensor c, and that $\Phi(x) = \lambda R$ is the product of a scaling $\lambda > 0$ and of a rotation R. Then $\widetilde{N}^*_x$ is defined by the Hooke tensor of components

$$\tilde{c}_{i'j'k'l'} = \lambda^{-2} \sum_{i,j,k,l} c_{ijkl} R_{ii'} R_{jj'} R_{kk'} R_{ll'}.$$

Another benefit of conformal transformations is that they leave invariant the length distortion and angular width of the metric, $\mu(\widetilde{N}_x) = \mu(N_{\phi(x)})$ and $\mu(\widetilde{N}_x) = \mu(N_{\phi(x)})$, see Definitions 6.2 and 6.4. Three dimensional conformal transformations include dilations, translations, rotations, the inversion $x \in \mathbb{R}^3 \setminus \{0\} \mapsto x/\|x\|^2$, and compositions of these.

In our experiments we use a "special conformal transformation", see (80, right) and Figure 29, which is smooth except for a singularity at $b/\|b\|^2$, where $b \in \mathbb{R}^3$ is a parameter. It is obtained as the composition of an inversion, a translation by $-b$, and another inversion. More precisely, we choose $b = (1/6, 1/9, 1/18)$ and let $\widetilde{\Omega} :=]-1, 1[^3$ with seed at the origin, so that the singular point $b/\|b\|^2 \notin \widetilde{\Omega}$, and the image domain $\Omega := \phi(\tilde{\Omega})$ is star shaped with respect to the origin, see Figure 29. Besides, we use the Hooke tensors of the olivine and mica as defined in Table 3, with a constant rotation of Euler axis $(2, 1, 3)$ and angle $3\pi/5$.

6.B Proof of proposition 6.7

We estimate in this appendix the quantity $\Theta(N)$, which measures the angular distortion associated with a norm N on $\mathbb{R}^d$, in terms of its length distortion, as announced in Proposition 6.7. Different proof techniques are used in the elliptic and anelliptic cases.

6.B.1 Anelliptic norms

The announced estimate, established in Corollary 6.24, follows from upper and lower bounds on the gradient of a norm, presented in the next lemma.

Lemma 6.23. *Let N be a norm on $\mathbb{R}^d$, differentiable at $v \in \mathbb{R}^d \setminus \{0\}$. Then*

$$\mu_*(N) \leq \|\nabla N(v)\| \leq \mu^*(N).$$

Proof. Since N is 1-homogeneous, one has $\langle \nabla N(v), v \rangle = N(v)$ by Euler's identity, and therefore

$$\mu_*(N) \leq \frac{N(v)}{\|v\|} = \frac{\langle \nabla N(v), v \rangle}{\|v\|} \leq \|\nabla N(v)\|.$$

On the other hand, for any vector w, one obtains using successively the convexity of N and the triangular inequality

$$\langle \nabla N(v), w \rangle \leq N(v + w) - N(v) \leq N(w).$$

Choosing $w := \nabla N(v)$ yields the announced upper estimate and concludes the proof:

$$\mu^*(N) \geq \frac{N(\nabla N(v))}{\|\nabla N(v)\|} \geq \frac{\langle \nabla N(v), \nabla N(v)\rangle}{\|\nabla N(v)\|} = \|\nabla N(v)\|. \qquad \square$$

Corollary 6.24. *For any norm N on $\mathbb{R}^d$, differentiable on $\mathbb{R}^d\backslash\{0\}$, one has $\mu(N)\cos\Theta(N) \geq 1$.*

Proof. Using Lemma 6.23 we obtain as announced

$$\cos\Theta(N) = \frac{\langle v, \nabla N(v)\rangle}{\|v\|\|\nabla N(v)\|} = \frac{N(v)}{\|v\|}\frac{1}{\|\nabla N(v)\|} \geq \mu_*(N)/\mu^*(N) = 1/\mu(N). \qquad \square$$

6.B.2 Elliptic norms

The announced estimate, established in Corollary 6.26, follows from a classical inequality in analysis, which proof is recalled in the next lemma.

Lemma 6.25 (Weighted Pólya-Szegö inequality)**.** *Let $(\lambda_i)_{i=1}^d$ be positive numbers, and $(\mu_i)_{i=1}^d$ be non-negative, let $\lambda_* = \min\{\lambda_1, \cdots, \lambda_d\}$ and let $\lambda^* := \max\{\lambda_1, \cdots, \lambda_d\}$. Then*

$$\sqrt{\sum_{1\leq i\leq d} \mu_i\lambda_i}\sqrt{\sum_{1\leq i\leq d} \frac{\mu_i}{\lambda_i}} \leq \frac{1}{2}\left(\sqrt{\frac{\lambda^*}{\lambda_*}} + \sqrt{\frac{\lambda_*}{\lambda^*}}\right)\sum_{1\leq i\leq d}\mu_i.$$

Proof. Without loss of generality, we may assume that $\sum_{1\leq i\leq d}\mu_i = 1$, and denote $E[\gamma] := \sum_{1\leq i\leq d}\mu_i\gamma_i$ for any sequence $(\gamma_i)_{i=1}^d$. Observing that $E[(\lambda^* - \lambda)(1/\lambda_* - 1/\lambda)] \geq 0$, and developping, we obtain

$$\frac{\lambda^*}{\lambda_*} + 1 \geq E\left[\frac{\lambda^*}{\lambda}\right] + E\left[\frac{\lambda}{\lambda_*}\right] \geq 2\sqrt{E\left[\frac{\lambda^*}{\lambda}\right]E\left[\frac{\lambda}{\lambda_*}\right]}.$$

The second inequality follows from the arithmetic-geometric mean inequality $\frac{a+b}{2} \geq \sqrt{ab}$, $\forall a, b \geq 0$. The announced result follows. $\square$

Corollary 6.26. *For any elliptic norm N one has $\frac{1}{2}(\mu(N) + \mu(N)^{-1})\cos\Theta(N) = 1$.*

Proof. Without loss of generality, up to a rotation, we may assume that for all $v \in \mathbb{R}^d$

$$N(v)^2 = \sum_{1\leq i\leq d}\lambda_i v_i^2, \qquad \text{thus} \quad N(v)\nabla N(v) = (\lambda_i v_i)_{i=1}^d,$$

where $\lambda_1, \cdots, \lambda_d > 0$. Denote $\lambda_* = \min\{\lambda_1, \cdots, \lambda_d\}$ and $\lambda_* = \max\{\lambda_1, \cdots, \lambda_d\}$, so that $\mu(N) = \sqrt{\lambda^*/\lambda_*}$. Then letting $\mu_i := \lambda_i v_i^2$ one obtains by Lemma 6.25

$$\frac{\|\nabla N(v)\|\|v\|}{\langle \nabla N(v), v\rangle} = \frac{\sqrt{\sum_i \lambda_i^2 v_i^2}\sqrt{\sum_i v_i^2}}{\sum_i \lambda_i v_i^2} = \frac{\sqrt{\sum_i \mu_i\lambda_i}\sqrt{\sum_i \mu_i/\lambda_i}}{\sum_i \mu_i} \leq \frac{1}{2}\left(\sqrt{\frac{\lambda^*}{\lambda_*}} + \sqrt{\frac{\lambda_*}{\lambda^*}}\right). \tag{81}$$

This proves $\frac{1}{2}(\mu(N) + \mu(N)^{-1})\cos\Theta(N) \geq 1$. Adequately choosing v turns (81) into an equality, which concludes the proof. More precisely, we may assume without loss of generality that $\lambda_* = \lambda_1$ and $\lambda^* = \lambda_2$, and then choose $v = (\sqrt{\lambda_2}, \sqrt{\lambda_1}, 0, \cdots, 0)$. $\square$

6.C Sequential quadratically constrained programming

The numerical implementation of our eikonal equation solver involves the solution to optimization problems of the form

$$\max\{\langle p, v\rangle;\ f(p) \leq 0\}, \tag{82}$$

where f is smooth and strongly convex, and the vector v is fixed. They arise in the definition of the norm (65), which is used in the source factorization (60), as well as in evaluation of the update operator on vertices and edges (78), with an additional linear constraint in the latter case. In order to solve (82), we use an approach known as Sequential Quadratically Constrained Quadratic Programming (SQCQP) [FLT03], which basic principle is to solve a sequence of simplified problems obtained by replacing the objective function and the constraints with their second-order Taylor expansion. We provide two basic results that are sufficient for our application, and refer to [FLT03] for more details on this rich theory. Our first observation, which proof is left to the reader, is that the problem (82) has a closed form solution when f is a suitable quadratic function.

Lemma 6.27 (Maximization of a linear function over an ellipsoid)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a quadratic function such that the set $\{f < 0\}$ is a non-empty ellipsoid, and let $p, v \in \mathbb{R}^d$. Then*

$$F(p) := p + M(p)(\lambda(p)v - V(p)) \tag{83}$$

is the unique solution to (82)*, where*

$$V(p) := \nabla f(p), \quad M(p) := (\nabla^2 f(p))^{-1}, \quad \lambda(p) := \sqrt{\frac{\langle V(p), M(p)V(p)\rangle - 2f(p)}{\langle v, M(p)v\rangle}}. \tag{84}$$

For convenience, the solution (83) is expressed in terms of the Taylor expansion of the quadratic function f at a given but arbitrary point p. Note however that, if f is a quadratic function as assumed in Lemma 6.27, then the matrix $M(p)$ in (84, center) is independent of p, and the value of $F(p)$ is independent of p since it solves (82). The basic SQCQP framework consists in repeatedly evaluating (83) with a *non-quadratic* function f, thus generating a sequence of points $p_{n+1} = F(p_n)$, $n \geq 0$. This yields an iterative method for the optimization problem (82), enjoying a quadratic (Newton-like) local convergence rate, as shown in Proposition 6.28. Variants of this method enjoy a global convergence guarantee [FLT03] under suitable assumptions, but in our numerical experiments the basic method was adequate.

Proposition 6.28. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be C^3 smooth, and let $v \in \mathbb{R}^d$. Assume that $p_* \in \mathbb{R}^d$ and $\lambda_* > 0$ are such that*

$$f(p_*) = 0 \qquad \nabla f(p_*) = \lambda_* v \qquad \nabla^2 f(p_*) \succ 0. \tag{85}$$

Then p_* is an isolated local maximum for the optimization problem (82)*. In addition there exists a constant $C > 0$ such that, for any $p_0 \in \mathbb{R}^d$ close enough to p_*, the sequence defined by $p_{n+1} = F(p_n)$, see* (83)*, satisfies for all $n \geq 0$*

$$\|p_n - p_*\| \leq C^{-1}(C\|p_0 - p_*\|)^{2^n}. \tag{86}$$

Sketch of proof. We recognize in (85) the second-order optimality conditions for the constrained optimization problem (82). A first-order Taylor expansion shows that $\lambda(p_*+h) = \lambda_* + \mathcal{O}(h^2)$, and then $F(p_*+h) = p_* + \mathcal{O}(h^2)$. The estimate (86) follows by induction on $n \geq 0$. □

Remark 6.29 (Exponential transformation, and numerical stability)**.** *Assume that the constraint in* (82) *takes the form $g \leq 0$, where $g = \exp(\alpha f) - 1$ is a strongly convex function, defined in terms of a smooth (but non-convex) f and a positive constant α. One can check that $(g, \nabla g, \nabla^2 g)$ is positively proportional to*

$$\tilde{f} := (1 - \exp(-\alpha f))/\alpha, \qquad \nabla f, \qquad \nabla^2 f + \alpha \nabla f \nabla f^T.$$

Note also that f and $\tilde{f}$ vanish at the same points, and if p_* is such a point then $\tilde{f}(p_*+h) = f(p_*+h)+\mathcal{O}(\|h\|^2)$ for small h. In the sequential quadratic iterations, see Proposition 6.28, one may thus replace $(g, \nabla g, \nabla^2 g)$ with $(f, \nabla f, \nabla^2 f + \alpha \nabla f \nabla f^T)$ and preserve the local quadratic convergence (86)*. This eliminates all exponentials, to the benefit of numerical stability.*

6.D Monotony and causality in fixed point problems

In this section, we review the properties of the numerical scheme considered in this paper, and derive the following guarantees: existence and uniqueness of a fixed point, convergence of an iterative method to find it, and validity of the fast marching method subject to an acuteness condition. We also discuss how these properties transfer to the source factored and high order scheme variants. Closely related arguments can be found in the literature devoted to semi-Lagrangian discretizations of the eikonal equation [Tsi95, SV01, AM12, BR06, Mir14b, Mir14a]. We fix the grid scale $h > 0$ in this appendix, and refer to [BR06] for a convergence analysis to the PDE solution as it is refined. Denote $X := \Omega_h \setminus \{x_*\}$ the discretization set (44) minus the source point, and let $\mathbb{U} := \mathbb{R}^X$ be the set of mappings from X to $\mathbb{R}$. Recall that the objective is to find $u \in \mathbb{U}$ such that for all $x \in X$

$$\Lambda u(x) = u(x), \qquad \text{where } \Lambda u(x) := \min_{y \in \partial \mathcal{V}_h^x} \mathrm{I}_h^x u(y) + N_x(x-y), \tag{87}$$

where $\mathcal{I}_h^x$ denotes the piecewise linear interpolation operator, on a polytope $\mathcal{V}_h^x$ enclosing x with vertices in the grid $h\mathbb{Z}^d$, see Section 6.2.2. By convention in (87, right), $u \in \mathbb{U}$ is extended to $h\mathbb{Z}^d \setminus X$ by $u(x_*) = 0$ and $u = +\infty$ elsewhere. We make the following connectedness assumption: for any $x_0 \in X$, one can find $n \geq 1$ and $x_1, \cdots, x_n \in X$, such that x_{i+1} is a vertex of $\mathcal{V}_h^{x_i}$, for all $i < n$, and x_* is a vertex of $\mathcal{V}_h^{x_n}$. Given $u, v \in \mathbb{U}$, the strict inequality "$u < v$" stands for "$\forall x \in X, u(x) < v(x)$"; and likewise for weak inequality $u \leq v$. Given $u \in \mathbb{U}$ and $\tau \in \mathbb{R}$ we define

$$u^{\leq \tau}(x) := \begin{cases} u(x) & \text{if } u(x) \leq \tau, \\ +\infty & \text{else.} \end{cases}$$

Proposition 6.30. *The operator $\Lambda : \mathbb{U} \to \mathbb{U}$ defined by (87, right) is continuous and obeys the following properties, where δ_0, δ_1 are positive constants, and where $u, v \in \mathbb{U}$, and $s, t \geq 0$, $\tau \in \mathbb{R}$ are arbitrary*

- *Monotone: if $u \le v$ then $\Lambda u \le \Lambda v$.*
- *Subadditive: $\Lambda(u+t) \le \Lambda u + t$.*
- *δ_0-submultiplicative: $\Lambda[(1+s)u] \le (1+s)\Lambda u - \delta_0 s$.*
- *Existence of a super-solution: there is $\underline{u} \in \mathbb{U}$ such that $\Lambda \underline{u} \le \underline{u}$.*

If in addition $\Theta(N_x, \mathcal{V}_h^x) < \pi/2$ for all $x \in X$, then the operator Λ is also

- *δ_1-causal: if $u^{\le \tau} = v^{\le \tau}$ then $(\Lambda u)^{\le \tau + \delta_1} = (\Lambda v)^{\le \tau + \delta_1}$.*

Proof. The monotony of Λ follows from the monotony of linear interpolation $\mathcal{I}_h^x$. Likewise, the subadditivitity of Λ follows from the same property of $\mathcal{I}_h^x$ (actually $\Lambda(u+t) = \Lambda u + t$ at all points $x \in X$ for which the stencil $\mathcal{V}_h^x$ does not contain the source x_*). Submutiplicativity is established as follows, using the 1-homogeneity of the interpolation operator $\mathcal{I}_h^x$

$$\min_{y \in \partial \mathcal{V}_h^x} (1+s)\, \mathrm{I}_h^x u(y) + N_x(x-y) \le (1+s) \Big[\min_{y \in \partial \mathcal{V}_h^x} \mathrm{I}_h^x u(y) + N_x(x-y) \Big] - s \min_{y \in \partial \mathcal{V}_h^x} N_x(x-y),$$

thus with $\delta_0 = \min\{N_x(x-y);\, x \in X,\, y \in \partial \mathcal{V}_h^x\}$. Consider the *directed graph*, with an edge (x,y) of length $N_x(x-y)$ whenever y is a vertex of $\mathcal{V}_h^x$. Then the distance from a given point $x_0 \in X$ to the source x_*, denoted by $\underline{u}(x_0)$, is finite by assumption and obeys $\Lambda \underline{u} \le \underline{u}$. Finally, Proposition 6.3 establishes δ_1-causality with δ_1 the minimal value of $\|y-x\| \mu_*(N_x) \cos \Theta(N_x, \mathcal{V}_h^x)$ among all $x \in X$ and all vertices y of the stencil $\mathcal{V}_h^x$. □

In the remainder of this appendix, we do not use the specific form (87, right) of the operator Λ, but only the properties established in Proposition 6.30. From monotony, subadditivity, and δ_0-submultiplicativity, one derives the discrete comparison principle.

Proposition 6.31 (Discrete comparison principle)**.** *Let $u, v \in \mathbb{U}$. If $u \le \Lambda u$ and $\Lambda v \le v$ then $u \le v$. In addition, if either inequality is strict then $u < v$.*

Proof. Let $x \in X$ be such that $t := u(x) - v(x)$ is maximal, so that $u \le v + t$ and $u(x) = v(x) + t$. Assuming that $t \ge 0$ we obtain $u(x) \le \Lambda u(x) \le \Lambda[v+t](x) \le \Lambda v(x) + t \le v(x) + t = u(x)$, by monotony and subadditivity. If either the first or last inequality is strict, we obtain a contradiction, thus $t < 0$ and therefore $u < v$ as announced. Otherwise note that $v_\varepsilon := (1+\varepsilon)v$ obeys $v_\varepsilon < \Lambda v_\varepsilon$ for any $\varepsilon > 0$ by δ_0-submultiplicativity, thus $u < v_\varepsilon$ by the previous argument, hence $u \le v$ by letting $\varepsilon \to 0$, which concludes the proof. □

Using in addition the continuity of Λ and the existence of a supersolution, one establishes that the fixed point problem (87, left) can be solved by iterating the operator. Finitely many iterations are sufficient if the operator is δ_1-causal.

Proposition 6.32 (Convergence of the global iterative method)**.** *The operator Λ admits a unique fixed point $\mathbf{u}$, and for any $u \in \mathbb{U}$ one has $\Lambda^n u \to \mathbf{u}$ as $n \to \infty$. If in addition Λ is δ_1-causal and $u > 0$, then $\Lambda^n u = \mathbf{u}$ for all $n \ge \max(\mathbf{u})/\delta_1$.*

Proof. Proposition 6.31 yields the uniqueness (but not the existence) of the fixed point $\boldsymbol{u}$. The null function $\overline{u} = 0$ satisfies $\Lambda\overline{u} \geq \delta_0 \geq 0 = \overline{u}$, by δ_0-submultiplicativity. Choose $t \geq 0$ sufficiently large so that $\overline{v} := \overline{u} - t \leq u \leq \underline{u} + t =: \underline{v}$, and note that $\overline{v} \leq \Lambda\overline{v}$ and $\Lambda\underline{v} \leq \underline{v}$ by subadditivity of Λ. Thus $\overline{v} \leq \cdots \leq \Lambda^n\overline{v} \leq \Lambda^n u \leq \Lambda^n\underline{v} \leq \cdots \leq \underline{v}$ by monotonicity of Λ, and induction on $n \geq 0$. By the monotone convergence theorem, $\Lambda^n\overline{v}$ and $\Lambda^n\underline{v}$ admit limits as $n \to \infty$. By continuity, these limits are fixed points of Λ, thus are equal to $\boldsymbol{u}$ by uniqueness. By the squeeze theorem we obtain $\Lambda^n u \to \boldsymbol{u}$ as announced.

Finally, assume that Λ is δ_1-causal, that $u > 0$, and note that $\mathbf{u} \geq \Lambda\overline{u} \geq \delta_0 > 0$. Then $u^{\leq 0} = \mathbf{u}^{\leq 0}$, and thus by induction $(\Lambda^n u)^{\leq n\delta_1} = (\Lambda^n \mathbf{u})^{\leq n\delta_1} = \mathbf{u}^{\leq n\delta_1}$ for all $n \geq 0$. The result follows. □

Global iteration is a poor way to allocate computational ressources in front propagation problems, and more efficient algorithms concentrate their efforts on a narrow band along the front. The convergence of iterative methods such as fast sweeping [QZZ07], the AGSI [BR06], or the FIM [JW08], follows from closely related arguments. The fast marching method, Algorithm 4 [Tsi95], solves the fixed point problem (87, left) in finitely many steps with complexity $\mathcal{O}(N \ln N)$, see [Mir19, Proposition A.2] for a proof based on the properties established in Proposition 6.30, causality included. We next establish that the properties of Proposition 6.30 are stable under perturbation.

Proposition 6.33 (Operator perturbation)**.** *Let $\alpha_*, \alpha^* \geq 0$, and for all $x \in X$ let $\alpha_x : X \to [-\alpha_*, \alpha^*]$. Define $\tilde{\Lambda} : \mathbb{U} \to \mathbb{U}$ by $\tilde{\Lambda}u(x) := \Lambda[u + \alpha_x](x)$. Then $\tilde{\Lambda}$ is continuous, monotone, subadditive, is $(\delta_0 - \alpha^*)$-submultiplicative if $\delta_0 > \alpha_*$, and admits the supersolution $(1 + \alpha^*/\delta_0)\underline{u}$. If Λ is δ_1-causal with $\delta_1 > \alpha_*$, then $\tilde{\Lambda}$ is $(\delta_1 - \alpha_*)$-causal.*

Proof. Fix $x \in X$, $u, v \in \mathbb{U}$, and $s, t \geq 0$. The continuity of $\tilde{\Lambda}$ immediately follows from the continuity of Λ. If $u \leq v$, then $u + \alpha_x \leq v + \alpha_x$, thus $\Lambda[u + \alpha_x] \leq \Lambda[v + \alpha_x]$ since Λ is monotone, therefore $\tilde{\Lambda}$ is monotone. One has $\Lambda[u+t+\alpha_x] \leq \Lambda[u+\alpha_x]+t$ by subadditivity of Λ, thus $\tilde{\Lambda}$ is subadditive. One has $\Lambda[(1 + s)u + \alpha_x] = \Lambda[(1 + s)(u + \alpha_x) - s\alpha_x] \leq \Lambda[(1 + s)(u + \alpha_x) + s\alpha_*] \leq \Lambda[(1 + s)(u + \alpha_x)] + s\alpha_* \leq (1 + s)\Lambda[u + \alpha_x] - \delta_0 s + s\alpha_*$, using successively the monotony, subadditivity, and δ_0-submultiplicativity of Λ, thus $\tilde{\Lambda}$ is $(\delta_0 - \alpha_*)$-submultiplicative if $\delta_0 > \alpha_*$. One has $\Lambda[(1 + s)\underline{u} + \alpha_x] \leq \Lambda[(1 + s)\underline{u}] + \alpha^* \leq (1 + s)\Lambda\underline{u} - \delta_0 s + \alpha^*$, by subadditivity and submultiplicativity of Λ, thus choosing $s = \alpha^*/\delta_0$ yields a supersolution of $\tilde{\Lambda}$. Finally, if Λ is δ_1-causal and $u^{\leq\tau} = v^{\leq\tau}$, then $(u + \alpha_x)^{\leq\tau-\alpha_*} = (v + \alpha_x)^{\leq\tau-\alpha_*}$ and therefore $(\Lambda[u + \alpha_x])^{\leq\tau-\alpha_*+\delta_1} \leq (\Lambda[v + \alpha_x])^{\leq\tau-\alpha_*+\delta_1}$, thus $\tilde{\Lambda}$ is $(\delta_1 - \alpha_*)$-causal. □

The *first-order source factorization* (61) falls in the framework of Proposition 6.33, with $\alpha_x(y) := u_*(x) - u_*(y) + \langle \nabla u_*(x), y - x \rangle$, which satisfies $\alpha_x(y) = \mathcal{O}(h^2/\|x - x_*\|)$, where u_* is the source factor and x_* is the source point. On the other hand, inspection of the proof of Proposition 6.30 yields that $\delta_0 = \hat{\delta}_0 h$ and $\delta_1 = \hat{\delta}_1 h$ where $\hat{\delta}_0$ and $\hat{\delta}_1$ are independent of the grid scale h. Thus $\delta_0 \geq \|\alpha_x\|_\infty$ and $\delta_1 \geq \|\alpha_x\|_\infty$ when h is sufficiently small (except for points x in a ball of radius $\mathcal{O}(h)$ around the source point x_*), and thus Proposition 6.33 applies to the factored scheme.

The *second and third-order schemes* define perturbations (62) and (63) which depend on the unknown u, and thus do *not* fall in the framework of Proposition 6.33. This

is the reason why, following [Set99], we use them in a cautious way: only in the post-processing step of the fast marching method right before the accepted value is frozen[6] see line 3 of Algorithm 4, and only if their magnitude does not exceed Ch^2 where C is an absolute constant. Together, these limitations ensure that the fast marching algorithm still terminates in a single pass over the domain, and produces an output obeying $\Lambda u = u + \mathcal{O}(h^2)$. Therefore $u = \mathbf{u} + \mathcal{O}(h)$ where $\mathbf{u}$ is the solution of the original scheme, by Proposition 6.34 below. In other words, we cannot prove that the high order variants of the scheme improve the solution accuracy, but at least they do not jeopardize first-order accuracy, and neither substantially increase computation time.

Proposition 6.34. *Let $u \in \mathbb{U}$ and let $k_*, k^* \geq 0$ be such that $k_* \leq \Lambda u - u \leq k^*$. Then $1 - k^*/\delta_0 \leq u/\boldsymbol{u} \leq 1 + k^*/\delta_0$.*

Proof. Let $s \geq 0$. Then $\Lambda[(1+s)u] \leq (1+s)\Lambda u - \delta_0 s \leq (1+s)(u+k^*) - \delta_0 s$, by δ_0-submultiplicativity. Choosing $s = k^*/(\delta_0 - k^*)$ yields $\Lambda[(1+s)u] \leq (1+s)u$ and thus $(1+s)u \geq \boldsymbol{u}$ by Proposition 6.31. On the other hand, $(1+s)\Lambda[u/(1+s)] \geq \Lambda u + s\delta_0 \geq u - k_* + s\delta_0$, again by δ_0-submultiplicativity. Choosing $s = k_*/\delta_0$ yields $\Lambda[u/(1+s)] \geq u/(1+s)$ and thus $u/(1+s) \leq \boldsymbol{u}$ by Proposition 6.31. The result follows. □

[6]Note that iterative methods, such as fast sweeping, lack the FMM specific concepts of accepted point, post-processing, and frozen value. For this reason, introducing high order finite differences can raise additional challenges, such as numerical instability along iterations.

7 Worst Case and Average Case Cardinality of Strictly Acute Stencils for Two Dimensional Anisotropic Fast Marching [MD20]

This section corresponds to the paper (with minor modification):

- J. M. Mirebeau and F. Desquilbet. Worst case and average case cardinality of strictly acute stencils for two dimensional anisotropic fast marching. In *Constructive Theory of Functions - 2019*, pages 157–180. Publishing House of Bulgarian Academy of Sciences, 2020

Abstract

We study a one dimensional approximation-like problem arising in the discretization of a class of Partial Differential Equations, providing worst case and average case complexity results. The analysis is based on the Stern-Brocot tree of rationals, and on a non-Euclidean notion of angles. The presented results generalize and improve on earlier work [Mir14b].

7.1 Introduction

This paper is devoted to the analysis of an approximation-like problem arising in the discretization of a class of Partial Differential Equations (PDE): eikonal equations, defined with respect to a possibly strongly anisotropic Finslerian metric. The results presented are related with the numerical solution of this equation on two dimensional cartesian grids, and their extension to higher dimension and/or to unstructured domains remains an open question. The unique viscosity solution to such an equation is a distance map, whose computation has numerous applications [Set99] in domains as varied as motion planning, seismic traveltime tomography [LBMV18], image processing [BC11], ... The construction studied in this paper is designed is to achieve a geometrical property - strict acuteness with respect to a given asymmetric norm - ensuring that the resulting numerical scheme is strictly causal [KS98, SV03, AM12, Mir14b, Mir14a]. This in turn enables efficient algorithms for solving the numerical scheme, in a single pass over the domain, with linear complexity, and possibly in parallel [Tsi95, RS09]. In order to better focus on the problem of interest, further discussion of the addressed PDE and of its discretization is postponed to §7.A.

We study in this paper a one dimensional approximation-like problem, involved in the construction of local stencils of minimal cardinality for a numerical solver of eikonal PDEs, see Definition 7.3 for a formal statement. The efficiency of the procedure is directly tied to the complexity of the numerical scheme. A few properties of this problem deviate from the common settings in approximation theory, and deserve to be discussed here.

- The main function $\varphi_F : \mathbb{R} \to]-\pi/2, \pi/2[$ considered benefits from regularity and integrability properties, derived from its geometrical interpretation §7.2.2. However these are fairly uncommon: $-\varphi_F$ is one-sided Lipschitz, and $\tan(\varphi_F)$ is bounded in the $L^1([0, 2\pi])$ norm.

- The approximation-like problem involves an interval subdivision procedure, that is reminiscent of e.g. dyadic splitting in non-linear approximation based on the Haar system [DeV98]. However, subdivision is here governed by the Stern-Brocot tree, and breaks the interval $[0, 2\pi]$ into unequal parts whose endpoints have rational tangents, see §7.3.

- We present a uniform "worst case" complexity result, but also an "average case" result under random shifts, see Theorem 7.4. Because of the peculiarities of the approximation procedure, a more favorable estimate is obtained in the average case.

In the rest of this introduction, we introduce the notations and concepts necessary to state our main result. Our first step is to equip the Euclidean space $\mathbb{R}^2$ with the anisotropic geometry defined by a (possibly) asymmetric norm. Here and below, all asymmetric norms are on $\mathbb{R}^2$.

Definition 7.1. *An asymmetric norm is a function $F : \mathbb{R}^2 \to \mathbb{R}_+$ which is 1-positively homogeneous, obeys the triangular inequality, and vanishes only at the origin:*

$$F(\lambda u) = \lambda F(u), \qquad F(u+v) \leq F(u) + F(v), \qquad F(u) = 0 \Leftrightarrow u = 0,$$

for all $u, v \in \mathbb{R}^2$, $\lambda \geq 0$. The anisotropy ratio of F is defined as $\mu(F) := \max_{|u|=|v|=1} \frac{F(u)}{F(v)}$.

Note that an asymmetric norm is always a continuous and convex function. We denote by $\measuredangle(u, v) \in [0, \pi]$ the *unoriented* Euclidean angle between two vectors $u, v \in \mathbb{R}^2 \setminus \{0\}$, which is characterized by the identity

$$\cos \measuredangle(u, v) = \frac{\langle u, v \rangle}{\|u\| \|v\|}.$$

The next definition introduces a generalized measure of angle, associated with an asymmetric norm. We only consider acute angles, since obtuse angles will not be needed, and because their definition raises issues. The notion of F-acute angle is similarly defined in [Mir14b, Vla08], but the related angular measure is new.

Definition 7.2. *Let F be an asymmetric norm, which is differentiable except at the origin, and let $u, v \in \mathbb{R}^2 \setminus \{0\}$. We say that u, v form an F-acute angle iff $\langle \nabla F(u), v \rangle \geq 0$. We define the F-angle $\measuredangle_F(u, v) \in [0, \pi/2] \cup \{\infty\}$ by*

$$\cos \measuredangle_F(u, v) := \langle \nabla F(u), v \rangle / F(v) \tag{88}$$

if u, v form an F-acute angle. Otherwise we let $\measuredangle_F(u, v) := +\infty$.

We show in Lemma 7.7 that the r.h.s. of (88) is at most 1, so that $\measuredangle_F(u, v)$ is well defined, with equality if $u = v$, so that $\measuredangle_F(u, u) = 0$. If F is the Euclidean norm, then one easily checks that the F-angle coincides with the usual Euclidean angle, when the latter is acute. More generally, if $F(u) = \|Au\|$ for some invertible linear map A, then $\measuredangle_F(u, v) = \measuredangle(Au, Av)$, when the latter is acute. In general however, one has $\measuredangle_F(u, v) \neq$

$\measuredangle_F(v,u)$, and F-acuteness is not a symmetric relation. The differentiability assumption in Definition 7.2 can be removed, see Definition 7.5.

The following definition introduces (F,α)-acute stencils, which are at the foundation of our numerical scheme, see Figure page 119. Their cardinality is directly proportional to the algorithmic complexity of our eikonal PDE solver, see §7.A, hence it is important to choose them as small as possible. When $\alpha = \pi/2$ one recovers the F-acute stencils of [Mir14b], and closely related concepts are considered in [KS98, SV03, Vla08, AM12].

Definition 7.3. *A* stencil *is a finite sequence of pairwise distinct vectors $u_1, \cdots, u_n \in \mathbb{Z}^2$, $n \geq 4$, such that*

$$\det(u,v) = 1, \qquad\qquad \langle u, v\rangle \geq 0,$$

for all $u = u_i$, $v = u_{i+1}$, $1 \leq i \leq n$, with the convention $u_{n+1} := u_n$. It is said (F,α)-acute, where F is an asymmetric norm and $\alpha \in]0,\pi/2]$, iff with the same notations one has

$$\measuredangle_F(u,v) \leq \alpha, \qquad\qquad \measuredangle_F(v,u) \leq \alpha. \tag{89}$$

We let $N(F,\alpha)$ denote the minimal cardinality of an (F,α)-acute stencil.

We provide in §7.3.2 a simple and efficient algorithm, based on a recursive refinement procedure and which is effectively used in our numerical implementation, for producing an (F,α)-acute stencil of minimal cardinality $N(F,\alpha)$. A similar method appears in [Mir14b] when $\alpha = \pi/2$. The main result of this paper is the following estimate of $N(F,\alpha)$, both in the worst case and in the average case over random rotations of the asymmetric norm F. The average case makes sense in view of our application to PDE discretizations §7.A, since the orientation of the grid can be set and modified arbitrarily.

Theorem 7.4. *For any asymmetric norm F and any $\alpha \in]0,\pi/2]$, one has*

$$N(F,\alpha) \leq C\frac{\mu}{\alpha^2}\ln\big(\frac{\ln\mu}{\alpha^2}\big), \qquad \int_0^{2\pi} N(F\circ R_\theta,\alpha)\,\mathrm{d}\theta \leq C\frac{\ln(\mu)}{\alpha^2}\ln\big(\frac{\mu}{\alpha^2}\big). \tag{90}$$

where $\mu = \max\{\mu(F), 12\}$, R_θ denotes the rotation of angle $\theta \in \mathbb{R}$, and C is an absolute constant.

In the intended applications, one typically has $\mu(F) \lesssim 100$. The most pronounced anisotropies $\mu(F) \approx 100$ are often encountered in image processing methods [BC11, Mir14a], and this bound is large enough that the asymptotic behavior of (90) w.r.t. μ is meaningful to our use cases. In contrast, we do confess that it seems pointless to let $\alpha \to 0$ in our applications (typically we set $\alpha = \pi/3$). If one fixes $\alpha_0 \in]0,\pi/2]$ then

$$N(F,\alpha_0) \leq C\mu\ln\ln\mu, \qquad \int_0^{2\pi} N(F\circ R_\theta,\alpha_0)\,\mathrm{d}\theta \leq C\ln^2(\mu). \tag{91}$$

uniformly w.r.t. μ. This improves on [Mir14b], whose arguments are limited to the case $\alpha_0 = \pi/2$, and where the sub-optimal bounds $\mu\ln(\mu)$ (resp. $\ln^3(\mu)$) are obtained for (91, left) (resp. right).

Outline. The notion of F-acute angle, see Definition 7.2, is described in more detail §7.2, where related tools are introduced. The Stern-Brocot tree, an arithmetic structure underlying concept of stencil in Definition 7.3, is discussed in §7.3. We conclude in §7.4 the proof of Theorem 7.4. Some context on the intended applications of the presented results is given in §7.A.

7.2 Anisotropic angle

This section is devoted to the study of the anisotropic measure of angle $\measuredangle_F(u,v)$ of Definition 7.2, where $u, v \in \mathbb{R}^2 \setminus \{0\}$ and F is an asymmetric norm. Some elementary comparison properties, with the Euclidean angle $\measuredangle(u,v)$ or with another angle $\measuredangle_F(u,w)$, are presented §7.2.1. We prepare in §7.2.2 (resp. §7.2.3) the proof of the average case (resp. worst case) estimate of Theorem 7.4, by introducing a helper function φ_F (resp. $\psi_F^{\pm}$) for which we show a $L^1([0,2\pi])$ norm estimate and a comparison principle with $\measuredangle_F$.

In the rest of this section, we fix an asymmetric norm F, assumed to be continuously differentiable on $\mathbb{R}^2 \setminus \{0\}$. That is with the exception of the following definition and proposition, where we briefly consider the case of non-differentiable norms, and show that the smoothness assumption holds without loss of generality. Closely related arguments are found in Lemma 2.11 of [Mir14b].

Definition 7.5 (Generalization of $\measuredangle_F(u,v)$ with no differentiability assumption)**.** *Let F be an asymmetric norm, and let $u, v \in \mathbb{R}^2 \setminus \{0\}$. We say that u, v form an F-acute angle iff $F(u+\delta v) \geq F(u)$ for all $\delta \geq 0$. In that case we let $\alpha = \measuredangle_F(u,v) \in [0, \pi/2]$ denote the smallest value such that*

$$F(u+\delta v) \geq F(u) + \delta \cos(\alpha) F(v), \tag{92}$$

for all $\delta \geq 0$. If u, v do not form an F-acute angle, then we let $\measuredangle_F(u,v) := \infty$.

Proposition 7.6. *Definitions 7.2 and 7.5 agree on differentiable norms. Also, if $F_n \to F$ locally uniformly as $n \to \infty$, where $(F_n)_{n\geq 0}$ and F are asymmetric norms, and $u, v \in \mathbb{R}^2 \setminus \{0\}$, then*

$$\measuredangle_F(u,v) \leq \liminf_{n\to\infty} \measuredangle_{F_n}(u,v). \tag{93}$$

If Theorem 7.4 holds under the additional assumption $F \in C^1(\mathbb{R}^2 \setminus \{0\})$, then it does without it.

Proof. Under the assumptions of Definition 7.2 one has $F(u+\delta v) = F(u) + \delta\langle \nabla F(u), v\rangle + o(\delta)$ by differentiability of F at u, and $F(u+\delta v) \geq F(u) + \delta\langle \nabla F(u), v\rangle$ by convexity of F, for any $\delta \geq 0$ and any $v \in \mathbb{R}^2 \setminus \{0\}$. Thus Definitions 7.5 and 7.2 agree. The lower semi-continuity property (93) follows from the fact that (92) is closed under uniform convergence. Therefore if a given stencil is (F_n, α)-acute for all $n \geq 0$, then it is also (F, α)-acute see Definition 7.3. Thus $N(F,\alpha) \leq \liminf_{n\to\infty} N(F_n, \alpha)$, and likewise for the l.h.s. of (90, right). Finally, we observe that any asymmetric norm F is the locally uniform limit of a sequence of asymmetric norms $F_n \in C^\infty(\mathbb{R}^2 \setminus \{0\})$, $n \geq 1$, defined as

$$F_n(u) := \int_{\mathbb{R}} F(R_\theta u)\, \rho_n(\theta)\, \mathrm{d}\theta,$$

where $\rho_n(\theta) := n\rho(n\theta)$, and the mollifier ρ is smooth, non-negative, compactly supported, and has unit integral. The statement regarding Theorem 7.4 follows, which concludes the proof. □

7.2.1 Elementary comparison properties

This subsection is devoted to elementary comparisons between $\measuredangle_F(u,v)$ and the angle between other vectors, see Lemma 7.8, or the Euclidean angle $\measuredangle(u,v)$, see Proposition 7.9, where $u, v \in \mathbb{R}^2 \setminus \{0\}$. In addition, Lemma 7.7 below was announced and used in the introduction to show that $\measuredangle_F(u,v)$ is well defined, and that $\measuredangle_F(u,u) = 0$. Throughout this subsection, F denotes a fixed asymmetric norm, assumed to be differentiable except at the origin.

Lemma 7.7. *For any $u, v \in \mathbb{R}^2$, with $u \neq 0$, one has*

$$\langle \nabla F(u), u\rangle = F(u), \qquad\qquad \langle \nabla F(u), v\rangle \leq F(v). \tag{94}$$

Proof. Euler's identity for the 1-homogeneous function F yields (94, left), whereas the triangular inequality $F(u+\delta v) \leq F(u) + \delta F(v)$ for all $\delta \geq 0$ yields (94, right). □

The next lemma shows that the F-angle is non-increasing when an angular sector is split.

Lemma 7.8. *Let u, v form an F-acute angle, and let $w := \alpha u + \beta v$ for some $\alpha, \beta > 0$. Then*

$$\max\{\measuredangle_F(u,w), \measuredangle_F(w,v)\} \leq \measuredangle_F(u,v)$$

Proof. Assume w.l.o.g. that $\alpha = 1$, and denote $\lambda := \cos \measuredangle_F(u,v)$. By convexity of F one has

$$\langle \nabla F(w), v\rangle = \langle \nabla F(u+\beta v), v\rangle = \langle \nabla F(u+\beta v) - \nabla F(u), v\rangle + \langle \nabla F(u), v\rangle \geq 0 + \lambda F(v).$$

On the other hand, one obtains noting that $\lambda \in [0,1]$ by assumption

$$\langle \nabla F(u), w\rangle = \langle \nabla F(u), u+\beta v\rangle \geq F(u) + \lambda\beta F(v) \geq \lambda(F(u)+\beta F(v)) \geq \lambda F(u+\beta v) = \lambda F(w).$$

□

The last proposition of this subsection is an upper bound on the F-angle in terms of the Euclidean angle and of the anisotropy ratio $\mu(F)$ of the asymmetric norm. This upper bound grows non-linearly and perhaps more quickly than one may expect, namely as the square root of the Euclidean angle, because we do not make any quantitative assumption on the smoothness of F. Here and below we denote $u^\perp := (-b, a)$ for any $u = (a,b) \in \mathbb{R}^2$.

Proposition 7.9. *For any $u, v \in \mathbb{R}^2 \setminus \{0\}$, one has assuming $\mu(F)\measuredangle(u,v) \leq 1/2$*

$$\measuredangle_F(u,v) \leq \sqrt{5\mu(F)\measuredangle(u,v)}. \tag{95}$$

Proof. Denote $\theta := \measuredangle(u,v)$, $\alpha := \measuredangle_F(u,v)$, and $\mu := \mu(F)$. Assume w.l.o.g. that $v = u + \tan(\theta)u^\perp$. Then

$$\langle \nabla F(u), v\rangle = \langle \nabla F(u), u\rangle + \tan(\theta)\langle \nabla F(u), u^\perp\rangle \geq F(u) - \tan(\theta)F(-u^\perp).$$
$$F(v) = F(u + \tan\theta u^\perp) \leq F(u) + \tan(\theta)F(u^\perp).$$

Observing that $F(u^\perp) \leq \mu F(u)$ and $F(-u^\perp) \leq \mu F(u)$, we obtain

$$\frac{1-\mu\tan\theta}{1+\mu\tan\theta} \leq \frac{F(u) - F(-u^\perp)\tan\theta}{F(u)+F(u^\perp)\tan\theta} \leq \frac{\langle \nabla F(u), v\rangle}{F(v)} = \cos\alpha = \frac{1-\tan^2(\alpha/2)}{1+\tan^2(\alpha/2)}. \tag{96}$$

This implies $\tan^2(\alpha/2) \leq \mu\tan\theta$. We conclude the proof of (95) observing that $\tan(\alpha/2) \geq \alpha/2$, and $\tan\theta \leq (5/4)\theta$, both estimates by convexity of tan on $[0,\pi/2[$ and since $\theta \leq 1/2$. Note also that $\mu\tan\theta \leq (5/4)\mu\theta \leq 5/8 < 1$ by assumption, which shows that the l.h.s. of (96) is positive, and thus excludes the case where $\measuredangle_F(u,v) = \infty$, see Definition 7.2. □

7.2.2 Gradient deviation

We describe and study a function φ_F attached to the asymmetric norm F of interest, introduced in [Mir14b] and used in the proof of the average case estimate in Theorem 7.4. More precisely, the quantity $\varphi_F(u)$ is the *oriented* Euclidean angle between a given vector $u \in \mathbb{R}^2 \setminus \{0\}$ and the gradient $\nabla F(u)$. Note that these two vectors are aligned if F is proportional to the Euclidean norm. The main results of this section are an L^1 estimate of φ_F, see Lemma 7.13, and a comparison with the F-angle, see Proposition 7.14.

Definition 7.10. *For each $u \in \mathbb{R}^2 \setminus \{0\}$, define a signed angle $\varphi_F(u) \in]-\pi/2, \pi/2[$ by*

$$\langle u^\perp, \nabla F(u)\rangle = F(u)\tan\varphi_F(u). \tag{97}$$

For $\theta \in \mathbb{R}$, we abusively denote $\varphi_F(\theta) := \varphi_F(\,(\cos\theta, \sin\theta)\,)$.

The next lemma shows, as announced, that $|\varphi_F(u)|$ is the Euclidean angle between the given vector u and its image by the gradient of F, and establishes a uniform upper bound for φ_F.

Lemma 7.11. *For any $u \in \mathbb{R}^2 \setminus \{0\}$, one has*

$$|\varphi_F(u)| = \measuredangle(u, \nabla F(u)), \qquad\qquad |\tan\varphi_F(u)| \leq \mu(F). \tag{98}$$

Proof. Equality (98, left) follows from Euler's identity (94, left) and the definition (97). Estimate (98, right) follows from $-F(-u^\perp) \leq \langle u^\perp, \nabla F(u)\rangle \leq F(u^\perp)$ see (94, right), and from the upper bound $F(\pm u^\perp) \leq \mu(F)F(u)$ which holds by Definition 7.1 of the anisotropy ratio. □

We recall in the next proposition, without proof, two key properties of the function φ_F established in [Mir14b]: a one-sided regularity property, and an upper bound on the integral of $\tan(\varphi_F)$ on any interval. See the plots of φ_F on Figure page 18.

Proposition 7.12 (Proposition 3.6 in [Mir14b]). *The function $\varphi_F : \mathbb{R} \to]-\pi/2, \pi/2[$ obeys:*

- *(Regularity) For all $\theta \in \mathbb{R}$, one has $\varphi_F'(\theta) \geq -1$.*
- *(Integral bound) One has $|\int_{\theta_*}^{\theta^*} \tan \varphi_F(\theta)\, \mathrm{d}\theta| \leq \ln \mu(F)$ for all $\theta_*, \theta^* \in \mathbb{R}$.*

Combining the one-sided regularity property and the integral bound, one obtains an L^1 estimate of $\tan(\varphi_F)$, as shown in the next lemma, which turns out to be a key ingredient of the proof of the average case estimate (90, right), see §7.4.2.

Corollary 7.13 (L^1 estimate of $\tan \varphi_F$). *One has with $C = 2\pi\sqrt{3}$*

$$\int_0^{2\pi} |\tan \varphi_F(\theta)| \mathrm{d}\theta \leq C(1 + \ln \mu(F)) \tag{99}$$

Proof. In view of Proposition 7.12 (Integral bound), of the continuity of φ_F, and of its 2π-periodicity, there exists $\alpha_0 \in \mathbb{R}$ such that $\varphi_F(\alpha_0) = 0$. Then inductively for $n \geq 0$ let

- β_n be the smallest $\beta \geq \alpha_n$ such that $|\varphi_F(\beta)| = \pi/3$,
- α_{n+1} be the smallest $\alpha \geq \beta_n$ such that $\varphi_F(\alpha) = 0$.

The sequences $(\alpha_n, \beta_n)_{n \geq 0}$ are well defined, thanks to the periodicity of φ_F, except if $|\varphi_F| < \pi/3$ uniformly, but in that case the announced result (99) clearly holds. If $\varphi_F(\beta_n) = \pi/3$ for some $n \geq 0$ then $\alpha_{n+1} - \beta_n \geq \pi/3$, whereas if $\varphi_F(\beta_n) = -\pi/3$ one has $\beta_n - \alpha_n \geq \pi/3$, by Proposition 7.12 (Regularity). Therefore $\alpha_{n+1} - \alpha_n \geq \pi/3$ for all $n \geq 0$, thus $\alpha_6 \geq \alpha_0 + 2\pi$, which implies

$$\int_0^{2\pi} |\tan \varphi_F(\theta)|\, \mathrm{d}\theta \leq \int_0^{2\pi} \tan(\pi/3)\, \mathrm{d}\theta + 6 \ln \mu(F) \leq 2\pi\sqrt{3} + 6 \ln \mu(F). \tag{100}$$

On each interval $[\alpha_n, \beta_n] \cap [0, 2\pi]$ we used the upper bound $|\varphi_F(\theta)| \leq \pi/3$, which holds by definition of β_n. On each interval $[\beta_n, \alpha_{n+1}] \cap [0, 2\pi]$ we used Proposition 7.12 (Integral bound) and the fact that φ_F does not change sign, which holds by definition of α_{n+1}. □

The last result of this subsection can be regarded as a refinement of Proposition 7.9.

Proposition 7.14 (Estimate of $\measuredangle_F$ in terms of φ_F). *Let $u, v \neq 0$ be such that $\measuredangle(u, v) \leq \pi/3$. Then one has, with $C = 32$*

$$\min\{\measuredangle_F(u, v), 2\}^2 \leq C \measuredangle(u, v) \max\{\measuredangle(u, v), |\tan \varphi_F(u)|, |\tan \varphi_F(v)|, \}. \tag{101}$$

Proof. Denote $\theta := \measuredangle(u, v)$ and $\alpha := \measuredangle_F(u, v)$. Assuming w.l.o.g. that $\|u\| = \|v\| = 1$ and $\det(u, v) > 0$ one has

$$v = (u + u^\perp \tan \theta) \cos \theta, \qquad \text{and } u = (v - v^\perp \tan \theta) \cos \theta.$$

Using linearity in the first line, and convexity in the second line, we obtain

$$\langle v, \nabla F(u)\rangle = \langle u + u^\perp \tan\theta, \nabla F(u)\rangle \cos\theta = F(u)(1 + \tan\varphi_F(u)\tan\theta)\cos\theta \tag{102}$$

$$\begin{aligned} F(u) &= F(v - v^\perp \tan\theta)\cos\theta \\ &\geq (F(v) - \langle v^\perp, \nabla F(v)\rangle \tan\theta)\cos\theta = F(v)(1 - \tan\varphi_F(v)\tan\theta)\cos\theta \end{aligned} \tag{103}$$

Assume for a moment that $-\tan\varphi_F(u)\tan\theta \geq 1/2$. Recalling that $\theta \leq \pi/3$, thus $\tan\theta \leq 2\theta$, we obtain $-\tan\varphi_F(u)\theta \geq 1/4$ and the announced result (101) is proved. Likewise if $\tan\varphi_F(v)\tan\theta \geq 1/2$. In particular, if $\alpha = +\infty$ then $\langle \nabla F(u), v\rangle \leq 0$ by Definition 7.2, and therefore $-\tan\varphi_F(u)\tan\theta \geq 1$ by (102), so that the result is proved.

In the following, we let $t_u := \tan\varphi_F(u)$, $t_v := \tan\varphi_F(v)$. Based on the previous argument we assume w.l.o.g. that $t_u\tan\theta \geq -1/2$, $t_v\tan\theta \leq 1/2$ and $\alpha \neq \infty$. We obtain from (102)

$$\cos\alpha = \frac{\langle v, \nabla F(u)\rangle}{F(v)} = \frac{\langle v, \nabla F(u)\rangle}{F(u)} \times \frac{F(u)}{F(v)} \geq (1 + t_u\tan\theta)(1 - t_v\tan\theta)\cos^2\theta.$$

Taking logarithms yields with $t := \max\{0, -t_u, t_v\}$

$$-\ln\cos\alpha \leq -2\ln(1 - t\tan\theta) - 2\ln\cos\theta. \tag{104}$$

An elementary function analysis shows that $-\ln\cos\alpha \geq \alpha^2/2$ for $\alpha \in [0, \pi/2[$, and $-\ln\cos\theta \leq \theta^2$ for $\theta \in [0, \pi/3]$. In addition $\tan\theta \leq 2\theta$ for $\theta \in [0, \pi/3]$, and $-\ln(1-x) \leq 2x$ for $x \in [0, 1/2]$. Inserting these bounds in (104) yields the announced result

$$\alpha^2/2 \leq -\ln\cos\alpha \leq 4\max\{-\ln(1 - t\tan\theta), -\ln\cos\theta\} \leq 4\max\{4t\theta, 2\theta^2\}. \qquad \square$$

7.2.3 Regularized gradient deviation

We consider in this subsection two 1-Lipschitz regularizations ψ_F^+ and ψ_F^- of the gradient deviation φ_F. See the plots of $\psi_F^\pm$ on Figure page 18. Note that $-\varphi_F$ is already (but also only) one-sided 1-Lipschitz, see Proposition 7.12 (Regularity). We extend to $\psi_F^\pm$ some of the results of §7.2.2, namely the L^1-norm estimate in Corollary 7.18 and the comparison with the F-angle in Proposition 7.14, which are used §7.4.1 in the proof of the worst case estimate in Theorem 7.4. We recall that $\varphi_F : \mathbb{R} \to]-\pi/2, \pi/2[$ is 2π-periodic.

Definition 7.15. *Define for any $\theta \in \mathbb{R}$*

$$\psi_F^+(\theta) := \max_{\eta\geq 0} \varphi_F(\theta + \eta) - \eta, \qquad \psi_F^-(\theta) := \min_{\eta\geq 0} \varphi_F(\theta - \eta) + \eta. \tag{105}$$

The functions ψ_F^- and ψ_F^+ define an upper and lower envelope of φ_F: for any $\theta \in \mathbb{R}$

$$-\pi/2 < \inf_{\mathbb{R}} \varphi_F \leq \ \psi_F^-(\theta) \leq \varphi_F(\theta) \leq \psi_F^+(\theta) \ \leq \sup_{\mathbb{R}} \varphi_F < \pi/2.$$

They play symmetrical roles, up to replacing φ_F with $\theta \mapsto -\varphi_F(-\theta)$, which amounts to reversing the orientation of the plane $\mathbb{R}^2$. Hence results established for ψ_F^+ automatically extend to ψ_F^-.

Lemma 7.16. *The map $\psi_F^+ : \mathbb{R} \to]-\pi/2, \pi/2[$ is 1-Lipschitz.*

Proof. By design (105, left) the function ψ_F^+ is one-sided 1-Lipschitz: for all $\theta \in \mathbb{R}$, $h \geq 0$

$$\psi_F^+(\theta + h) = \sup_{\eta \geq h} \varphi_F(\theta + \eta) - (\eta - h) \leq \psi_F^+(\theta) + h.$$

On the other hand one has $\psi_F^+(\theta - h) \leq \psi_F^+(\theta) + h$, for all $h \geq 0$, as follows from the same property of the function φ_F, see Proposition 7.12 (Regularity). Combining these two estimates, we obtain $\psi_F^+(\theta + h) \leq \psi_F^+(\theta) + |h|$, for all $\theta \in \mathbb{R}$ and all $h \in \mathbb{R}$ (positive or negative), hence ψ_F^+ is 1-Lipschitz as announced. □

The next lemma and corollary are devoted to estimating the $L^1([0, 2\pi])$ norm of ψ_F^+. We denote by $|A|$ the Lebesgue measure of a measurable set $A \subset \mathbb{R}$.

Lemma 7.17. *Let $\theta_0, \theta_1 \in \mathbb{R}$ be such that $\psi_F^+(\theta_0) = \psi_F^+(\theta_1)$. Then*

$$|\{\theta \in [\theta_0, \theta_1];\, \psi_F^+(\theta) > \varphi_F(\theta)\}| \leq |\{\theta \in [\theta_0, \theta_1];\, \psi_F^+(\theta) = \varphi_F(\theta)\}|. \tag{106}$$

Proof. Denote by A_0 (resp. A_1) the set appearing in (106, left) (resp. (106, right)). Then

$$0 = \psi_F^+(\theta_1) - \psi_F^+(\theta_0) = \int_{\theta_0}^{\theta_1} \frac{\mathrm{d}}{\mathrm{d}\theta} \psi_F^+ = \int_{A_0} \frac{\mathrm{d}}{\mathrm{d}\theta} \psi_F^+ + \int_{A_1} \frac{\mathrm{d}}{\mathrm{d}\theta} \psi_F^+ \geq |A_0| - |A_1|,$$

where we used the observation that $\frac{d}{d\theta}\psi_F^+(\theta) = 1$ for all $\theta \in A_0$, whereas $\frac{d}{d\theta}\psi_F^+(\theta) \geq -1$ for a.e. $\theta \in A_1$. The result follows. □

Corollary 7.18 (L^1 estimate of ψ_F^+)**.**

$$\int_0^{2\pi} \max\{0, \tan \psi_F^+ - 1\} \leq 2 \int_0^{2\pi} \max\{0, \tan \varphi_F - 1\}$$

Proof. As observed in Corollary 7.13 there exists $\theta_0 \in \mathbb{R}$ such that $\varphi_F(\theta_0) = 0$. Thus $\varphi_F(\theta) \leq \pi/4$ for all $\theta \in [\theta_0 - \pi/4, \theta_0]$, and therefore $\psi_F^+(\theta_0 - \pi/4) \leq \pi/4$. As a result, the level sets

$$\Psi(\lambda) := \{\theta \in \mathbb{R};\, \psi_F^+(\theta) > \lambda\}, \qquad \Phi(\lambda) := \{\theta \in \mathbb{R};\, \varphi_F(\theta) > \lambda\},$$

are strict subsets of $\mathbb{R}$ for any $\lambda \geq \pi/4$. They are also 2π-periodic sets, and for that reason we denote $\widetilde{\Phi}(\lambda) := \Phi(\lambda) \cap [0, 2\pi[$ and $\widetilde{\Psi}(\lambda) := \Psi(\lambda) \cap [0, 2\pi[$. Applying Lemma 7.17 to the closure $[\theta_0, \theta_1]$ of each connected component of $\Psi(\lambda)$, and using periodicity, we obtain $|\widetilde{\Psi}(\lambda) \setminus \widetilde{\Phi}(\lambda)| \leq |\widetilde{\Phi}(\lambda)|$. Thus $|\widetilde{\Psi}(\lambda)| \leq 2|\widetilde{\Phi}(\lambda)|$, and therefore, as announced

$$\begin{aligned}
\int_0^{2\pi} \max\{0, \tan \psi_F^+ - 1\} &= \int_{\pi/4}^{\pi/2} |\widetilde{\Psi}(\lambda)| \tan \lambda \, \mathrm{d}\lambda \\
&\leq 2 \int_{\pi/4}^{\pi/2} |\widetilde{\Phi}(\lambda)| \tan \lambda \, \mathrm{d}\lambda \\
&= 2 \int_0^{2\pi} \max\{0, \tan \varphi_F - 1\}
\end{aligned}$$

□

From Corollaries 7.13 and 7.18 we obtain

$$\int_0^{2\pi} \max\{1, \tan\psi_F^+\} \leq C \ln \mu, \tag{107}$$

where $\mu := \max\{2, \mu(F)\}$ and C is an absolute constant. The same result holds for $\max\{1, -\tan\psi_F^-\}$, by a similar argument, see the comment after Definition 7.15. Finally, we compare the F-angle of two vectors with such integral quantities.

Proposition 7.19 (Estimate of $\measuredangle_F$ in terms of $\psi_F^\pm$)**.** *Let $u, v \in \mathbb{R}^2 \setminus \{0\}$, with $\measuredangle(u,v) \leq \pi/3$. Let $\Theta \subset \mathbb{R}$ be a corresponding angular sector, with $|\Theta| = \measuredangle(u,v)$. Then*

$$\min\{\measuredangle_F(u,v), 2\}^2 \leq C' \int_\Theta \max\{1,\ \tan\psi_F^+,\ -\tan\psi_F^-\}, \tag{108}$$

where $C' = 4C$ and C is from Proposition 7.14.

Proof. By angular sector, we mean that up to exchanging u and v one has $\Theta = [\theta_u, \theta_v[$ where u (resp. v) is positively proportional to $(\cos\theta_u, \sin\theta_u)$ (resp. $(\cos\theta_v, \sin\theta_v)$). By Proposition 7.14 one has

$$\measuredangle_F(u,v)^2 \leq C \max\{|\Theta|^2,\ |\Theta|\,|\tan\varphi_F(\theta_u)|,\ |\Theta||\tan\varphi_F(\theta_v)|\}.$$

If $\measuredangle_F(u,v)^2 \leq C|\Theta|^2$ then the announced result (108) is proved, since $|\Theta| \leq \pi/3$. Otherwise we may assume w.l.o.g that $\measuredangle_F(u,v)^2 \leq |\Theta|\,|\tan\varphi_F(\theta_u)|$. Denoting $\psi_F := \max\{\psi_F^+, -\psi_F^-\}$ one has $\psi_F(\theta_u + h) \geq \varphi_* - h$ for all $h \geq 0$, where $\varphi_* := |\varphi_F(\theta_u)|$. We conclude by case elimination:

- If $\varphi_* \leq \pi/3$, then $\measuredangle_F(u,v)^2 \leq |\Theta|\tan(\pi/3)$, hence (108) holds as announced.
- Otherwise if $\varphi_* + \measuredangle(u,v) \geq \pi/2$, we obtain

$$\begin{aligned}\int_\Theta |\tan\psi_F| &\geq \int_0^{\frac{\pi}{2}-\varphi_*} \tan(\varphi_* - h)\,\mathrm{d}h = \ln\Big(\frac{\sin(2\varphi_*)}{\cos\varphi_*}\Big) = \ln(2\sin\varphi_*)\\ &\geq \ln(2\sin(\pi/3)) = \frac{\ln 3}{2}.\end{aligned}$$

  Thus the r.h.s. of (108) is bounded below by $C\ln(3)/2 \geq 2^2$, hence (108) holds as announced.

- Otherwise if $\varphi_* + \measuredangle(u,v) \leq \pi/2$, then we obtain for all $\theta \in \Theta$

$$\tan\psi_F(\theta) \geq \tan(\varphi_* - (\pi/2 - \varphi_*)) = -\cot(2\varphi_*) = \frac{1}{2}(\tan\varphi_* - \cot\varphi_*) \geq \frac{1}{4}\tan\varphi_*,$$

  using that $\varphi_* \geq \pi/3$ in the last inequality. Therefore $\int_\Theta \tan\psi_F \geq \frac{1}{4}|\Theta|\tan(\varphi_*)$, which implies (108) and concludes the proof. □

7.3 The Stern-Brocot tree

We describe a variant of the Stern-Brocot tree [Niq07], an arithmetic structure which allows to effectively construct and study the minimal (F, α)-acute stencil considered in Definition 7.3. We formally introduce the Stern-Brocot tree in this introduction, and then we relate it in §7.3.2 with the stencils of Definition 7.3. We estimate in §7.3.3 the cardinality of a subtree, based on the number of its inner leaves and on a measure of their depth, for use in the proof §7.4.1 of the worst case estimate of Theorem 7.4.

Let $\mathcal{Z}$ collect all elements of $\mathbb{Z}^2$ whose coordinates are co-prime, and $\mathcal{T}$ all elements of $\mathcal{Z}^2$ with unit determinant and a non-negative scalar product.

$$\mathcal{Z} := \{(a,b) \in \mathbb{Z}^2 \setminus \{0\};\ \gcd(a,b) = 1\}, \quad \mathcal{T} := \{(u,v) \in \mathcal{Z}^2;\ \langle u, v\rangle \geq 0,\ \det(u,v) = 1\}.$$

We often denote $T = (u, v)$ the elements of the set $\mathcal{T}$.

Definition 7.20. *For any $T = (u, v) \in \mathcal{T}$, we refer to $T' = (u, u + v) \in \mathcal{T}$ and $T'' = (u + v, v) \in \mathcal{T}$ as its children, and we denote this relation by $T \lhd T'$ and $T \lhd T''$. We also let*

$$S(T) = \langle u, v\rangle, \qquad \Delta(T) = \min\{\|u\|^2, \|v\|^2\}.$$

By construction one has for any $T \lhd T' \in \mathcal{T}$

$$S(T) \geq 0, \qquad \Delta(T) \geq 1, \qquad S(T') \geq S(T) + \Delta(T), \qquad \Delta(T') \geq \Delta(T). \tag{109}$$

Definition 7.21. *A chain in $\mathcal{T}$ is a finite sequence $T_0 \lhd \cdots \lhd T_n$, where $n \geq 0$. We write $T_* \preceq T^*$ iff there exists a chain $T_* = T_0 \lhd \cdots \lhd T_n = T^*$ in $\mathcal{T}$ for some $n \geq 0$.*

The next lemma fully describes the graph $(\mathcal{T}, \lhd)$. For that purpose, denoting by (e_1, e_2) the canonical basis of $\mathbb{R}^2$ we let

$$\mathcal{T}_0 := \{(e_1, e_2), (e_2, -e_1), (-e_1, -e_2), (-e_2, e_1)\}.$$

Lemma 7.22 (Lemma 2.3 in [Mir14b])**.**
- *Let $T = (u, v) \in \mathcal{T}$. The following are equivalent: (i) $T \in \mathcal{T}_0$, (ii) $\|u\| = \|v\|$, (iii) $S(T) < \Delta(T)$, (iv) T has no parent.*
- *The graph $(\mathcal{T}, \lhd)$ is the disjoint union of four complete infinite binary trees, whose roots lie in $\mathcal{T}_0$.*

The tree rooted in (e_1, e_2) is isomorphic to the classical Stern-Brocot tree [Niq07], an infinite binary tree labeled with rationals, via the mapping $(u, v) \mapsto p/q$ where $(p, q) = u + v$. Each positive rational appears exactly once as a label, in its irreducible form, as follows from the first statement of the next proposition. See also [Niq07].

Proposition 7.23. *For each $u \in \mathcal{Z}$ with $\|u\| > 1$, there exists a unique $(u_-, u_+) \in \mathcal{T}$ such that $u = u_- + u_+$. By convention we let $(u_-, u_+) := (-u^\perp, u^\perp)$ if $\|u\| = 1$. For any $u, v \in \mathcal{Z}$*

$$(u, v) \in \mathcal{T} \Leftrightarrow \exists k \geq 0,\ v = u_+ + ku, \qquad (v, u) \in \mathcal{T} \Leftrightarrow \exists k \geq 0,\ v = u_- + ku.$$

Furthermore, $\|u_\pm + ku\| > k\|u\|$ for all $k \geq 0$. Also, $\|u_\pm\| \leq \|u\|$ with equality iff $\|u\| = \|u_\pm\| = 1$.

Proof. See Proposition 1.2 in [Mir16] for the existence and uniqueness of (u_-, u_+).

The announced properties are obvious if $\|u\| = 1$, hence w.l.o.g. we assume $\|u\| > 1$. One has $\|u\|^2 = \|u_+\|^2 + 2\langle u_+, u_-\rangle + \|u_-\|^2 \geq \|u_+\|^2 + 0 + 1$, hence $\|u\| > \|u_+\|$ as announced, and likewise for u_-. One has $\|u_+ + ku\|^2 = k^2\|u\|^2 + 2k\langle u, u_+\rangle + \|u_+\|^2 \geq k^2\|u\|^2 + 0 + 1$ for all $k \geq 0$, hence $\|u_+ + ku\| > k\|u\|$ and likewise for u_- as announced.

If $(u, v) \in \mathcal{T}$, then $\det(u, v) = \det(u, u_+)$, hence $v = u_+ + ku$ for some $k \in \mathbb{R}$. Since u_+, v have integer coordinates, and u has co-prime coordinates, one has $k \in \mathbb{Z}$. By definition $0 \leq \langle u, v\rangle = \langle u, u_+\rangle + k\|u\|^2 < (k+1)\|u\|^2$, showing that $k \geq 0$ as announced. Likewise for u_-, and the reverse implication is obvious. □

7.3.1 Angular partitions

To each element $T = (u, v)$ of (our variant of) the Stern-Brocot tree one can associate an angular sector, whose width and covering properties are the object of this short subsection.

Lemma 7.24. *For all $(u, v) \in \mathcal{T}$ one has $(\|u\|\|v\|)^{-1} \leq \measuredangle(u, v) \leq \frac{\pi}{2}(\|u\|\|v\|)^{-1}$.*

Proof. One has $\sin(\measuredangle(u, v)) = \det(u, v)/(\|u\|\|v\|) = (\|u\|\|v\|)^{-1}$. Also, by concavity, one has $\frac{2}{\pi}\varphi \leq \sin\varphi \leq \varphi$ for all $\varphi \in [0, \pi/2]$, hence $t \leq \arcsin t \leq \frac{\pi}{2}t$ for all $t \in [0, 1]$. □

Definition 7.25. *Given $T = (u, v) \in \mathcal{T}$ we let $\Theta(T) := [\theta_u, \theta_v[$, where u is positively proportional to $(\cos\theta_u, \sin\theta_u)$ and $\theta_u \in [0, 2\pi[$, and likewise for v and $\theta_v \in]0, 2\pi]$.*

If $T \in \mathcal{T}_0$, then $\Theta(T) = [k\pi/2, (k+1)\pi/2[$ for some $0 \leq k \leq 3$. By construction, $\Theta(T) = \Theta(T') \sqcup \Theta(T'')$ if T' and T'' are the children of T, where $\sqcup$ denotes the disjoint union. In addition $|\Theta(T)| = \measuredangle(u, v)$ for all $T = (u, v) \in \mathcal{T}$.

Definition 7.26.

- *A sub-forest is a set $\mathcal{T}_* \subset \mathcal{T}$ which contains the parent, if any, of each of its elements: for all $T \lhd T'$ with $T' \in \mathcal{T}_*$ one has $T \in \mathcal{T}_*$.*
- *An outer leaf of $\mathcal{T}_*$ is an element of $\mathcal{T} \setminus \mathcal{T}_*$ whose parent, if any, lies in $\mathcal{T}_*$. An inner leaf of $\mathcal{T}_*$ is an element of $\mathcal{T}_*$ whose two children lie outside $\mathcal{T}_*$. Their sets are respectively denoted*

$$\mathcal{L}^o(\mathcal{T}_*) \subset \mathcal{T} \setminus \mathcal{T}_* \qquad\qquad \mathcal{L}^i(\mathcal{T}_*) \subset \mathcal{T}_*.$$

Said otherwise, an element $T \in \mathcal{T} \setminus \mathcal{T}_*$ (resp. $T \in \mathcal{T}_*$) is an outer leaf (resp. inner leaf) of a sub-forest $\mathcal{T}_* \subset \mathcal{T}$, iff $\mathcal{T}_* \cup \{T\}$ (resp. $\mathcal{T}_* \setminus \{T\}$) also is a sub-forest. In addition one easily checks that the angular sectors associated with the outer leaves define a partition of the angular space $[0, 2\pi[$, and that the angular sectors associated with the inner leaves are pairwise disjoint:

$$\bigsqcup_{T \in \mathcal{L}^o(\mathcal{T}_*)} \Theta(T) = [0, 2\pi[, \qquad\qquad \bigsqcup_{T \in \mathcal{L}^i(\mathcal{T}_*)} \Theta(T) \subset [0, 2\pi[. \tag{110}$$

7.3.2 Stencil construction

We show in Proposition 7.27 that *stencils* are in one to one correspondance with *finite sub-forests* of $\mathcal{T}$, see Definitions 7.3 and 7.26. This yields an efficient construction of stencils with minimal cardinality, and a way of counting their elements, see Corollary 7.28.

Proposition 7.27. *Let $(u_1, \cdots, u_n)$, $n \geq 4$, be a stencil in the sense of Definition 7.3, and let*

$$\mathcal{L}_* := \{(u_i, u_{i+1});\ 1 \leq i \leq n\}, \tag{111}$$

collect the pairs of consecutive elements, with $\alpha_{n+1} := \alpha_n$. Then $\mathcal{L}_*$ is the set of outer leaves of some finite sub-forest $\mathcal{T}_* \subset \mathcal{T}$, and in particular $\#\mathcal{L}_* = 4 + \#\mathcal{T}_*$. Any finite sub-forest $\mathcal{T}_*$ of $\mathcal{T}$ can be obtained in this way.

Proof. We proceed by induction on the cardinality of $\mathcal{L}_*$. For initialization, we note that $\#\mathcal{L}_* \geq 4$, with equality iff $\mathcal{L}_* = \mathcal{T}_0$, in which case it collects the outer leaves of the empty sub-forest $\mathcal{T}_* = \emptyset$. Otherwise denote $u = u_i$ the element of $\mathcal{L}_*$ with maximal norm, and observe that $u_{i+1} = u_+$ and $u_{i-1} = u_-$ by Proposition 7.23. Since $\mathcal{L}_* \subsetneq \mathcal{T}_0$ one has $\|u\| > 1$, and therefore $(u_{i-1}, u_{i+1}) = (u_-, u_+) \in \mathcal{T}$, showing that $(u_1, \cdots, u_{i-1}, u_{i+1}, \cdots, u_n)$ is also a stencil in the sense of Definition 7.3. Thus by induction $\mathcal{L}_* \cup \{(u_{i-1}, u_{i+1})\} \setminus \{(u_{i-1}, u_i), (u_i, u_{i+1})\} = \mathcal{L}^o(\mathcal{T}'_*)$ for some sub-forest $\mathcal{T}'_*$ of $\mathcal{T}$, and therefore $\mathcal{L}_* = \mathcal{L}^o(\mathcal{T}'_* \cup \{(u_{i-1}, u_{i+1})\})$ as announced.

Conversely, we observed in (110) that the set of outer leaves of a finite sub-forest of $\mathcal{T}$ defines a partition the angular space, and thus yields a stencil.

Recall that a finite complete rooted binary tree has one more outer leaf than inner nodes. Since $\mathcal{T}_*$ collects the inner nodes (possibly none) of four such trees, and $\mathcal{L}_*$ their leaves, one has $\#\mathcal{L}_* = 4 + \#\mathcal{T}_*$ as announced. □

Corollary 7.28. *Let F be an asymmetric norm, and let $\alpha \in]0, \pi/2]$. Define*

$$\mathcal{T}(F, \alpha) := \{(u, v) \in \mathcal{T};\ \measuredangle_F(u, v) > \alpha \text{ or } \measuredangle_F(v, u) > \alpha\}. \tag{112}$$

Then $\mathcal{T}(F, \alpha)$ is a finite sub-forest of $\mathcal{T}$, and $N(F, \alpha) = 4 + \#\mathcal{T}(F, \alpha)$.

Proof. The set $\mathcal{T}(F, \alpha)$ is a sub-forest of $\mathcal{T}$ by Lemma 7.8, and is finite by Proposition 7.9. Denote by $\mathcal{L}^o(F, \alpha)$ the collection of its outer leaves, and by $u_1, \cdots, u_n$ the corresponding stencil, see Proposition 7.27. One has $(u_i, u_{i+1}) \in \mathcal{L}^o(F, \alpha) \subset \mathcal{T} \setminus \mathcal{T}(F, \alpha)$, for any $1 \leq i \leq n$, implying the (F, α)-acuteness property (89) by definition of $\mathcal{T}(F, \alpha)$. This implies the upper bound $N(F, \alpha) \leq n = \#\mathcal{L}^o(F, \alpha) = 4 + \#\mathcal{T}(F, \alpha)$.

Conversely, let $u_1, \cdots, u_n$ be an (F, α)-acute stencil with minimal cardinality, and let $\mathcal{L}_*$ and $\mathcal{T}_*$ be as in Proposition 7.27. By Lemma 7.8, and recalling Definition 7.21, all elements of the set

$$\mathcal{E} := \{T' \in \mathcal{T};\ \exists T \in \mathcal{L}_*,\ T \preceq T'\}$$

obey the acuteness condition (89), hence $\mathcal{E} \subset \mathcal{T} \setminus \mathcal{T}(F, \alpha)$. On the other hand, one has $\mathcal{E} = \mathcal{T} \setminus \mathcal{T}_*$, hence $\mathcal{T}(F, \alpha) \subset \mathcal{T}_*$, which yields the lower bound $\#\mathcal{T}(F, \alpha) \leq \#\mathcal{T}_* = \#\mathcal{L}_* - 4 = N(F, \alpha)$. □

Thanks to the tree structure, the set $\mathcal{T}(F, \alpha)$ can easily be computed in practice, as well as the corresponding minimal (F, α)-acute stencil, by e.g. depth first search as in [Mir14b].

7.3.3 Cardinality of a sub-forest

We estimate the cardinality of a sub-forest of $\mathcal{T}$ based on the number of inner leaves and on their depth as measured by the function S, see Corollary 7.32 and Definition 7.20. The proof is based on a decomposition of the sub-forest into a disjoint union of chains. We state, without proof, a lower bound on the depth of the last element of a chain, which immediate follows from (109).

Lemma 7.29. *If $T_0 \lhd \cdots \lhd T_n$ is a chain in $\mathcal{T}$, then $S(T_n) \geq n\Delta(T_0)$.*

Definition 7.30. *Let $\mathcal{T}_*$ be a finite sub-forest of $\mathcal{T}$. Then $\mathcal{T}_*$ is the union of a finite family of chains $C_1, \cdots, C_I$, each denoted $C_i = \{T_0^i \lhd \cdots \lhd T_{n_i}^i\}$, and defined as follows:*

- *(Main loop, iteration variable: i the chain index) Choose an element T_0^i minimizing S in $\mathcal{T}_* \setminus C_0 \sqcup \cdots \sqcup C_{i-1}$. If this set is empty, then the algorithm ends.*
- *(Inner loop, iteration variable: k the chain element index) Consider the two children T', T'' of T_k^i. If both lie in $\mathcal{T}_*$, then define T_{k+1}^i as the one minimizing S (any in case of tie). If only one lies in $\mathcal{T}_*$, then define it as T_{k+1}^i. If none lies in $\mathcal{T}_*$ then the inner loop ends.*

Lemma 7.31. *With the notations and assumptions of Definition 7.30. The chains are disjoint and their number I is also the number of inner leaves of $\mathcal{T}_*$. Denote by $(u_i, v_i) = T_0^i$, $1 \leq i \leq I$, the first element of each chain. Then the vectors $\{u_i;\, 1 \leq i \leq I,\, \|u_i\| < \|v_i\|\}$ are pairwise distinct, and likewise $\{v_i;\, 1 \leq i \leq I,\, \|u_i\| > \|v_i\|\}$.*

Proof. Assume for contradiction that $T_k^i = T_l^j$ for some $0 \leq i < j \leq I$, $k \leq n_i$, $l \leq n_j$, where (i, j, k, l) is minimal for lexicographic ordering. By construction of the first element of each chain, one has $l \geq 1$. One has $k = 0$, since otherwise $T_{k-1}^i = T_{l-1}^j$ contradicting the minimality of (i, j, k, l). Thus $S(T_0^j) < S(T_l^j) = S(T_0^i)$, contradicting the definition of T_0^i.

By construction, the chains exhaust $\mathcal{T}_*$, are disjoint as shown in the above paragraph, and each one ends at an inner leaf. Hence their number is the number of inner leaves, as announced.

Assume that (u, v_i) and (u, v_j) are the first element of the chains C_i and C_j, with $\|u\| < \min\{\|v_i\|, \|v_j\|\}$ and $i < j$. Then $v_i = u_+ + ku$ and $v_j = u_+ + lu$ for some $1 \leq k < l$, by Proposition 7.23. Since $(u, u_+) \lhd \cdots \lhd (u, u_+ + lu)$, one has $(u, u_+ + ru) \in \mathcal{T}_*$ for all $0 \leq r \leq l$. The two children of $T = (u, u_+ + ru)$ are $T' = (u, u_+ + (r+1)u)$ and $T'' = (u_+ + (r+1)u, u_+ + ru)$, and satisfy $S(T'') - S(T') = \langle u_+ + (r-1)u, u_+ + (r+1)u \rangle > 0$ for all $r \geq 1$. Hence $T_0^j \in C_i$, by construction of C_i, see the inner loop, which is a contradiction. The result follows. □

Corollary 7.32. *Let $\mathcal{T}_*$ be a sub-forest of $\mathcal{T}$. Then for some absolute constant C*

$$\#\mathcal{T}_* \leq C \left(1 + \max_{T \in \mathcal{T}_*} S(T)\right) \ln \max\{2, \#\mathcal{L}^i(\mathcal{T}_*)\}.$$

Proof. Denote by $I := \#\mathcal{L}^i(\mathcal{T}_*)$ the number of inner leaves, and $s := \max\{S(T); T \in \mathcal{T}_*\}$ the depth of $\mathcal{T}_*$ as measured by S. By Lemma 7.31, $\mathcal{T}_*$ is the disjoint union of I chains,

with $n_1, \cdots, n_i$ elements, and whose first element we denote $(u_1, v_1), \cdots, (u_I, v_I)$. By Lemma 7.29 one has

$$\#\mathcal{T}_* = \sum_{1\le i\le I} n_i \le \sum_{1\le i\le I} \frac{s+1}{\min\{\|u_i\|, \|v_i\|\}^2} \le (s+1)\Big(4 + 2\sum_{1\le i\le I} \frac{1}{\|w_i\|^2}\Big), \tag{113}$$

where $(w_n)_{n\ge 1}$ is an enumeration of $\mathbb{Z}^2 \setminus \{0\}$ sorted by non-decreasing norm. In (113, r.h.s.) the constant 4 corresponds to the case $\|u_i\| = \|v_i\|$ and thus to chains rooted in $\mathcal{T}_0$ by Lemma 7.22. The sum comes from the cases $\|u_i\| < \|v_i\|$ or $\|u_i\| > \|v_i\|$ and from the injectivity property of Lemma 7.31. Observing that $\|w_I\| \le C\sqrt{I}$ and using (114, left) below, we conclude the proof. □

7.4 Complexity estimates

This section concludes the proof of Theorem 7.4, dealing with the worst case and average case complexity estimates in §7.4.1 and §7.4.2 respectively. Most of the material has been prepared in §7.2 and §7.3. The following elementary estimate serves in several occasions.

Lemma 7.33 (Lemma 2.7 in [Mir14b])**.** *For all $r \ge 2$, one has with C an absolute constant*

$$\sum_{\substack{0<\|u\|\le r\\ u\in\mathbb{Z}^2}} \frac{1}{\|u\|^2} \le C \ln r. \tag{114}$$

Corollary 7.34. *For any $r \ge 2$, one has with C an absolute constant*

$$\#\{(u,v) \in \mathcal{T}; \|u\|\|v\| \le r\} \le Cr \ln r \tag{115}$$

Proof. We distinguish the cases $\|u\| = \|v\|$, $\|u\| < \|v\|$, and $\|u\| > \|v\|$. In the first case one has $(u,v) \in \mathcal{T}_0$, see Lemma 7.22, so that the contribution of these terms is 4. Otherwise, assuming w.l.o.g. that $\|u\| < \|v\|$, one has $v = u_+ + ku$ for some $k \ge 1$, see Proposition 7.23. Therefore $\|v\| \ge k\|u\|$, thus $k \le r/\|u\|^2$, which is an upper bound for the number of possible choices of v for a given u. Eventually we conclude the proof using (114)

$$\#\{(u,v) \in \mathcal{T}; \|u\|\|v\| \le r\} - 4 \le 2\sum_{\|u\|\in\mathcal{Z}} \Big\lfloor \frac{r}{\|u\|^2} \Big\rfloor \le 2\sum_{0<\|u\|\le\sqrt{r}} \frac{r}{\|u\|^2} \le Cr\ln r. \qquad \square$$

7.4.1 Worst case

We establish the upper bound on the cardinality $N(F,\alpha)$ of a minimal (F,α)-acute stencil, announced in Theorem 7.4. The asymmetric norm F and parameter $\alpha \in]0, \pi/2]$ are fixed throughout this section.

Lemma 7.35. *$\mathcal{T}(F,\alpha)$ has at most $C\ln(\mu)/\alpha^2$ inner leaves, each obeying $S(T) \le 5\mu/\alpha^2$, where $\mu := \max\{2, \mu(F)\}$ and C is an absolute constant.*

Proof. The set $\mathcal{T}(F,\alpha)$ is introduced in Corollary 7.28, and the quantity $S(T)$ in Definition 7.20. Denoting by $\mathcal{L}^i(F,\alpha)$ the set of inner leaves of $\mathcal{T}(F,\alpha)$, see Definition 7.26, we obtain

$$\begin{aligned}\alpha^2 \#\mathcal{L}^i(F,\alpha) &\le \sum_{\substack{T\in\mathcal{L}^i(F,\alpha)\\ T=(u,v)}} \max\{\measuredangle_F(u,v),\measuredangle_F(v,u)\}^2\\ &\le C\int_0^{2\pi} \max\{1,\tan\psi_F^+,-\tan\psi_F^-\}\\ &\le C'\ln\mu.\end{aligned}$$

We successively used (i) the inclusion $\mathcal{L}^i(F,\alpha)\subset\mathcal{T}(F,\alpha)$ and definition (112) of $\mathcal{T}(F,\alpha)$, (ii) Proposition 7.19 and (110), (iii) the integral upper bound (107). The first announced point follows.

On the other hand, for each $T=(u,v)\in\mathcal{T}(F,\alpha)$ one has by (112) and Proposition 7.9

$$\alpha < \measuredangle_F(u,v) \le \sqrt{5\mu\measuredangle(u,v)},$$

and therefore since $\det(u,v)=1$

$$\alpha^2/(5\mu) \le \measuredangle(u,v) = \arctan(1/\langle u,v\rangle) \le 1/\langle u,v\rangle,$$

implying as announced that $S(T):=\langle u,v\rangle \le 5\mu/\alpha^2$. □

Corollary 7.36. *$\#\mathcal{T}(F,\alpha)\le C\frac{\mu}{\alpha^2}\ln(\frac{\ln\mu}{\alpha^2})$, with $\mu:=\max\{12,\mu(F)\}$ and C an absolute constant.*

Proof. The announced estimate immediately follows from Lemma 7.35 and Corollary 7.32. Note that $\frac{\ln\mu}{\alpha^2}\ge\frac{\ln(12)}{(\pi/2)^2}>1$. □

7.4.2 Average case

Throughout this section, we denote by F an asymmetric norm, which is continuously differentiable except at the origin. In the following, $\chi_{\ge1}:\mathbb{R}\to\{0,1\}$ denotes the indicator function of the set $[1,\infty[$.

Recall that $\mathcal{T}(F,\alpha)$ is a family of pairs (u,v) of vectors, playing symmetrical roles, see (112). Our first lemma breaks this symmetry, and lets u (or v) play a preferred role through the introduction of auxiliary sets $\mathcal{Z}_\sigma(F,\delta,u)$, for suitable $\delta\ge0$, $\sigma\in\{+,-\}$.

Definition 7.37. *For each $u\in\mathcal{Z}$, $\sigma\in\{+,-\}$, $\delta>0$, let $\mathcal{Z}_\sigma(F,\delta,u)$ collect all $v\in\mathcal{Z}$ such that*

$$|\tan\varphi_F(u)|\ge\delta\|u\|\|v\| \qquad \langle u,v\rangle\ge0,\quad \det(u,v)=\sigma 1. \tag{116}$$

Lemma 7.38. *Let $\delta=\alpha^2/C$, where C is from Proposition 7.14. Then $\mathcal{T}(F,\alpha)$ is a subset of*

$$\{(u,v)\in\mathcal{T};\alpha\|u\|\|v\|\le C\}\cup\{(u,v)\in\mathcal{T};v\in\mathcal{Z}_+(F,\delta,u)\}\cup\{(u,v)\in\mathcal{T};u\in\mathcal{Z}_-(F,\delta,v)\}.$$

Therefore for some absolute constant C'

$$\#\mathcal{T}(F,\alpha)\le\frac{C'}{\alpha}|\ln\alpha|+\sum_{u\in\mathcal{Z}}\sum_{\sigma\in\{+,-\}}\#\mathcal{Z}_\sigma(F,\delta,u). \tag{117}$$

Proof. Let $(u,v) \in \mathcal{T}(F,\alpha)$, so that $\measuredangle_F(u,v) > \alpha$ or $\measuredangle_F(v,u) > \alpha$, see (112). Then by Proposition 7.14, and recalling that $\|u\|\|v\|\measuredangle(u,v) \leq 1$ see Lemma 7.24, we obtain

$$\|u\|\|v\|\alpha^2 \leq C \max\{\frac{1}{\|u\|\|v\|}, |\tan\varphi_F(u)|, |\tan\varphi_F(v)|\}.$$

The announced inclusion of follows, implying the cardinality estimate by Corollary 7.34. □

The next lemma estimates the cardinality of each $\mathcal{Z}_\sigma(F,\delta,u)$ individually. Recall that $u_\pm$ is defined in Proposition 7.23.

Lemma 7.39. *For each $u \in \mathcal{Z}$, $\sigma \in \{+,-\}$, $\delta > 0$, one has $\mathcal{Z}_\sigma(F,\delta,u) = \emptyset$ if $\|u\| > \mu(F)/\delta$, and else*

$$\#\mathcal{Z}(F,\delta,u) \leq \frac{|\tan\varphi_F(u)|}{\delta\|u\|^2} + \chi_{\geq 1}\Big(\frac{|\tan\varphi_F(u)|}{\delta\|u\|\|u_\sigma\|}\Big). \tag{118}$$

Proof. From definition (116, right) we obtain $v = u_\sigma + ku$ for some $k \geq 0$. One has $\|u_\sigma + ku\| \geq \max\{\|u_\sigma\|, k\|u\|\}$, see Proposition 7.23, hence $k \leq (\tan\varphi_F(u))/(\delta\|u\|^2)$ which accounts for the first contribution in (118). The second contribution corresponds to the case $k = 0$.

Finally, if $\|u\| > \mu(F)/\delta$ then (116, left) yields $|\tan\varphi_F(u)| \geq \delta\|u\|\|v\| > \mu(F)$, since $\|v\| \geq 1$, in contradiction with $|\tan\varphi_F(u)| \leq \mu(F)$ see Lemma 98. This concludes the proof. □

In view of (117) and towards the average case estimate of $\mathcal{T}(F\circ R_\theta)$, where R_θ denotes the rotation of angle $\theta \in [0,2\pi]$, we consider the following integral. Let $0 < \delta \leq 1$ be fixed.

$$\begin{aligned}\sum_{u\in\mathcal{Z}}\int_0^{2\pi} \#\mathcal{Z}_\sigma(F\circ R_\theta,\delta,u)\,\mathrm{d}\theta &\leq \sum_{\|u\|\leq\mu(F)/\delta}\int_0^{2\pi} \frac{|\tan\varphi_F(R_\theta u)|}{\delta\|u\|^2} + \chi_{\geq 1}\Big(\frac{|\tan\varphi_F(R_\theta u)|}{\delta\|u\|\|u_\sigma\|}\Big)\,\mathrm{d}\theta \\ &= \sum_{\|u\|\leq\mu(F)/\delta}\int_0^{2\pi} \frac{|\tan\varphi_F(\theta)|}{\delta\|u\|^2} + \chi_{\geq 1}\Big(\frac{|\tan\varphi_F(\theta)|}{\delta\|u\|\|u_\sigma\|}\Big)\,\mathrm{d}\theta,\end{aligned} \tag{119}$$

where implicitly $u \in \mathcal{Z}$ in each of the sums. Recall that φ_F is defined both on non-zero vectors and on reals, by taking the argument see Definition 7.10, and that on $\mathbb{R}$ it is 2π-periodic.

The first contribution of (119) is separable w.r.t. θ and u, hence can be bounded as follows:

$$\sum_{\|u\|\leq\mu(F)/\delta}\int_0^{2\pi} \frac{|\tan\varphi_F(\theta)|}{\delta\|u\|^2}\mathrm{d}\theta = \frac{1}{\delta}\int_0^{2\pi} |\tan\varphi_F(\theta)|\mathrm{d}\theta \sum_{0<\|u\|\leq\mu(F)/\delta}\frac{1}{\|u\|^2} \leq \frac{C}{\delta}\ln(\mu)\ln(\frac{\mu}{\delta}), \tag{120}$$

where $\mu := \max\{2,\mu(F)\}$. We used Corollary 7.13 to upper bound the integral w.r.t. θ, and Lemma 7.33 for the summation over u.

In contrast, the second contribution in (119) is non-separable, motivating the following lemma.

Lemma 7.40. *For all $r \geq 2$, $\sigma \in \{+,-\}$, one has with C an absolute constant*

$$\sum_{u \in \mathcal{Z}} \chi_{\geq 1}\Big(\frac{r}{\|u\|\|u_\sigma\|}\Big) \leq Cr \ln r. \tag{121}$$

Proof. For each $u \in \mathcal{Z}$ one has $(u, u_+) \in \mathcal{T}$ and $(u_-, u) \in \mathcal{T}$. Hence (121) is bounded by the cardinality of $\{(u,v) \in \mathcal{T}; \|u\|\|v\| \leq r\}$, which is estimated in Corollary 7.34. □

The second contribution of (119) is bounded as follows, denoting $r(\theta) := \max\{2, |\tan\varphi_F(\theta)|/\delta\}$

$$\sum_{\|u\| \leq \mu(F)/\delta} \int_0^{2\pi} \chi_{\geq 1}\Big[\frac{|\tan\varphi_F(\theta)|}{\delta\|u\|\|u_\sigma\|}\Big] \mathrm{d}\theta \leq C \int_0^{2\pi} r(\theta) \ln r(\theta) \mathrm{d}\theta$$

$$\leq C \int_0^{2\pi} \max\{2, \frac{|\tan\varphi_F(\theta)|}{\delta}\} \ln(\frac{\mu}{\delta}) \,\mathrm{d}\theta \leq C \frac{\ln\mu}{\delta} \ln(\frac{\mu}{\delta}),$$

where we used successively (i) Lemma 7.40, (ii) the uniform upper bound $|\tan\varphi_F(\theta)| \leq \mu(F)$ see Lemma 7.11, and (iii) the L^1 estimate of $|\tan\varphi_F|$ established in Corollary 7.13. Together with (120), this proves that (119) is bounded by $C\frac{\ln\mu}{\delta}\ln(\frac{\mu}{\delta})$. In view of Lemma 7.38, this concludes the proof of Theorem 7.4.

7.A Semi-Lagrangian discretization of Finslerian eikonal equations

We present an elementary introduction to numerical methods for the computation of generalized traveltimes and distance maps, focusing on single pass semi-Lagrangian methods [Tsi95, KS98, SV03, BR06, AM12, Mir14b, Mir14a], at the expense of alternative approaches such as [LQ12, BR06], which is the context underlying of the problem studied in this paper. An open source code implementing this method is available on the author's webpage `github.com/Mirebeau`.

The main result of this section is Proposition 7.41 known as *acuteness implies causality* [SV03]. It requires that the numerical method be based upon strictly acute stencils, in the sense of Definition 7.3 with $\alpha < \pi/2$. Under this condition, one can compute an approximate travel time $T_h(x)$, at a given discretization point $x \in \Omega_h$ where h is the grid scale, in terms of suitable neighbor values $T_h(x + hu_i)$ and $T_h(x + hu_{i+1})$ no greater than $T_h(x) - h\varepsilon$, where $\varepsilon > 0$ is uniform over the domain. As a result, T_h can be efficiently computed in a single pass over the domain using the fast-marching algorithm, similar to Dijkstra's method on graphs, which deals with vertices in the order of increasing values of T_h. In addition let us mention that uniform causality, a.k.a. $\varepsilon > 0$, is a stable property which is also satisfied by suitably small perturbations of the numerical scheme, such as those related to second order accuracy [Set99] and to source factorization [LQ12].

Consider a bounded domain $\Omega \subset \mathbb{R}^2$, equipped with a Finslerian metric $\mathcal{F} : \overline{\Omega} \times \mathbb{R}^2$, $(x,u) \mapsto \mathcal{F}_x(u)$. In other words, $\mathcal{F}$ is a continuous mapping, and $\mathcal{F}_x(\cdot)$ is an asymmetric norm for each $x \in \overline{\Omega}$ in the sense of Definition 7.1. The Finslerian distance from x to $y \in \overline{\Omega}$ is defined as

$$d_{\mathcal{F}}(x,y) := \inf_{\gamma \in \Gamma_{x\to y}} \int_0^1 \mathcal{F}_{\gamma(t)}(\gamma'(t)) \,\mathrm{d}t, \quad \Gamma_{x\to y} := \{\gamma \in \mathrm{Lip}([0,1], \overline{\Omega});\, \gamma(0) = x,\, \gamma(1) = y\}.$$

One is interested in the distance from the boundary, $T(x) := \min\{d_{\mathcal{F}}(x,y);\, y \in \partial\Omega\}$ often referred to as the "escape time" from the domain, which under mild assumptions is the unique viscosity solution [BCD08] to the following (generalized) eikonal Partial Differential Equation (PDE), written in Bellman form:

$$\inf_{u\in S^1} \mathcal{F}_x(u) + \langle \nabla T(x), u\rangle = 0,\ \forall x \in \Omega, \qquad\qquad T(x) = 0,\ \forall x \in \partial\Omega. \tag{122}$$

Note that the PDE remains equivalent if the unit circle S^1 is replaced with any curve enclosing the origin. In particular, we can consider the closed polygonal line defined by a stencil, see Definition 7.3, possibly depending on $x \in \Omega$ and denoted $u_1(x), \cdots, u_{n(x)}(x)$ where $n(x) \geq 4$. In the following, the explicit dependency $u_i = u_i(x)$ w.r.t. the base point $x \in \Omega$ is often omitted readability, and by convention $u_{n(x)+1} := u_1$.

Consider a grid scale $h > 0$, and introduce the sets $\Omega_h := \Omega \cap h\mathbb{Z}^2$ and $\partial\Omega_h := (\mathbb{R}^2 \setminus \Omega) \cap h\mathbb{Z}^2$ devoted to the discretization of Ω and $\partial\Omega$. Semi-Lagrangian numerical schemes for the eikonal equation mimick (122) as follows: find $T_h : h\mathbb{Z}^2 \to \mathbb{R}$ such that

$$\min_{1\leq i\leq n(x)} \min_{s\in[0,1]} \mathcal{F}_x((1-s)u_i + su_{i+1}) + \frac{(1-s)T_h(x+hu_i) + sT_h(x+hu_{i+1}) - T_h(x)}{h} \tag{123}$$

equals 0 for all $x \in \Omega_h$, with again the boundary condition $T_h(x) = 0$ for all $x \in \partial\Omega_h$.

Proposition 7.41 (Acuteness implies causality [SV03])**.** *Assume that $u_1(x), \cdots, u_{n(x)}(x)$ is an $(\mathcal{F}_x, \alpha)$-acute stencil, where $\alpha \in]0, \pi/2[$. Assume also that* (123) *vanishes, and that the minimum is attained for some $1 \leq i \leq n(x)$ and $s \in]0,1[$. Then*

$$T_h(x) \geq h\cos(\alpha)\mathcal{F}_x(u_i) + T_h(x+hu_i), \quad T_h(x) \geq h\cos(\alpha)\mathcal{F}_x(u_{i+1}) + T_h(x+hu_{i+1}).$$

Proof. A standard analysis based on Lagrange's optimality conditions shows that

$$hA^T\nabla\mathcal{F}_x((1-s)u_i + su_{i+1}) + \begin{pmatrix} T_h(x+hu_i) \\ T_h(x+hu_{i+1}) \end{pmatrix} = T_h(x)\begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

where A is the matrix of columns u_i and u_{i+1}, see the Appendix of [SV03] or the Appendix of [Mir14b]. Considering this vector equality componentwise yields the announced result. □

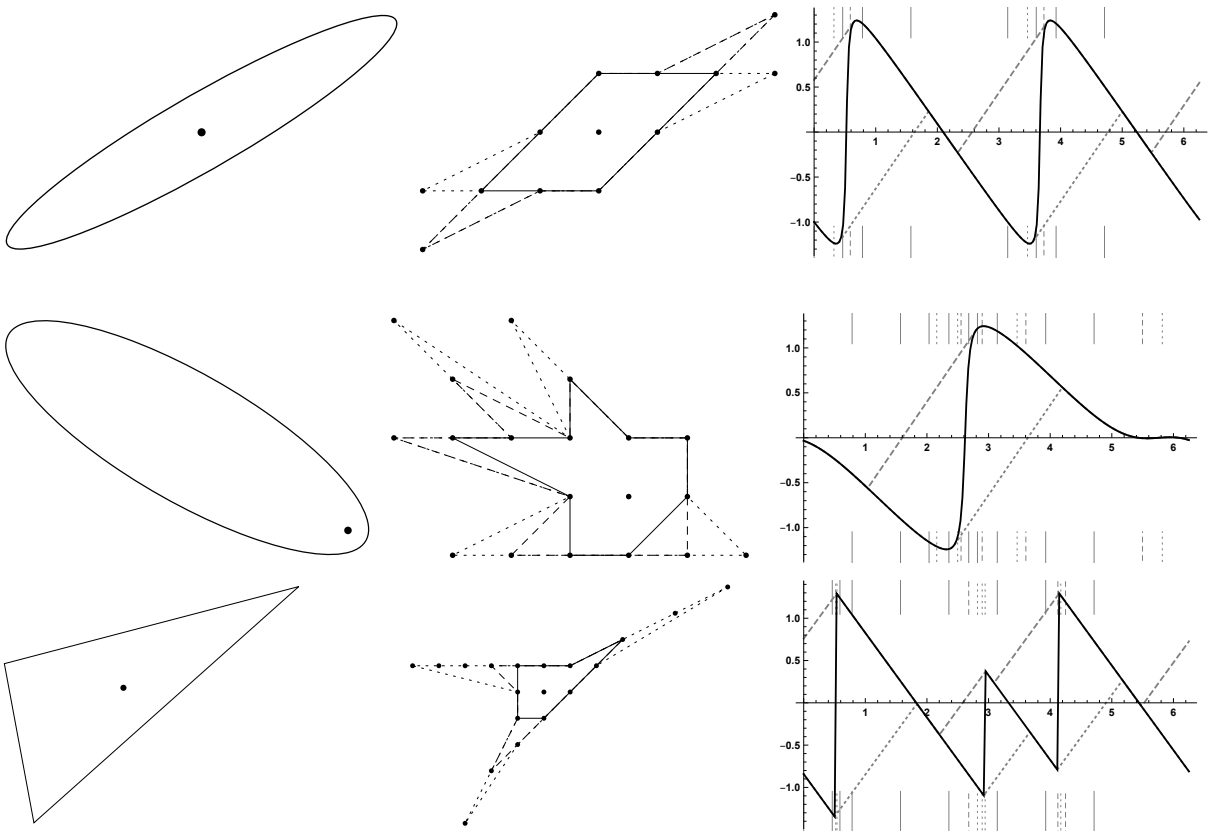

Figure 35: Left: Unit sphere $\{F = 1\}$ of a norm F, which is asymmetric in the second and third row. The origin is marked with a point. Center: Minimal (F, α)-acute stencil for $\alpha = \pi/2$ (solid), $\pi/3$ (dashed), $\pi/4$ (dotted). Right: Function φ_F (solid), ψ_F^+ (dashed, above), ψ_F^- (dotted, below). Vertical bars correspond to the angles of the stencil points.

Part III
Eulerian scheme for the eikonal equation

Contents

8 Single pass computation of first seismic wave travel time in three dimensional heterogeneous media for the TTI anisotropy [DMM22]

This section corresponds to the paper, currently submitted and under reviewing:

- François Desquilbet, Jean-Marie Mirebeau, and Ludovic Métivier. Single pass eikonal solver in tilted transversely anisotropic media. 2022

Abstract

We present a numerical scheme to solve the eikonal equation in a Tilted Transversely Isotropic (TTI) medium. The solution to this equation corresponds to the first arrival time of seismic pressure waves in the high frequency asymptotic regime, whose propagation speed is neither isotropic nor elliptic. Instead, the speed profile is characterized by a fourth degree polynomial equation in a rotated frame, defined in terms of the Thomsen or Hooke elasticity coefficients of the geophysical medium.

We show that TTI eikonal equations can be expressed as the maximum or minimum of a family of Riemannian eikonal equations, for which efficient discretizations are known. Based on this observation, we propose an original scheme that is causal, thus solvable in a single pass over the domain, and Eulerian, hence also mapping well to massively parallel architectures. Numerical experiments illustrate the method's accuracy, speed and robustness, on both a problem with analytical solution and a realistic synthetic instance, and compare a CPU with a GPU implementation, with the GPU being fifty times faster than the CPU implementation.

8.1 Introduction

The eikonal Partial Differential Equation (PDE) characterizes the first arrival time of a front, whose propagation speed is locally dictated by a metric. Classical examples include isotropic metrics, which define a propagation speed depending only on the position of the front, as well as Riemannian metrics, whose propagation speed also depends on the normal to the front according to an ellipsoidal profile. In this paper we focus on the more complex *tilted transversely isotropic* (TTI) model, which commonly accounts for the velocity profiles of seismic pressure waves in complex media [LBMV18]. Such anisotropy may originate from a variety of causes, at various physical scales: from the atomic layout in crystals, through the small scale layered structure of rocks produced by sedimentation, to homogenisation effects along geophysical fault lines [BC91, CMA$^+$20]. We introduce a numerical scheme to solve the eikonal equation in TTI media, which is both very general - able to handle anisotropy of arbitrary strength, and to include the effects of topography, see Remark 8.43 - and highly efficient - solvable in a single pass over the domain, and efficiently portable to massively parallel accelerators. The scheme requires a Cartesian discretization grid, involves adaptive discretization stencils designed using algorithmic geometry for greater efficiency, and relies on a characterization of the TTI speed profile as a union or an intersection of ellipsoids, depending on the PDE coefficients, see Figure 36. We establish the wellposedness of the method, including the scheme causality, monotony,

the quasi-convexity or quasi-concavity of the involved optimization problems, and the convergence analysis. Numerical experiments illustrate our results, and include a comparison of a CPU and a GPU implementation, the validation of second order accuracy in synthetic test cases achieved using source factorization and multi-scale computations, and the fast resolution of a large and realistic three dimensional instance.

For concreteness, let us readily state the PDE that is addressed in this paper. Denote by $\Omega \subset \mathbb{R}^3$ an open connected and bounded domain, by $\sigma = (a, b, c, d, e) \in C^0(\overline{\Omega}, \mathbb{R}^5)$ some coefficients which are subject to the admissibility condition described in Theorem 8.2, and by $R \in C^0(\overline{\Omega}, \mathrm{GL}_3(\mathbb{R}))$ a continuous field of invertible matrices. Our objective is to numerically compute the viscosity [BCD08] solution $u : \Omega \to \mathbb{R}$ of the following static first order Hamilton-Jacobi-Bellman PDE :

$$ap_r^4 + bp_z^4 + cp_r^2 p_z^2 + dp_r^2 + ep_z^2 = 1, \qquad \text{where } (p_x, p_x, p_z) = R\nabla u \text{ and } p_r^2 := p_x^2 + p_y^2, \tag{124}$$

on $\Omega \setminus \{q_0\}$, subject to the additional condition $u(q_0) = 0$ at a point source $q_0 \in \Omega$, and to outflow boundary conditions on $\partial\Omega$. The coefficients $\sigma = (a, b, c, d, e)$ are derived from the local geophysical properties of the medium, and define an anelliptic (non-Riemannian) speed propagation profile, see Section 8.A. The model is said *tilted* in view of the coordinate transformation R, which is usually a rotation, and *transversely isotropic* in view of symmetry in p_x and p_y. A Riemannian, or elliptic, geometry is recovered in the special case where the equation is quadratic, i.e. $a = b = c = 0$. Following the geophysical terminology, we refer to (124, left) as the P-SV equation and note that it defines two *slowness* surfaces, see Figure 36, corresponding to the pressure and vertical shear wave propagation. Only the inner surface, associated with pressure waves which are the fastest, is considered in this paper.

Our numerical approach involves rephrasing the highly non-linear eikonal PDE (124), defined by a fourth degree polynomial, as a maximum or a minimum of a varying family of Riemannian eikonal PDEs, defined by quadratic polynomials and for which efficient numerical schemes have been developed [Mir19], see Figure 36 and Figure 8.1.1. For this reason, this work is related to multi-stencil fast marching methods [HF07], and more generally to discretizations of Hamilton-Jacobi-Bellman equations written in the Bellman extremal form [BBM20, BM21], see Section 8.4. For concreteness, we readily state our numerical scheme, which is presented in more detail in Sections 8.1.1, 8.1.2 and 8.1.3: find the solution to a finite differences equation denoted $\mathfrak{F}u = 1$, whose unknown $u : \Omega_h \to \mathbb{R}$ is discretized on the Cartesian grid $\Omega_h := \Omega \cap h\mathbb{Z}^d$ of scale $h > 0$ with the appropriate boundary conditions, and where

$$\mathfrak{F}u(q) := \underset{\alpha \in [\alpha_*, \alpha^*]}{\mathrm{mix}} \frac{1}{\mu(\alpha)} \sum_{1 \leq i \leq I} \rho_i(\alpha) \max\big\{0, \frac{u(q) - u(q + he_i)}{h}, \frac{u(q) - u(q - he_i)}{h}\big\}^2. \tag{125}$$

The notation "mix" stands for the (max) or (min) operator, if the P slowness surface is obtained as an *intersection* or as a *union* of ellipses respectively, see Figure 36. The optimization interval $[\alpha_*, \alpha^*]$ and multiplier $\mu(\alpha) > 0$, which are related to the anisotropy bounds and dilation coefficients of the ellipses respectively, have explicit algebraic expressions presented in Section 8.1.1. The weights $\rho_i(\alpha) \geq 0$ and offsets $e_i \in \mathbb{Z}^d$ are obtained

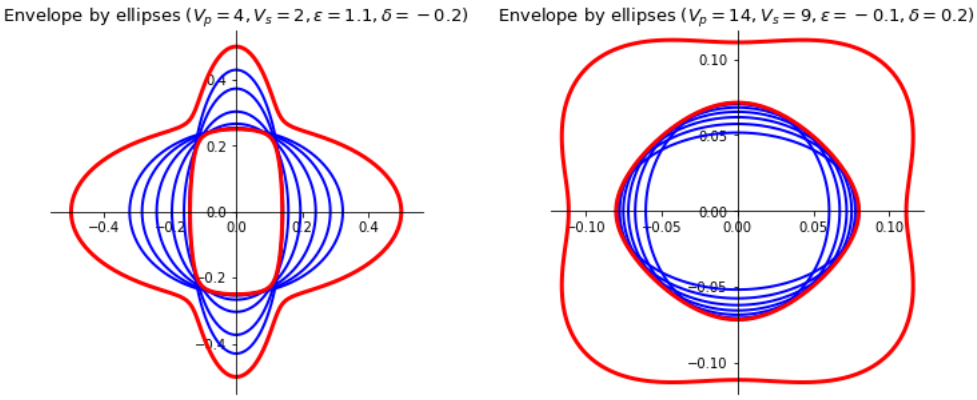

Figure 36: Slowness surfaces (red) defined by equation (124, left), in the (p_r, p_z) plane. The coefficients (a, b, c, d, e) are derived from the supplied Hooke parameters $(V_p, V_s, \varepsilon, \delta)$. Only the inner slowness surface is considered, and our numerical method involves its approximation by an intersection of ellipses (left) or a union of ellipses (right), shown blue. Subfigures (left) and (right) correspond respectively to the (max) and (min) alternative cases in (125) and Theorem 8.3.

using Selling's matrix decomposition [Sel74], similarly to the Riemannian scheme [Mir19], see Section 8.1.2. The scheme properties, and two strategies for the numerical optimization over $\alpha \in [\alpha_*, \alpha^*]$, are investigated in Section 8.1.3. The parameters $\alpha_*, \alpha^*, \mu(\alpha), \rho_i(\alpha)$ are derived from the coefficients a, b, c, d, e, R of the TTI eikonal PDE (124), and similarly they implicitly depend on the position $q \in \Omega$ according to the variations of the medium in the domain, see Remark 8.1.

The general framework of this study is seismic imaging: inferring the geophysical properties of the subsurface and its structure from the physical recordings of the seismic waves. In this context, accessing to travel-times is crucial in many steps of the workflow: macro-velocity model building through tomography, high resolution reflectivity estimation through migration techniques, and quality control along different stages of waveform based inversion techniques to estimate the time-shifts between recorded and simulated data. For this reason, efficient numerical methods to solve the high frequency asymptotics of the elastic wave equations in 3D general media are of particular interest for the seismic imaging community.

Distance maps are ubiquitous in mathematics and their applications, hence a sustained research effort and a wide variety of methods have been developed for their computation. The solution to the eikonal PDE (124) falls in this framework, since it admits an interpretation as the geodesic distance map from the source point q_0 and with respect to a metric defined in terms of the parameters (a, b, c, d, e) and R, see Remark 8.5. Graph based methods can compute distances while avoiding the PDE formalism [CHK13], but they often lack the stability and high order accuracy required for seismic imaging applications. Some other approaches exploit indirect connections with different PDEs, such as the heat method which is based on the small time asymptotics of the heat or Poisson kernels [CWW13]; however, it is limited to metrics featuring a quadratic structure, and

does not appear to scale well to three dimensional problems. Yet another approach to the computation of geodesics and geodesic distances is ray tracing [Sla03].

In the rest of this discussion we limit our attention to numerical methods which solve directly the eikonal PDE, either the one (124) associated to the TTI geometry, or a variant defined by another metric structure. We categorize these methods based on two criteria: *causal (single-pass) vs multi-pass solvers*, and *Eulerian vs semi-Lagrangian* approximation schemes - the method here proposed being causal and Eulerian. The first distinction is tied to a property, referred to as causality, of the coupled system of non-linear equations arising from the PDE discretization.

- Causal, single pass methods, are often referred to as fast marching methods (FMM), see Algorithm 5. The causality property is the translation at the discrete level of a principle underlying the front propagation: the front arrival time at a given point only depends on *earlier* arrival times, see Definition 8.6. Common advantages of FMMs include faster computation times (on sequential machines), easier back propagation (thanks to the triangular structure of the Jacobian of the scheme), and opportunities for modification (adaptive stopping criteria, high order schemes, etc), see the discussion in [DCC$^+$21]. Originally limited to isotropic eikonal equations [Set96], FMMs have been generalized to a variety of metrics [SV03, Mir14b, Mir18, Mir19, DCC$^+$21].

- Multi-pass methods rely on fast sweeping [Zha05], the fast iterative method [FKW13], or adaptive Gauss-Seidel iterations [BR06], to solve the discretized PDE. These iterative methods miss some of the advantages of FMMs, but also avoid the severe constraints associated with the design of a causal scheme. This shift in compromises enables a wider variety of numerical approaches, and thus possibly (if properly exploited) methods with narrower stencils, or addressing more complex geometries. They are also easier to parallelize.

The second distinction is between Eulerian and semi-Lagrangian PDE discretization schemes.

- Eulerian discretizations use finite differences (or finite elements, possibly discontinuous) to approximate the derivatives of the unknown arrival time function, and to produce a consistent approximation of the eikonal PDE operator. A variety of Eulerian schemes have been developed, for isotropic [Set96, HF07], Riemannian [Mir19], TTI [LBMV18], and curvature penalized [Mir18] geometry models.

- Semi-Lagrangian schemes mimic Bellman's optimality principle at the discrete level, which is derived from the shortest path interpretation of the solution to the eikonal equation, see e.g. [BR06, SV03, Mir14b, Mir14a]. These methods require maintaining a complex neighborhood structure around each point, and for this reason implementing them on GPUs, while feasible [FKW13], is more cumbersome and usually less efficient than for Eulerian schemes. In addition, implementing semi-Lagrangian schemes for TTI and related models in seismology requires solving complicated algebraic equations, due to the high degree of the PDEs (124) or (196), such as an

optimization problem subject to a polynomial constraint of degree six in three variables in [DCC+21]. This has a significant computational cost and typically requires double precision floating point arithmetic for stability.

Note that in the context of Eulerian methods, causality can be rephrased as a structural constraint on the numerical scheme, see Definition 8.6. Only few Eulerian schemes obey this condition beyond the standard isotropic one [Set96, Mir19, Mir18] and in particular the discretization of the TTI eikonal PDE proposed in [LBMV18] is not causal, in contrast to the one presented in this paper. For comparison, causality is in the context of semi-Lagrangian schemes is equivalent to a geometric acuteness property of the stencils [SV03], which can be ensured by refinement in two dimensions [KS98, Mir14b], but in three dimensions either requires a Riemannian structure [Mir14a], or poses a limit on the strength of the anisotropy [DCC+21], or requires impractically large stencils [SV03].

Summary of contributions. We present a *causal* and *Eulerian* discretization of the TTI eikonal PDE. Our approach is based on a new methodology, expressing the TTI speed profile as a union or an intersection of ellipsoids, established in Theorem 8.3. The scheme update operator is defined by a one-dimensional optimization problem, which can be efficiently solved numerically thanks to a quasi-convexity or quasi-concavity property established in Theorem 8.11. A proof of convergence of the scheme numerical solutions is presented in Theorem 8.33, together with regularity and growth estimates established using some fine properties of Selling's decomposition, which is a tool from discrete geometry involved in the scheme construction. The resulting scheme can be solved in a single pass over the domain using a CPU solver, but a massively parallel GPU solver is also demonstrated. Numerical experiments include a smooth synthetic test case for validating the scheme accuracy, as well as a large realistic instance.

Paper organization. The rest of this introduction is organized as follows: we present in Section 8.1.1 a reformulation of the TTI metric as an extremum of a family of Riemannian metrics, we recall in Section 8.1.2 an efficient discretization of the Riemannian eikonal PDE, and we combine in Section 8.1.3 the previous elements to obtain a discretization of the TTI eikonal equation (124). Beyond the introduction, the rest of this paper is organized as follows. Section 8.2 establishes the results, announced in Section 8.1.1, relating TTI and Riemannian geometry. We prove in Section 8.3 a property of the update operator of our scheme, announced in Section 8.1.3, which makes it numerically easy to evaluate. We establish in Section 8.4 some regularity and growth estimates for the scheme solutions, and prove their convergence to the viscosity solution of the PDE (124). Numerical experiments are presented in Section 8.5. Section 8.A describes Hooke tensors, Thomsen's elastic parameters, and their relation to the coefficients of (124). Section 8.B describes Selling's decomposition, a tool from discrete geometry used in the design of our numerical scheme in Section 8.1.3. Section 8.C discusses various heuristic enhancements designed to improve the accuracy of the scheme solutions.

Remark 8.1 (Varying material coefficients)**.** *For readability, we present in Section 8.1.1, Section 8.1.2 and Section 8.1.3 the construction of our numerical scheme in the setting where the coefficients $\sigma = (a, b, c, d, e)$ and the linear transformation R defining the eikonal*

PDE (124) *are fixed over the domain* Ω. *It must be clear however that, in the intended applications including the numerical experiments in Section 8.5, the parameters* $\sigma : \Omega \to \mathbb{R}^5$ *and* $R : \Omega \to \mathrm{GL}_3(\mathbb{R})$ *vary over the domain, and the definitions below are applied independently at each discretization point.*

Notations. We denote by $\mathrm{CC}_x(X)$ the connected component of a point x in the topological set X. Position variables are usually named q, impulsions named p, and velocities named v. The set of non-negative reals is denoted $\mathbb{R}_+ := [0, \infty[$. Let $\langle \cdot, \cdot \rangle$ denote the Euclidean scalar product, $|\cdot|$ the Euclidean norm. S_d^{++} stands for the set of symmetric positive definite matrices of shape $d \times d$, and we let $\|v\|_D := \sqrt{\langle v, Dv \rangle}$ for any $D \in S_d^{++}$, $v \in \mathbb{R}^d$.

8.1.1 The eikonal equation associated to a TTI model

We study in this subsection the algebraic structure of the TTI eikonal PDE (124): we characterize in Theorem 8.2 a family of coefficients for which it is physically meaningful and mathematically well posed, and we reformulate in Theorem 8.3 and Corollary 8.4 the PDE operator in a form related to the Riemannian setting that is amenable to discretization. Theorem 8.2 and Theorem 8.3 are proved in Section 8.2, and Corollary 8.4 is established in Section 8.3.2.

Our first step is to disambiguate the PDE (124), by distinguishing the role of the inner slowness surface. For that purpose we introduce given coefficients $\sigma = (a, b, c, d, e) \in \mathbb{R}^5$ the quadratic function $\mathcal{Q}_\sigma$ and the set $\mathcal{B}_\sigma$ defined as follows

$$\mathcal{Q}_\sigma(r, z) := ar^2 + bz^2 + crz + dx + ez, \tag{126}$$

$$\mathcal{B}_\sigma := \mathrm{CC}_0\{(p_x, p_y, p_z) \in \mathbb{R}^3; \mathcal{Q}_\sigma(p_x^2 + p_y^2, p_z^2) \leq 1\}. \tag{127}$$

By considering only the connected component of the origin, denoted CC_0 in (127), we obtain that ∂B_σ is the inner slowness surface defined by $\mathcal{Q}_\sigma$, as illustrated on Figure 36. We assume that the coefficients σ are admissible in the sense of Theorem 8.2 below, in such way that the set $\mathcal{B}_\sigma$ is compact and convex, and therefore this construction is well posed and physically meaningful. As a result, the eikonal PDE (124) can be reformulated in an unambiguous way

$$\mathcal{F}_\sigma^*(R\nabla u) = 1, \qquad \text{where } \mathcal{F}_\sigma^*(p) := \min\{\nu > 0;\, p/\nu \in \mathcal{B}_\sigma\}, \tag{128}$$

with $\mathcal{F}_\sigma^*(0) := 0$ by convention. In the geophysical context, the reformulation (128) characterizes the travel-time of the pressure wave (the fastest wave), and disregards the shear wave. Note that the eikonal PDE (128) is imposed on $\Omega \setminus \{q_0\}$, similarly to (124), and is combined with the point source constraint $u(q_0) = 0$ and outflow boundary conditions on $\partial\Omega$. The PDE (128) admits a unique viscosity solution [BCD08], provided[7] the parameters σ and R vary continuously over the domain $\overline{\Omega}$ and obey the admissibility condition

[7]Indeed, these conditions imply that $R(z)^{-1}\mathcal{B}_{\sigma(z)}$ is a convex and compact neighborhood of the origin, depending continuously on $z \in \overline{\Omega}$, so that the theory of viscosity solutions for optimal control problems is applicable.

described below in equations (130) and (131). The solution $u(q)$ at $q \in \Omega$ can be characterized as the minimal path length from the source point q_0, as measured by the Finsler metric $\mathcal{F}_\sigma$ dual to $\mathcal{F}_\sigma^*$, see Remark 8.5.

A fundamental special case of TTI geometry is when the equation is derived from the coefficients $(c_{11}, c_{13}, c_{33}, c_{44})$ of a Hooke tensor with the appropriate hexagonal symmetry, see Section 8.A. In this case the eikonal equation (124) reads:

$$-c_{11}c_{44}p_r^4 - c_{33}c_{44}p_z^4 + (2c_{13}c_{44} + c_{13}^2 - c_{11}c_{33})p_r^2 p_z^2 + (c_{11} + c_{44})p_r^2 + (c_{33} + c_{44})p_z^2 = 1, \quad (129)$$

with $p_r^2 := p_x^2 + p_y^2$. In other words, $\mathcal{Q}_\sigma(p_r^2, p_z^2) = 1$ where the parameters $\sigma = (a, b, c, d, e)$ can be recovered by identification with (124), namely

$$\sigma = (-c_{11}c_{44},\ -c_{33}c_{44},\ 2c_{13}c_{44} + c_{13}^2 - c_{11}c_{33},\ c_{11} + c_{44},\ c_{33} + c_{44}). \quad (130)$$

Our first result uses the algebraic properties of Hooke tensors and a criterion on the coefficients to ensure that the ball $\mathcal{B}_\sigma$ is well shaped. Some limit cases are illustrated on Figure 40.

Theorem 8.2. *Define $C_{\text{adm}} \subset \mathbb{R}^4$ as the set of Hooke tensor coefficients $c_{11}, c_{13}, c_{33}, c_{44}$ obeying*

$$c_{11} > c_{44}, \qquad c_{33} > c_{44}, \qquad c_{44} > 0, \qquad c_{13} + c_{44} > 0, \qquad c_{11}c_{33} > c_{13}^2, \quad (131)$$

which is open and convex. The PDE coefficients $\sigma \in \mathbb{R}^5$ are said admissible if they take the form (130) for some $(c_{11}, c_{13}, c_{33}, c_{44}) \in C_{\text{adm}}$, and in that case the ball $\mathcal{B}_\sigma$ is compact and convex.

Geophysical elasticity properties are also often described through the Thomsen parameters $(V_P, V_S, \varepsilon, \delta)$, but this turns out to be equivalent to specifying $(c_{11}, c_{13}, c_{33}, c_{44})$, with explicit conversion formulas between the parameters, see Section 8.A. We checked that all the material elasticity parameters listed in [Tho86] obey the admissibility condition. In addition, the convexity of the set C_{adm} means that admissibility is preserved when one 'interpolates' between those materials, hence the criterion of Theorem 8.2 does not appear to be excessively restrictive for applications in geophysics. As discussed in Remark 8.24, the eikonal PDE (124) could be studied under weaker assumptions, but in that case the slowness surfaces may not be separated see Figure 40, the scheme would need to be adapted and the solution u may have lower regularity. See also Remark 8.20 on the condition $c_{13} + c_{44} > 0$.

In order to design our numerical scheme, we need a description of $\mathcal{F}_\sigma^*$ more tractable than (128, right). For that purpose, we consider the following set, illustrated on Figure 37

$$\mathcal{A}_\sigma := \text{CC}_0\{(h_r, h_z) \in \mathbb{R}_+^2;\ \mathcal{Q}_\sigma(h_r, h_z) \leq 1\}, \quad (132)$$

which corresponds to the change of variables $h_r = p_r^2 = p_x^2 + p_y^2$ and $h_z = p_z^2$ in (129). The following result shows that $\mathcal{A}_\sigma$ is either a union or an intersection of triangular regions.

Theorem 8.3. *Let $\sigma \in \mathbb{R}^5$ be admissible. Then there exists $0 < \alpha_* \leq \alpha^* < 1$ and $\mu \in C^\infty([\alpha_*, \alpha^*],]0, \infty[)$ such that one of the following "max" and "min" cases holds:*

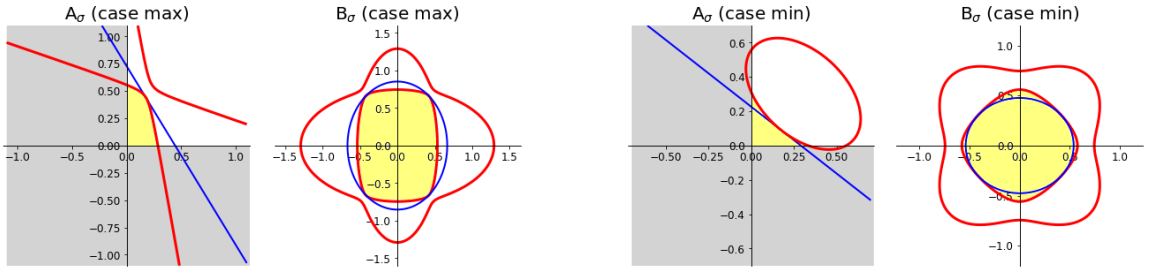

Figure 37: The set $\mathcal{A}_\sigma \subset \mathbb{R}^2_+$ (resp. $\mathcal{B}_\sigma \subset \mathbb{R}^2$ as defined from (p_r, p_z)), in yellow, is bounded by a conic curve (resp. a quartic curve), in red. Tangent lines to $\partial\mathcal{A}_\sigma$ correspond to tangent ellipses to $\partial\mathcal{B}_\sigma$, in blue. If the conic curve defines a convex (resp. concave) boundary, in case (max) see left (resp. case (min) see right) then the ellipses are exterior (resp. interior) tangent.

(max) μ *is convex, and* $\mathcal{A}_\sigma = \{(h_r, h_z) \in \mathbb{R}^2_+;\ \forall \alpha \in [\alpha_*, \alpha^*],\ (1-\alpha)h_r + \alpha h_z \leq \mu(\alpha)\}$.

(min) μ *is concave, and* $\mathcal{A}_\sigma = \{(h_r, h_z) \in \mathbb{R}^2_+;\ \exists \alpha \in [\alpha_*, \alpha^*],\ (1-\alpha)h_r + \alpha h_z \leq \mu(\alpha)\}$.

We deduce that the TTI unit ball (127) can be obtained as a union or an intersection of ellipsoids, depending on the alternative of Theorem 8.3 and as illustrated on Figure 36 and Figure 37 :

$$
\begin{aligned}
(max) : \mathcal{B}_\sigma &= \bigcap_{\alpha \in [\alpha_*, \alpha^*]} E(\alpha), \\
(min) : \mathcal{B}_\sigma &= \bigcup_{\alpha \in [\alpha_*, \alpha^*]} E(\alpha), \\
E(\alpha) &:= \{(1-\alpha)(p_x^2 + p_y^2) + \alpha p_z^2 \leq \mu\}.
\end{aligned}
$$

The denomination (max) and (min) reflects the expression of the numerical scheme (125) that we eventually obtain. The treatment of the (max) and (min) cases is remarkably symmetric in the results presented below, despite some technical differences in the proof of Corollary 8.4 and more notably in the convergence analysis of Theorem 8.33. In order to favor a unified treatment, we thus introduce a notation "mix" which stands for either "max" or "min", according to the alternative in Theorem 8.3. However, because a number of mathematical operations (such as negation or Legendre-Fenchel duality) turn a maximum into a minimum, and conversely, we also need to introduce a complementary $\overline{\text{mix}}$ notation. Summarizing, we denote

$$
\text{Case (max):}\ \ \text{mix} := \max\ \text{ and }\ \overline{\text{mix}} := \min\,. \quad \text{Case (min):}\ \ \text{mix} := \min\ \text{ and }\ \overline{\text{mix}} := \max\,. \tag{133}
$$

The proof of Theorem 8.3, presented in Section 8.2, relies on the observation that the boundary of $\mathcal{A}_\sigma$ is defined by a portion of the conic curve $\mathcal{C}_\sigma := \{(h_z, h_z) \in \mathbb{R}^2;\ \mathcal{Q}_\sigma(h_z, h_z) = 1\}$. The function μ and the bounds $\alpha_* \leq \alpha^*$ admit a simple closed form expression, established in (167), which is welcome for implementation purposes, but is not particularly

enlightening for the mathematical analysis. Denoting $\boldsymbol{\alpha} := (1-\alpha, \alpha)$ and rewriting the quadratic function (126) as $\mathcal{Q}_\sigma(p) = \langle l, p\rangle + \frac{1}{2}\langle p, Qp\rangle$ one has

$$\mu(\alpha) = \varepsilon\sqrt{\langle \boldsymbol{\alpha}, Q^{-1}\boldsymbol{\alpha}\rangle(2 + \langle l, Q^{-1}l\rangle)} - \langle \boldsymbol{\alpha}, Q^{-1}l\rangle,$$
$$\{\alpha_*, \alpha^*\} = \{\frac{l_2 + Q_{12}x}{l_1 + l_2 + Q_{11}x + Q_{12}x}, \frac{l_2 + Q_{22}y}{l_1 + l_2 + Q_{12}y + Q_{22}y}\},$$

where $\varepsilon := \operatorname{sign}(2 + \langle l, Q^{-1}l\rangle)$, where x solves $l_1 x + \frac{1}{2}Q_{11}x^2 = 1$, and $l_2 y + \frac{1}{2}Q_{22}y^2 = 1$; both x and y are the smallest root of their defining quadratic equation. An alternative expression of μ applies when Q is degenerate (168), see Section 8.2.3.

As a consequence of Theorem 8.3, we obtain a new expression of the TTI eikonal equation (128) operator $\mathcal{F}^*_\sigma(R\,\cdot)$ as an extremum of Riemannian norms (134). We also derive its gradient and dual norm (135), which are involved in the shortest path interpretation of the PDE, see Remark 8.5.

Corollary 8.4. *Let $\sigma \in \mathbb{R}^5$ be admissible, and let $R \in \mathrm{GL}_3(\mathbb{R})$. Denote by $\mathrm{mix} \in \{\max, \min\}$ the corresponding case of Theorem 8.3. Then for any $p \in \mathbb{R}^3$*

$$\mathcal{F}^*_\sigma(Rp) = \min_{\alpha\in[\alpha_*,\alpha^*]} \mu(\alpha)^{-\frac{1}{2}}\|p\|_{D(\alpha)}, \quad \text{where } D(\alpha) := R^\top \begin{pmatrix} 1-\alpha & & \\ & 1-\alpha & \\ & & \alpha \end{pmatrix} R. \tag{134}$$

Introducing the norm $\mathcal{F}^*(p) := \mathcal{F}^*_\sigma(Rp)$, we have the following expressions of its gradient at $p \neq 0$, and of the dual norm defined as $\mathcal{F}(v) := \max\{\langle p, v\rangle;\ \mathcal{F}^*(p) \le 1\}$

$$\nabla\mathcal{F}^*(p) = \mu(\alpha')^{-\frac{1}{2}}\frac{D(\alpha')p}{\|p\|_{D(\alpha')}}, \qquad \mathcal{F}(v) = \overline{\min_{\alpha\in[\alpha_*,\alpha^*]}}\, \mu(\alpha)^{\frac{1}{2}}\|v\|_{D(\alpha)^{-1}}, \tag{135}$$

where α' in (135, left) is the optimal parameter in (134, left), and where $\{\mathrm{mix}, \overline{\mathrm{mix}}\} = \{\min, \max\}$.

Remark 8.5 (Shortest path interpretation of the TTI eikonal PDE)**.** *We assume in this remark that the TTI parameters σ and R vary continuously on the PDE domain. Denote $\mathcal{F}^*_q(p) := \mathcal{F}^*_{\sigma(q)}(R(q)p)$ for all $q \in \overline{\Omega}$ and $p \in \mathbb{R}^d$, and likewise the dual norm $\mathcal{F}$. Then the unique viscosity solution to the eikonal equation* (128) *is the geodesic distance map from the source point q_0 [BCD08]:*

$$u(q) = \min\big\{\,\mathrm{length}_{\mathcal{F}}(\gamma);\ \gamma(0) = q_0, \gamma(1) = q\big\}, \qquad \mathrm{length}_{\mathcal{F}}(\gamma) := \int_0^1 \mathcal{F}_{\gamma(t)}(\gamma'(t))\mathrm{d}t,$$

where the infimum is over Lipschitz paths $\gamma : [0,1] \to \overline{\Omega}$ with the given endpoints. Conversely, the optimal path can be obtained from the value function by solving the backtracking ODE

$$\gamma'(t) = V(\gamma(t)) \qquad \text{where } V(q) := \nabla\mathcal{F}^*_q(\nabla u(q)), \tag{136}$$

backwards in time*, with terminal boundary condition $\gamma(T) = q$ where $T = u(q)$. Numerically the geodesic flow V is estimated in an upwind manner, using (135, left) and adapting the Riemannian case presented in [MP19, Section 3.2.1].*

8.1.2 Monotony, causality, and the fast marching algorithm

We recall in this subsection the concept of finite difference scheme $\mathfrak{F}$, and two key structural properties known as discrete degenerate ellipticity[8] (DDE) and causality that enable a stable and fast numerical solution, see Definition 8.6. We also introduce Selling's matrix decomposition, see Proposition 8.7, a tool from the field of discrete geometry previously used for the discretization of the Riemannian eikonal PDE [Mir19].

Definition 8.6. *Let $X, \overline{X}$ be finite sets with $X \subset \overline{X}$. Consider a finite difference scheme $\mathfrak{F}$ on X taking the form*

$$\mathfrak{F}u(q) := \hat{\mathfrak{F}}(q, [u(q) - u(r)]_{r \in \overline{X}}), \tag{137}$$

where $q \in X$, $u \in \mathbb{R}^{\overline{X}}$, and $\hat{\mathfrak{F}} : X \times \mathbb{R}^{\overline{X}} \to \mathbb{R}$ is continuous. The scheme $\mathfrak{F}$ is said:

- Discrete Degenerate Elliptic *(DDE), if $\hat{\mathfrak{F}}$ is non-decreasing w.r.t. its second argument.*
- Causal*, if $\hat{\mathfrak{F}}$ only depends on the positive part of its second argument.*

The DDE property implies a comparison principle for the equation $\mathfrak{F}u = 1$, and thus plays a key role in ensuring the stability and convergence of the scheme solutions, following techniques introduced in [Obe06]. Causality, on the other hand, enables the Fast Marching Method (FMM) to solve the system $\mathfrak{F}u = 1$; this property was initially used in [Set96], and formalized as above in [Mir19]. These consequences are summarized in [Mir19, Theorem 2.3], and also discussed here in Section 8.4.3. Our GPU massively parallel solver, on the other hand, is based on a variant of the fast iterative method [JW08], which requires the DDE property but not causality, although it benefits from it.

In order to fully determine the solution, the scheme $\mathfrak{F}$ is complemented with Dirichlet boundary conditions, of the form $u = \psi$ on the discrete boundary $\partial X := \overline{X} \setminus X$, where $\psi : \partial X \to \mathbb{R}$ is given data. The description of the FMM solver Algorithm 5 involves two additional objects derived from the scheme $\mathfrak{F}$: an update operator Λ defined implicitly, used to reformulate the system of equations $\mathfrak{F}u = 1$ into the fixed point problem $\Lambda u = u$, and a stencil $\mathcal{V}$ which describes the scheme local dependency structure, used to guide the update order. For each $q \in X$ and $u \in \mathbb{R}^{\overline{X}}$, the update operator value $\Lambda u(x) = \lambda \in \mathbb{R}$ is defined by the equation

$$\hat{\mathfrak{F}}(q, [\lambda - u(r)]_{r \in \overline{X}}) = 1. \tag{138}$$

Note that (138, l.h.s.) is a non-decreasing function of λ under the DDE property. In the applications of interest, one easily checks that (138) admits a unique solution, as discussed below in our case, so that the update operator Λ is well defined. On the other hand, for each $q \in X$, the stencil $\mathcal{V}(q) \subset \overline{X}$ is defined as the collection of neighbors r such that the expression of $\mathfrak{F}u(q)$ depends on $u(r)$.

In the following, we fix a grid scale $h > 0$ for the discretization of the PDE domain Ω, and we assume w.l.o.g. that the source point is $q_0 = 0$. Consistently with the addressed problem (128), we can assume that

$$\Omega_h := \Omega \cap h\mathbb{Z}^d, \qquad X := \Omega_h \setminus \{q_0\}, \qquad \partial X := \{q_0\},$$

[8] This terminology is closely related to monotony, but more precise in our context.

with boundary data $u(q_0) = \psi(q_0) = 0$. Note that alternative boundary conditions may be considered, such as the null Dirichlet boundary conditions on $\partial\Omega$ used in the convergence analysis Section 8.4 for simplicity. We reproduce in the rest of this section a DDE and causal scheme for Riemannian eikonal equations originally presented in [Mir19], which acts as a building block for the construction of our scheme in Section 8.1.3 in combination with the PDE operator description (134). For that purpose, we introduce a tool from lattice geometry, known as Selling's decomposition of positive quadratic forms [Sel74, CS92], which is particularly convenient for the design of DDE discretizations of non-linear and anisotropic PDEs on Cartesian grids, both of first [Mir18, Mir19] and second order [FM14, BBM20]. We gather in the next result two properties of Selling's decomposition that are useful for our scheme, consistency (139) and piecewise linearity (140), and we refer to Section 8.B for the proof and for a more constructive and detailed presentation. A function is said affine if it is the sum of a constant and of a linear map.

Proposition 8.7 (Selling's decomposition)**.** *Let $D \in S_d^{++}$, where $d \in \{2, 3\}$. Then Selling's decomposition defines weights $\rho_i \geq 0$, and offsets $e_i \in \mathbb{Z}^d \setminus \{0\}$, where $1 \leq i \leq I = d(d+1)/2$, such that*

$$D = \sum_{1 \leq i \leq I} \rho_i e_i e_i^\top. \tag{139}$$

In addition, assume that $D(\alpha) \in S_d^{++}$ depends in an affine manner of a parameter $\alpha \in [\alpha_*, \alpha^*]$, and let $(\rho_i(\alpha), e_i)_{i=1}^I$ be the weights and offsets of Selling's decomposition[9]. Then there exists $\alpha_* = \alpha_0 < \cdots < \alpha_K = \alpha^*$, such that for all $0 \leq k < K$ and all $1 \leq i \leq I$,

$$\rho_i(\alpha) \text{ is affine as } \alpha \in [\alpha_k, \alpha_{k+1}]. \tag{140}$$

If D is a diagonal matrix, then Selling's decomposition (139) is particularly simple: $\rho_1, \cdots, \rho_d$ are the diagonal coefficients, $\rho_{d+1} = \cdots = \rho_I = 0$, and $e_1, \cdots, e_d$ is the canonical basis of $\mathbb{R}^d$. In contrast, if D is not diagonal, then Selling's decomposition differs for the eigenvalue-eigenvector decomposition, and crucially it only involves offsets $(e_i)_{i=1}^I$ with integer coordinates. The piecewise linearity of Selling's decomposition is used in Theorem 8.11 below to establish that the numerical scheme proposed in this paper benefits from a property, known as quasi-convexity or quasi-concavity, which allows to evaluate it efficiently numerically.

We devote the rest of this subsection to the description and analysis of a discretization scheme denoted $\mathfrak{F}^D$ for the Riemannian eikonal equation [Mir19], which generalizes the classical isotropic fast marching scheme [Set96] using Selling's decomposition. Given $D \in S_d^{++}$, and $u : \Omega_h \to \mathbb{R}$, $q \in \Omega_h \setminus \{q_0\}$, we define

$$\mathfrak{F}^D u(q) := \sum_{1 \leq i \leq I} \rho_i \max\big\{0, \frac{u(q) - u(q + he_i)}{h}, \frac{u(q) - u(q - he_i)}{h}\big\}^2 \tag{141}$$

where $D = \sum_{i=1}^I \rho_i e_i e_i^\top$ is Selling's decomposition. Since the offsets e_i have integer coordinates, the scheme $\mathfrak{F}^D$ only involves values of the unknown u on the Cartesian discretization

[9]Here I is arbitrary. Yet for each $\alpha \in [\alpha_*, \alpha^*]$, at most $d(d+1)/2$ of the weights $\rho_i(\alpha)$, $1 \leq i \leq I$, are positive.

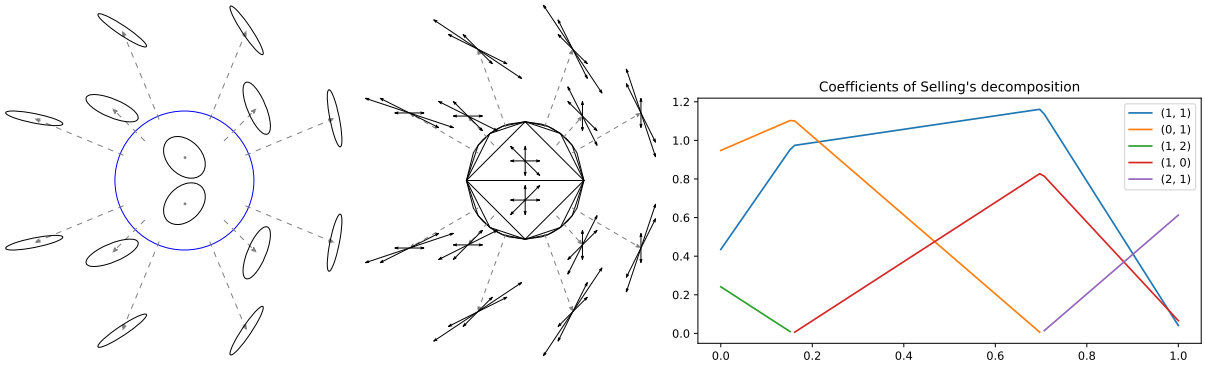

Figure 38: (Left) To each point (x, y) of the open unit disk (blue boundary), we attach the Pauli matrix $D = \begin{pmatrix} 1+x & y \\ y & 1-x \end{pmatrix} \in S_2^{++}$, and show the ellipse $\{p \in \mathbb{R}^2;\ \langle p, D(x,y)p\rangle = 1\}$. (Center) Offsets $\pm e_i$, $1 \le i \le 3$, of Selling's decomposition (139) of the matrix $D(x, y)$. (Right) Coefficients and offsets of Selling's decomposition of $(1-\alpha)D_0+\alpha D_1$, as $\alpha \in [0, 1]$, where $D_0, D_1 \in S_2^{++}$ are randomly chosen. The coefficient associated to a given offset is piecewise affine (140).

grid: $q + \varepsilon h e_i \in h\mathbb{Z}^d$ for all $1 \le i \le I$ and $\varepsilon \in \{-1, 1\}$. By convention, the terms associated to points $q + \varepsilon h e_i \notin \Omega_h$ are discarded[10], which implements outflow boundary conditions. The scheme stencil

$$\mathcal{V}^D(q) := \{q + \varepsilon h e_i;\ 1 \le i \le I,\ \rho_i > 0,\ \varepsilon = \pm 1\} \cap \Omega_h,$$

is reasonably small since $\|e_i\| \le C\sqrt{\|D\|\|D^{-1}\|}$, by Proposition 8.47 (Offset boundedness).

The DDE property of the scheme $\mathfrak{F}^D$ follows from the non-negativity of the weights $(\rho_i)_{i=1}^I$, and the observation that $s \in \mathbb{R} \mapsto \max\{0, s\}^2$ is non-decreasing. Causality holds as well, since by construction $\mathfrak{F}^D u(q)$ only depends on the positive part of the finite differences $u(q) - u(r)$, where $r \in \mathcal{V}^D(q)$. Observing that (141) defines a strictly increasing function of $u(q)$ over the interval $[u_{\min}, +\infty[$, where $u_{\min} := \min\{u(r);\ r \in \mathcal{V}(q)\}$, we obtain that the corresponding update operator Λ^D is uniquely defined by (138). It practice, solving (138) amounts to computing the roots of I univariate quadratic equations, see [MGB+21].

Finally, recalling that $D := \sum_{i=1}^I \rho_i e_i e_i^\top$, we obtain for smooth u the consistency relation

$$\|\nabla u(q)\|_D^2 = \sum_{1\le i\le I} \rho_i \langle \nabla u, e_i\rangle^2 = \sum_{1\le i\le I} \rho_i \max\{0,\ \langle \nabla u, e_i\rangle, -\langle \nabla u, e_i\rangle\}^2 = \mathfrak{F}^D u(q) + \mathcal{O}(h^r), \tag{142}$$

with $r = 1$ for the straightforward implementation (141). Some scheme modifications improve its accuracy, such as source factorization [LQ12], multiscale computation [WFNBZ20], and second order finite differences [Set99] which yield $r = 2$. However they break the DDE and causality properties, hence must be used carefully, see Section 8.C.

[10]Formally, we use the boundary condition $u = +\infty$ on $h\mathbb{Z}^d \setminus \Omega_h$.

Algorithm 5 The Fast Marching algorithm, solving $\mathfrak{F}u = 1$ for a DDE and causal scheme $\mathfrak{F}$.

Input: The update operator Λ and stencils $\mathcal{V}$ associated to $\mathfrak{F}$. Boundary conditions ψ.
Initialize: $u = +\infty$ on the domain X, and $u = \psi$ on ∂X. Tag all points as non-accepted.
While a non-accepted point remains: 1.
  Denote by $q \in \overline{X}$ the non-accepted point minimizing $u(q)$. 2.
  Tag q as accepted. (And optionally, for e.g. higher order methods: PostProcess(q)). 3.
  For each non-accepted point $r \in X$ such that $q \in \mathcal{V}(r)$: 4.
    $u(r) \leftarrow \tilde{\Lambda} u(r)$ (modified operator using only the values from accepted points). 5.

8.1.3 Discretization scheme for the TTI model

We combine the description of the TTI norm as an extremum of Riemannian norms (134), with the Riemannian scheme (141), to obtain a DDE and causal discretization of the TTI eikonal PDE. We then discuss its efficient numerical implementation, and the convergence of its solutions as the grid scale $h > 0$ is refined. Specifically, and consistently with (125), define for any $u : \Omega_h \to \mathbb{R}$ and any $q \in \Omega_h$, where $\Omega_h := \Omega \cap h\mathbb{Z}^d$ is a Cartesian discretization grid

$$\mathfrak{F}u(q) := \min_{\alpha \in [\alpha_*, \alpha^*]} \mathfrak{F}^\alpha u(q), \qquad \text{where } \mathfrak{F}^\alpha := \frac{1}{\mu(\alpha)} \mathfrak{F}^{D(\alpha)}. \tag{143}$$

We used the notations $0 < \alpha_* \leq \alpha^* < 1$ and $\mu(\alpha) > 0$ from Theorem 8.3, the matrix $D(\alpha)$ and extremum operator $\mathrm{mix} \in \{\min, \max\}$ from (134), and the Riemannian scheme $\mathfrak{F}^D$ defined in (141). For sufficiently smooth u one has

$$\mathcal{F}_\sigma(R\nabla u(q))^2 = \min_{\alpha \in [\alpha_*, \alpha^*]} \frac{1}{\mu(\alpha)} \|\nabla u(q)\|^2_{D(\alpha)} = \mathfrak{F}u(q) + \mathcal{O}(h^r), \tag{144}$$

using the consistency relation in the Riemannian case (142), with the same order $r \in \{1, 2\}$, and the expression of the TTI norm (134).

The proposed scheme $\mathfrak{F}$ for the TTI eikonal PDE is defined as the maximum or minimum of the infinite family of schemes $\mathfrak{F}^{D(\alpha)}$, $\alpha \in [\alpha_*, \alpha^*]$, hence inherits their DDE and causality properties as shown in Proposition 8.8 below. The existence and uniqueness of the scheme solutions, and their convergence as $h \to 0$ to the viscosity solution of (128), can then be established following a common scheme of proof, see Theorem 8.33 below. In a similar fashion, a DDE scheme for the second order fully non-linear Monge-Ampere [BM21] and Pucci [BBM20] PDEs, in two dimensions, is obtained as the maximum of an infinite family of linear schemes. In the same spirit again, and in the context of fast marching methods, multi-stencil schemes [HF07] are defined as the maximum of a finite number of discretizations of the eikonal equation (with identical anisotropy, unlike here). Note however that considering the *minimum* of several schemes, as we do in (143) for case (min) of Theorem 8.3, is uncommon and leads to a few additional difficulties in the analysis in comparison with the (max) case, see Section 8.4.

Proposition 8.8. *Let A be a compact set and let $\mathfrak{F}^\alpha$, for each $\alpha \in A$, be a finite difference scheme on a finite set X, depending continuously on the parameter α. Define*

$$\mathfrak{F}u(q) := \max_{\alpha\in A} \mathfrak{F}^\alpha u(q) \qquad \Big(\textit{resp. } \mathfrak{F}u(q) := \min_{\alpha\in A} \mathfrak{F}^\alpha u(q)\Big)$$

for all $u \in \mathbb{R}^{\overline{X}}$ and all $q \in X$. If $\mathfrak{F}^\alpha$ is DDE (resp. causal) for all $\alpha \in A$, then so is $\mathfrak{F}$. Furthermore, denoting by Λ^α the update operator for $\mathfrak{F}^\alpha$ (which is assumed to exist and to depend continuously on $\alpha \in A$), one has

$$\Lambda u(q) = \min_{\alpha\in A} \Lambda^\alpha u(q) \qquad \Big(\textit{resp. } \Lambda u(q) = \max_{\alpha\in A} \Lambda^\alpha u(q)\Big). \tag{145}$$

Proof. A maximum or a minimum over a compact set of a continuously depending family of functions which are continuous (resp. non-decreasing) (resp. depend only on the positive part of their arguments), clearly obeys the same property. The first claim follows.

We focus on (145, left) since the other case is proved similarly, and denote $\lambda_* := \min_{\alpha\in A} \Lambda^\alpha u(q)$. If $\lambda < \lambda_*$, then $\hat{\mathfrak{F}}^\alpha(q, [\lambda - u(r)]_{r\in X}) < 0$ for all $\alpha \in A$ by the DDE property, hence $\max_{\alpha\in A} \hat{\mathfrak{F}}^\alpha(q, [\lambda - u(r)]_{r\in X}) < 0$ by compactness. If $\lambda > \lambda_*$ on the other hand, then $\max_{\alpha\in A} \hat{\mathfrak{F}}^\alpha(q, [\lambda - u(r)]_{r\in X}) > 0$ by definition. Thus, by continuity, λ_* is the unique solution to $\hat{\mathfrak{F}}(q, [\lambda - u(r)]_{r\in X})$, hence $\lambda_* = \Lambda u(q)$ as announced. □

Proposition 8.8 immediately implies that the TTI scheme (143) is DDE and causal, and thus solvable using the FMM Algorithm 5. The numerical implementation however requires to efficiently evaluate the update operator Λ associated with the scheme, which is defined as an extremum (145) over a continuous set of parameters $A = [\alpha_*, \alpha^*]$. We compare in the following two strategies for solving this optimization problem, used respectively in our GPU and CPU eikonal solver.

Optimization by grid search. In this approach, the maximization or minimization problem (143) over $[\alpha_*, \alpha^*]$, is approximated using an exhaustive search over a regular sampling of this real interval with $K + 1$ elements, where the integer $K \geq 1$ is fixed by the user. More explicitly, we introduce the scheme $\mathfrak{F}_K$ and update operator Λ_K defined as

$$\mathfrak{F}_K u(q) := \operatorname*{mix}_{0\leq k\leq K} \mathfrak{F}^{\alpha_k} u(q), \quad \Lambda_K u(q) := \operatorname*{\overline{mix}}_{0\leq k\leq K} \mathfrak{F}^{\alpha_k} u(q), \quad \text{with } \alpha_k := (1 - \frac{k}{K})\alpha_* + \frac{k}{K}\alpha^*, \tag{146}$$

following the notations of (143). In particular $\operatorname{mix} \in \{\max, \min\}$ is the suitable extremum, and $\overline{\operatorname{mix}} \in \{\min, \max\}$ is the opposite extremum, following (133) and Proposition 8.8. The use of an equispaced sampling of parameters $\alpha_0 \leq \cdots \leq \alpha_K$ in the interval $[\alpha_*, \alpha^*]$ is quasi-optimal for consistency, see Proposition 8.9, and corresponds to an envelope of the TTI slowness surface by a family of ellipses with regularly varying aspect ratios, which is visually pleasing, see Figure 36 and Figure 40.

The eikonal solvers [MP19, MGB+21], originally limited to $K = 2$ and to the (max) case, are easily adapted to address the scheme $\mathfrak{F}_K$, using an exhaustive search over $0 \leq k \leq K$ to evaluate the update operator Λ_K. This approach is well suited to massively parallel

accelerators such as GPUs, since those have (i) enough horsepower to accommodate the computational cost of exhaustive search, and (ii) a SIMT[11] architecture that is not well suited to the multiple conditional branchings found in more sophisticated optimization procedures.

Proposition 8.9. *For smooth u, one has $\mathfrak{F}_K u(q) = \mathfrak{F}u(q) + \mathcal{O}(h^r + K^{-2})$.*

Proof. By the consistency relation in the Riemannian case (142),

$$\min_{0\leq k\leq K} g(\alpha_k) = \mathfrak{F}_K u(q) + \mathcal{O}(h^r), \qquad \text{where } g(\alpha) := \frac{1}{\mu(\alpha)}\|\nabla u(q)\|^2_{D(\alpha)}. \tag{147}$$

The function $g : [\alpha_*, \alpha^*] \to \mathbb{R}$ is smooth by (134) and Theorem 8.3. Since $\alpha_0 = \alpha_*$, $\alpha_K = \alpha^*$, and $\alpha_{k+1} - \alpha_k = (\alpha^* - \alpha_*)/K = \mathcal{O}(1/K)$ for all $0 \leq k < K$, one has $\min\{g(\alpha_k); 0 \leq k \leq K\} = \min\{g(\alpha); \alpha \in [\alpha_*, \alpha^*]\} + \mathcal{O}(1/K^2)$. From this point, the announced result follows from (144). □

For numerical efficiency, one usually balances the errors $\mathcal{O}(h^r + K^{-2})$ associated to the discretization scale h and to the consistency of the operator approximation with K terms. This suggests the parameter choice $K \approx h^{-\frac{1}{2}}$ with a first order scheme ($r = 1$), and $K \approx h^{-1}$ with a second order scheme ($r = 2$). For instance, in the synthetic numerical experiment presented on Figure 45, the TTI scheme needs to be defined as the extremum of $K + 1 = 26$ Riemannian schemes to ensure good accuracy. In the second order case, the evaluation cost of the update operator (146) thus becomes non-negligible, and for this reason an optimization procedure more efficient than exhaustive search is described in the next paragraph.

Quasi-convex optimization. This approach relies on a fine property of Selling's matrix decomposition, namely the piecewise linearity of its coefficients (140) established in Proposition 8.7, which is used in this paper for the first time in the discretization of a three dimensional PDE. The same property is exploited in [BM21, BBM20] to obtain a DDE, second order consistent, and numerically efficient scheme for the Pucci and Monge-Ampere PDEs in two dimensions. In those previous works, the non-linear PDE operator can be expressed as the maximum of an infinite family of linear operators, each discretized using Selling's decomposition, in a spirit similar to (143); a closed form expression is then obtained using the piecewise linear property of Selling's decomposition. We do not obtain such a closed form here, but nevertheless we derive a property known as quasi-convexity, allowing for an efficient implementation.

Definition 8.10. *A function $f : A \to \mathbb{R}$, where A is a convex subset of a vector space, is said quasi-convex if for each $\lambda \in \mathbb{R}$ the set $\{x \in I;\ f(x) \leq \lambda\}$ is convex.*

By construction, the set of minimizers of a quasi-convex function is convex, and in particular there is at most one isolated local minimum. If $A = [a, b]$ is a segment of $\mathbb{R}$, as in our application, and if f is continuous and quasi-convex, then the classical golden search method [PTVF07, Section 10.2] produces an interval of length $(b-a)\phi^N$ containing

[11] Single instruction multiple threads

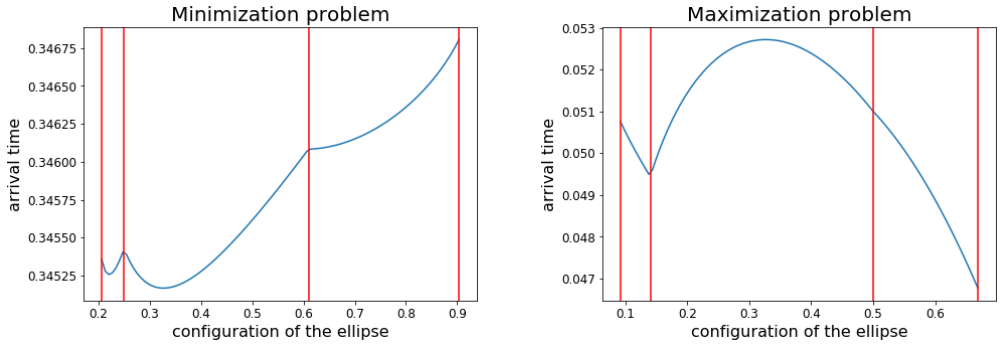

Figure 39: Mapping $\alpha \in [\alpha_*, \alpha^*] \mapsto f(\alpha) := \Lambda^\alpha u(q)$ obtained for some TTI parameters σ, R, a point $q \in h\mathbb{Z}^d$, and an arbitrary mapping $u : \Omega_h \to \mathbb{R}$. The vertical red lines correspond to the abcissas $\alpha_0 \leq \cdots \leq \alpha_K$ of Proposition 8.7, here with $K = 3$. Left (resp. Right) subfigure illustrates case (max) (resp. case (min)), where by Theorem 8.11 the function f is quasi-convex (resp. quasi-concave) on each sub-interval $[\alpha_k, \alpha_{k+1}]$, $0 \leq k \leq K$, and must be minimized (resp. maximized).

its minimizer using $N+1$ evaluations of f, where $\phi^{-1} = \frac{1+\sqrt{5}}{2}$ is the golden ratio. This is considerably more efficient than optimization by grid search, considered previously, which only yields an interval of length $2(b-a)/N$ for the same numerical cost. A function f is said quasi-concave if $-f$ is convex, and in that case by the previous discussion it can be efficiently *maximized* numerically.

Theorem 8.11. *Let $\alpha_* \leq \alpha_0 \leq \cdots \alpha_K = \alpha^*$ be such that Selling's decomposition of the matrix $D(\alpha)$, see (134, right), is piecewise linear on each interval $[\alpha_k, \alpha_{k+1}]$, $0 \leq k < K$, in the sense of (140). Fix $u : \Omega_h \to \mathbb{R}$, $q \in \Omega_h$ and define $f(\alpha) := \Lambda^\alpha u(q)$ for all $\alpha \in [\alpha_*, \alpha^*]$. Then the following alternative holds, whose cases match those of Theorem 8.3*

(max) f is quasi-convex on each interval $[\alpha_k, \alpha_{k+1}]$, $0 \leq k < K$.

(min) f is quasi-concave on each interval $[\alpha_k, \alpha_{k+1}]$, $0 \leq k < K$.

This result allows to extremize the function $f(\alpha) := \Lambda^\alpha u(q)$ over the interval $[\alpha_*, \alpha^*]$ in a numerically efficient manner, and thus to evaluate the update operator (145). A possible allure of f is illustrated on Figure 39.

8.1.4 Summary of the numerical method

This paper defines a numerical method designed to solve the eikonal equation in a geological medium with a TTI Hooke tensor. We present a summary of each of its steps. The computer code which implements the method in this study can also be found at: `https://github.com/Mirebeau`, with illustrative Python Notebooks.

Numerical solver. We consider the numerical scheme defined in (134) to solve the eikonal equation for a TTI metric. First, we need to determine if we are in the

(max) or (min) case. A criterion is presented in (163), by computing the sign of $\det(\nabla\mathcal{Q}(p_*), \nabla\mathcal{Q}(p^*))$. Explicit formulas for the computation of the bounds α_*, α^* from $\nabla\mathcal{Q}(p^*)$ are also presented in Corollary (8.28) and Corollary (8.29), and for the parameter $\mu(\alpha)$ in (168).

Computation of the maximum or minimum. The computation of (134) for all $\alpha \in [\alpha_*, \alpha^*]$ leads to an envelope by ellipses (either by the outside or from the outside) of the P-slowness surface, see Figure 36: each $\alpha \in [\alpha_*, \alpha^*]$ corresponds to a tangent lines to $\partial\mathcal{A}_\sigma$ (and a tangent ellipse to $\partial\mathcal{B}_\sigma$), as represented in Figure 37.

In Figure 39, we present an example of the minimum and the maximum on $\alpha \in [\alpha_*, \alpha^*]$ that we need to compute for the update operator. We consider two possibilities to compute this optimum:

- Quasi-convex optimization, see Theorem 8.11: from studying the stencils required in the numerical scheme based on the Eulerian scheme, we prove that the optimization problem can be divided into a finite number of intervals with at most one optimum in each interval. Therefore, a Newton-like search algorithm is possible in each of these intervals. An illustration is presented in Figure 39, with the red vertical lines indicating the different intervals.

- Optimization by grid-search, see (146): we can also consider the optimum over k ellipses only. This method is much less precise compared with the previous algorithm, but can be done much faster, with a GPU implementation.

8.2 Properties and guarantees of TTI models

This section is devoted to the proof of the results announced in Section 8.1.1. We introduce in Section 8.2.1 several properties of Hooke tensors, known as ellipticity, positivity and separability, and relate them with the admissibility conditions (131) of the TTI parameters σ. We establish in Section 8.2.2 that the TTI unit ball $\mathcal{B}_\sigma$ is convex and compact, thus concluding the proof of Theorem 8.2. We prove Theorem 8.3 in Section 8.2.3, where we also derive the closed form expression of the weight function $\mu : [\alpha_*, \alpha^*] \to]0, \infty[$.

Notations. The TTI parameters $\sigma \in \mathbb{R}^5$ are regarded as fixed throughout this section, and thus for readability the sets $\mathcal{A}_\sigma, \mathcal{B}_\sigma, \mathcal{C}_\sigma, \mathcal{Q}_\sigma$ introduced in Section 8.1.1 are simply denoted $\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{Q}$. The symbol $\propto$ denotes *positive proportionality*, i.e. $v \propto w$ iff $v = \lambda w$ for some $\lambda > 0$.

8.2.1 Admissible coefficients, and properties of Hooke tensors

We relate the TTI eikonal PDE with a two-dimensional Hooke tensor (151), and investigate its algebraic properties as a preliminary step to Theorem 8.2. See Section 8.A and references therein for a more physically oriented discussion of these elasticity parameters. A Hooke tensor is a 4-th order tensor, $\mathfrak{c} = (\mathfrak{c}_{ijkl})$ where $i, j, k, l \in \{1, \cdots, d\}$ in dimension d, which characterizes the anisotropy properties of a linear elastic material, hence also the propagation speed of elastic waves through it (196). Hooke tensors are subject to

the major and minor symmetry relations $\mathfrak{c}_{ijkl} = \mathfrak{c}_{jikl} = \mathfrak{c}_{klij}$, and for this reason a Hooke tensor $\mathfrak{c}$ can be represented compactly as a symmetric matrix $\mathfrak{C}$ of shape 3×3 if $d = 2$ (resp. 6×6 if $d = 3$) using Voigt's matrix of indices denoted v:

$$\mathfrak{c}_{ijkl} := \mathfrak{C}_{\mathrm{v}_{ij}\mathrm{v}_{kl}} \qquad \text{where } \mathrm{v} = \begin{pmatrix} 1 & 3 \\ 3 & 2 \end{pmatrix} \qquad \Big(\text{resp. } \mathrm{v} = \begin{pmatrix} 1 & 6 & 5 \\ 6 & 2 & 4 \\ 5 & 4 & 3 \end{pmatrix}\Big). \tag{148}$$

Following [BST83] we recall the notion of a positive or elliptic Hooke tensor in Definition 8.12, and the relation between these properties in Lemma 8.13.

Definition 8.12. *A Hooke tensor $\mathfrak{c}$ is said strictly positive (resp. strictly elliptic) if*

$$\sum_{i,j,k,l} \mathfrak{c}_{ijkl}\, m_{ij}\, m_{kl} > 0 \qquad \Big(resp. \sum_{i,j,k,l} \mathfrak{c}_{ijkl}\, p_i\, q_j\, p_k\, q_l > 0\Big), \tag{149}$$

for all $m \in S_d \setminus \{0\}$ (resp. $p, q \in \mathbb{R}^d \setminus \{0\}$), where the sums implicitly range over $i, j, k, l \in \{1, \cdots, d\}$.

In order to describe further properties of Hooke tensors, we introduce for all $p \in \mathbb{R}^d$ a symmetric matrix $\mathfrak{c}(p) \in S_d$ defined as follows: for all $j, l \in \{1, \cdots, d\}$

$$\mathfrak{c}(p)_{jl} := \sum_{i,k \in \{1,\cdots,d\}} \mathfrak{c}_{ijkl}\, p_i\, p_k, \qquad \text{thus } \langle q, \mathfrak{c}(p)q\rangle = \sum_{i,j,k,l} \mathfrak{c}_{ijkl}\, p_i\, q_j\, p_k\, q_l. \tag{150}$$

The following lemma rephrases the positivity and ellipticity properties of Hooke tensors in terms of usual matrix positive definiteness.

Lemma 8.13. *A Hooke tensor $\mathfrak{c}$ is strictly positive iff $\mathfrak{C} \in S_D^{++}$ with $D = d(d+1)/2$, where $\mathfrak{C}$ is defined by Voigt's notation* (148)*. It is strictly elliptic iff $\mathfrak{c}(p) \in S_d^{++}$ for all $p \neq 0$. Strict positivity implies strict ellipticity.*

Proof. By definition (149, left) a Hooke tensor is strictly positive iff it defines a positive definite quadratic form over the space S_d of $d \times d$ symmetric matrices, which has dimension D. Noting that $\mathfrak{C}$ is the matrix of this quadratic form, in the basis E_{11}, E_{22}, $(E_{12} + E_{21})/2$ if $d = 2$ where E_{ij} is the null matrix except for a single coefficient 1 at position (i, j), and likewise in the case $d = 3$, we establish the first point. On the other hand, the definition (149, right) of ellipticity can be rephrased using the identity (150, right) as $\langle q, \mathfrak{c}(p)q\rangle > 0$ for all $p, q \neq 0$, in other words $\mathfrak{c}(p) \in S_d^{++}$ for all $p \neq 0$, as announced. Finally, given a strictly positive Hooke tensor and $p, q \in \mathbb{R}^d \setminus \{0\}$, define $m \in S_d$ by $m_{ij} = p_i q_j + q_j p_i$, equivalently $m = pq^\top + qp^\top$, and note that $\mathrm{Tr}(m^2) = 2(\langle p, q\rangle^2 + \|p\|^2\|q\|^2) > 0$. Thus $m \neq 0$ and therefore $0 < \sum_{i,j,k,l} \mathfrak{c}_{ijkl} m_{ij} m_{kl} = 4 \sum_{i,j,k,l} \mathfrak{c}_{ijkl} p_i q_j p_k q_l$, showing that $\mathfrak{c}$ is strictly elliptic as announced. □

We introduce in Definition 8.14 a non-degeneracy property of Hooke tensors referred to as separability [DCC+21]. This property ensures that the slowness surfaces of the pressure and shear waves are topologically separated from each other, see Figure 36 and Figure 37 for examples and Figure 40 for counter-examples.

Definition 8.14. *A Hooke tensor $\mathfrak{c}$ is said separable iff the largest eigenvalue of $\mathfrak{c}(p)$ has multiplicity one, for all $p \in \mathbb{R}^d \setminus \{0\}$.*

In the rest of this section, we limit our attention to the following two dimensional Hooke tensor, whose coefficients are assumed to belong to the admissible set C_{adm}, and are related to the TTI coefficients σ by (130):

$$\mathfrak{c} := \begin{pmatrix} c_{11} & c_{13} & 0 \\ c_{13} & c_{33} & 0 \\ 0 & 0 & c_{44} \end{pmatrix}, \qquad \begin{cases} c_{11} > c_{44},\ c_{33} > c_{44}, \\ c_{44} > 0,\ c_{13} + c_{44} > 0, \\ c_{11}c_{33} > c_{13}^2. \end{cases} \tag{151}$$

We establish in Proposition 8.15 that $\mathfrak{c}$ is strictly positive, hence strictly elliptic by Lemma 8.13. We also prove that admissible coefficients form a convex set, as announced in Theorem 8.2.

Proposition 8.15. *The following sets of coefficients $(c_{11}, c_{13}, c_{33}, c_{44}) \in \mathbb{R}^4$ are open and convex:*

$$C^1_{\text{adm}} := \{c_{11} > 0,\ c_{33} > 0,\ c_{11}c_{33} > c_{13}^2,\ c_{44} > 0\},$$
$$C^2_{\text{adm}} := \{c_{11} > c_{44},\ c_{33} > c_{44},\ c_{13} > -c_{44}\},$$

thus also their intersection $C_{\text{adm}} = C^1_{\text{adm}} \cap C^2_{\text{adm}}$. In addition the Hooke tensor $\mathfrak{c}$ defined by (151, left) is strictly positive for any $(c_{11}, c_{13}, c_{33}, c_{44}) \in C^1_{\text{adm}} \supset C_{\text{adm}}$.

Proof. The openness properties follow from the definition of C^1_{adm} and C^2_{adm} by strict inequalities. Recall that $[c_{11} > 0,\, c_{33} > 0$ and $c_{11}c_{33} > c_{13}^2]$ iff $\begin{pmatrix} c_{11} & c_{13} \\ c_{13} & c_{33} \end{pmatrix} \in S_2^{++}$. Thus C^1_{adm} characterizes the positive definiteness of the block matrix (151, left), as announced. This also shows that C^1_{adm} is in linear bijection with $S_2^{++} \times]0, \infty[$, hence is a convex set. The set C^2_{adm} is convex since it is defined by linear inequalities. □

The rest of this subsection is devoted to the proof that $\mathfrak{c}$ is separable, which is concluded in Corollary 8.19. For that purpose, we introduce the quadratic function $F : \mathbb{R}^2 \to \mathbb{R}$ defined as

$$F(x, z) := 1 + c_{11}c_{44}x^2 + c_{33}c_{44}z^2 - (2c_{13}c_{44} + c_{13}^2 - c_{11}c_{33})xz - (c_{11} + c_{44})x - (c_{33} + c_{44})z. \tag{152}$$

The next identity (153, left) relates the function F with the Hooke tensor $\mathfrak{c}$, whereas (153, right) links it with the TTI eikonal equation (128). For all $p = (p_x, p_z) \in \mathbb{R}^2$ and all $x, z \in \mathbb{R}$

$$\det(\mathfrak{c}(p) - \mathrm{Id}) = F(p_x^2, p_z^2), \qquad F(x, z) = 1 - \mathcal{Q}(x, z). \tag{153}$$

Note that the quadratic function $\mathcal{Q}$ defined in (126) is based on the admissible coefficients (130).

Remark 8.16 (Validation of the polynomial identities)**.** *Checking polynomial identities such as* (153) *by hand is simultaneously trivial, tedious, and error prone. For this reason, a companion notebook is provided containing those verifications in the Wolfram Mathematica® language.*

Proposition 8.18 below, which is the most technical result of this subsection, establishes that the function F vanishes exactly twice in each direction of the positive quadrant. In the following, the letter r is used to denote a radius in the two-dimensional (x, z) plane. This usage is *distinct* from the three dimensional cylindrical coordinates considered in Section 8.1.

Lemma 8.17. *Define the polynomial $P(r) := ar^2 + br + c$, for all $r \in \mathbb{R}$. If $a > 0$, $b < 0$, $c > 0$, and the discriminant $\Delta := b^2 - 4ac > 0$ is positive, then P has two distinct positive roots.*

Proof. Noting that $0 < \Delta < b^2$ we obtain $\sqrt{\Delta} < |b| = -b$, and thus $(-b \pm \sqrt{\Delta})/(2a) > 0$. □

Proposition 8.18. *For any $\theta \in [0, \pi/2]$, the quadratic polynomial $r \in \mathbb{R} \mapsto F(r\cos\theta, r\sin\theta)$ has two distinct positive roots. Denoting by $R(\theta)$ the smallest root, one has $R \in C^\infty([0, \pi/2],]0, \infty[)$.*

Proof. By symmetry, we may assume that $0 \leq \theta < \pi/2$. Using the change of variables $r' = r/\cos\theta$, and denoting $\alpha := \tan\theta \geq 0$, we can limit our attention to the following polynomial:

$$F_\alpha(r) := F(r, \alpha r) =$$
$$1 + r^2(c_{11}c_{44} + \alpha^2 c_{33}c_{44} - 2\alpha c_{13}c_{44} - \alpha c_{13}^2 + \alpha c_{11}c_{33}) - r(c_{11} + c_{44} + \alpha c_{33} + \alpha c_{44}).$$

One has $F_\alpha(0) = 1 > 0$ and $F'_\alpha(0) = -(c_{11} + c_{44} + \alpha c_{33} + \alpha c_{44}) < 0$. The coefficient of r^2 in $F_\alpha(r)$ reads

$$c_{44}(c_{11} + \alpha^2 c_{33} - 2\alpha c_{13}) + \alpha(c_{11}c_{33} - c_{13}^2)$$

which is positive, by the admissibility conditions $c_{44} > 0$ and $c_{11}c_{33} > c_{13}^2$. Indeed one has $(c_{11} + \alpha^2 c_{33})/2 \geq \sqrt{c_{11}\alpha^2 c_{33}} \geq \alpha|c_{13}|$, by the arithmetic geometric mean inequality. The discriminant of $F_\alpha(r)$ reads (after suitably grouping the terms)

$$\Delta(\alpha) = \alpha^2(c_{33} - c_{44})^2 + (c_{11} - c_{44})^2 + 2\alpha c_*,$$
$$c_* := 2c_{13}^2 - c_{11}c_{33} + c_{44}(c_{11} + 4c_{13} + c_{33} + c_{44}).$$

Distinguishing two cases, depending on the sign of c_*, we establish below that $\Delta(\alpha) > 0$.

- Case $c_* \geq 0$. Then $\Delta(\alpha) \geq (c_{11} - c_{44})^2 > 0$ for any $\alpha \geq 0$, as announced.
- Case $c_* < 0$. Then we consider the discriminant of the polynomial $\alpha \mapsto \Delta(\alpha)$, which reads
$$16 \times (c_{13} + c_{44})^2(c_* - (c_{13} + c_{44})^2) \tag{154}$$
(after factorization), and is thus negative since $c_{13} + c_{44} > 0$ by admissibility. Therefore $\Delta(\alpha) \neq 0$ for all $\alpha \in \mathbb{R}$, and thus $\Delta(\alpha)$ has the same sign as $\Delta(0) = (c_{11} - c_{44})^2 > 0$.

By Lemma 8.17, the polynomial F_α admits two positive roots, as announced. Finally we note that the smallest root $(-b - \sqrt{\Delta})/(2a)$ of a polynomial of degree two $P(r) = a + br + cr^2$ is a smooth function of its coefficients so long as the discriminant $\Delta = b^2 - 4ac$ and the dominant coefficient a remain positive. By composition the smallest root of F_α depends smoothly on $\alpha \in [0, \infty[$, which concludes. □

Corollary 8.19. *The Hooke tensor* $\mathfrak{c}$ *is separable.*

Proof. Fix an arbitrary $p = (p_x, p_z) \in \mathbb{R}^2 \setminus \{0\}$. Then for any $r > 0$, one has

$$F(rp_x^2, rp_z^2) = \det(\mathfrak{c}(\sqrt{r}p) - \mathrm{Id}) = \det(r\mathfrak{c}(p) - \mathrm{Id}) = r^2 \det(\mathfrak{c}(p) - r^{-1}\,\mathrm{Id}). \tag{155}$$

By Proposition 8.18, the polynomial $r \mapsto F(rp_x^2, rp_z^2)$ admits two positive roots $0 < r_1 < r_2$, and thus the matrix $\mathfrak{c}(p)$ admits two positive eigenvalues $0 < r_2^{-1} < r_1^{-1}$, which concludes. □

Remark 8.20 (The condition $c_{13}+c_{44} > 0$)**.** *Consider a material obeying the admissibility conditions* C_{adm}*, except that* $c_{13} + c_{44} < 0$ *rather than the opposite. Define* $c'_{13} := -c_{13} - 2c_{44}$*, in such way that* $c'_{13} + c_{44} = -(c_{13} + c_{44}) > 0$*.*

The modified Hooke tensor coefficients $(c_{11}, c'_{13}, c_{33}, c_{44})$ *yield the same eikonal PDE as* $(c_{11}, c_{13}, c_{33}, c_{44})$*, since* $c_{13}^2 + 2c_{13}c_{44} - c_{11}c_{33} = (c_{13} + c_{44})^2 - c_{44}^2 - c_{11}c_{33}$ *only depends on* $(c_{13}+c_{44})^2$ *and the other terms of* (129) *are independent of* c_{13}*. The modified coefficients also meet the totality of the admissibility conditions* (131)*, noting that* $c'^2_{13} = c_{13}^2+4c_{44}(c_{44}+c_{13}) \leq c_{13}^2$*.*

On the positive side, this discussion shows that our numerical method can handle (hypothetical) materials such that $c_{13} + c_{44} < 0$*, through modified coefficients. On the negative side, this phenomenon illustrates an invariance of the TTI eikonal PDE, which therefore cannot be used to reconstruct the sign of* $c_{13} + c_{44}$ *in a tomography context.*

In the degenerate case where $c_{13} + c_{44} = 0$*, the eikonal equation factors as* $F(x, z) = (1-c44x-c33z)(1-c11x-c44z)$*. Subject to the other admissibility conditions, the conic* $\mathcal{C}$ *is then a union of two lines intersecting at the point* $(c_{33} - c_{44}, c_{11} - c_{44})/(c_{11}c_{33} - c_{13}^2)$ *of the positive quadrant, as illustrated on Figure 40 (right).*

8.2.2 Region delimited by a conic

In this section we conclude the proof of Theorem 8.2, which describes the shape of slowness profile $\mathcal{B}$ of the pressure waves, see Corollary 8.23. The ellipticity and separability of Hooke tensors defining TTI models, established in Section 8.2.1, are the key ingredient of the first result Proposition 8.21. In this section, we assume that $(c_{11}, c_{13}, c_{33}, c_{44}) \in C_{\mathrm{adm}}$ are admissible TTI parameters, see Theorem 8.2, and that $\mathfrak{c}$ and σ are the corresponding Hooke tensor (151) and coefficients (130) of the eikonal equation. The quadratic form $\mathcal{Q} = \mathcal{Q}_\sigma$ and set $\mathcal{B} = \mathcal{B}_\sigma$ are defined in (127). The regions $\mathcal{A}$, $\mathcal{B}$, and conic $\mathcal{C}$ are illustrated on Figure 41.

Proposition 8.21. *The set* $\mathcal{B}' := \mathrm{CC}_0\{(p_x, p_z) \in \mathbb{R}^2;\, \mathcal{Q}(p_x^2, p_z^2) \leq 1\}$ *is compact and convex.*

Proof. Define $N_{\mathfrak{c}}(p) := \sqrt{|||\mathfrak{c}(p)|||}$ for all $p \in \mathbb{R}^2$, where $|||m|||$ denotes the spectral norm of a matrix m, which is also the largest eigenvalue if $m \in S_2^{++}$. Since $\mathfrak{c}$ is a strictly elliptic Hooke tensor, as shown in Proposition 8.15, the function $N_{\mathfrak{c}}$ defines a norm over $\mathbb{R}^2$, by [DCC+21, Theorem 3.3]. As a result, the set $B_{\mathfrak{c}} := \{p \in \mathbb{R}^2;\, N_{\mathfrak{c}}(p) \leq 1\}$ is compact and convex.

Since $\mathfrak{c}$ is separable, as shown in Corollary 8.19, one has $B_{\mathfrak{c}} = \mathrm{CC}_0\{p \in \mathbb{R}^2;\, \det(\mathfrak{c}(p) - \mathrm{Id}) \geq 0\}$, by [DCC+21, Proposition 3.7]. Recalling the identity (153) we conclude the proof. □

We present in Lemma 8.22 a simple criterion for the convexity of axisymmetric sets, which is applied to the slowness profile $\mathcal{B}$ in Corollary 8.23, thus concluding the proof of Theorem 8.2.

Lemma 8.22. *Let E, F be normed vector spaces, and let $K \subset \mathbb{R} \times F$ be convex and such that $(-s, z) \in K$ for all $(s, z) \in K$. Then $\{(x, z) \in E \times F; (|x|, z) \in K\}$ is convex.*

Proof. Let $(x_1, z_1), (x_2, z_2) \in E \times F$, and let $t \in]0, 1[$. Define

$$s := \frac{|(1-t)x_1 + tx_2|}{(1-t)|x_1| + t|x_2|} \in [0, 1],$$
$$\alpha = (1-t)\frac{1+s}{2},\ \beta = (1-t)\frac{1-s}{2},\ \gamma = t\frac{1+s}{2},\ \delta = t\frac{1-s}{2},$$

choosing $s \in [0, 1]$ arbitrarily if $|x_1| = |x_2| = 0$. Then $\alpha, \beta, \gamma, \delta \geq 0$, $\alpha + \beta + \gamma + \delta = 1$, and

$$\big(|(1-t)x_1 + tx_2|, (1-t)z_1 + tz_2\big) = \alpha(|x_1|, z_1) + \beta(-|x_1|, z_1) + \gamma(|x_2|, z_2) + \delta(-|x_2|, z_2),$$

which establishes the announced convexity property. □

Corollary 8.23. *The set $\mathcal{B} := \mathrm{CC}_0\{(p_x, p_y, p_z);\ \mathcal{Q}(p_x^2 + p_y^2, p_z^2) \leq 1\}$ is compact and convex.*

Proof. The closedness and boundedness of $\mathcal{B}$, hence compactness, follow immediately from the same properties of $\mathcal{B}'$, established in Proposition 8.21. Convexity follows from Lemma 8.22 applied to the set $K = \mathcal{B}'$ from Proposition 8.21, choosing $E = \mathbb{R}^2$ equipped with the Euclidean norm, and $F = \mathbb{R}$. □

Remark 8.24 (Positivity without separability)**.** *If one weakens the admissibility condition for the TTI coefficients* (131)*, assuming only that $(c_{11}, c_{13}, c_{33}, c_{44}) \in C^1_{\mathrm{adm}}$, see Proposition 8.15, then the Hooke tensor (151, left) remains positive but may not be separable. As a result, the P and SH slowness surfaces may intersect each other, see Figure 40. Under these weaker assumptions, the open TTI unit ball $\mathrm{CC}_0\{(p_x, p_y, p_z) \in \mathbb{R}^3;\ \mathcal{Q}_\sigma(p_x^2 + p_y^2, p_z^2) < 1\}$ is bounded and convex but may have a non-smooth boundary, and likewise the solution u of the eikonal PDE* (128) *has lower regularity. Since no common geophysical material appears to fail the stronger C_{adm} conditions, see Section 8.A, we limit our attention to those, eliminating a few mathematical technicalities in the process.*

The rest of this section is a preparation for the proof of Theorem 8.3, achieved in Section 8.2.3. In particular, the alternative between the (max) and (min) cases arises in Corollary 8.26 from the fact that a (connected component of a non-degenerate) conic curve has no inflexion point, and therefore has a convex side and concave side. We recall from Section 8.1.1 that

$$\mathcal{A} := \mathrm{CC}_0\{(x, z) \in \mathbb{R}^2_+;\ \mathcal{Q}(x, z) \leq 1\}, \qquad \mathcal{C} := \{(x, z) \in \mathbb{R}^2; \mathcal{Q}(x, z) = 1\}. \tag{156}$$

The set $\mathcal{C}$ is a conic, in other words an algebraic set of degree two - which can thus be an ellipse, a hyperbola, a parabola, the union of two lines, etc, depending on the choice

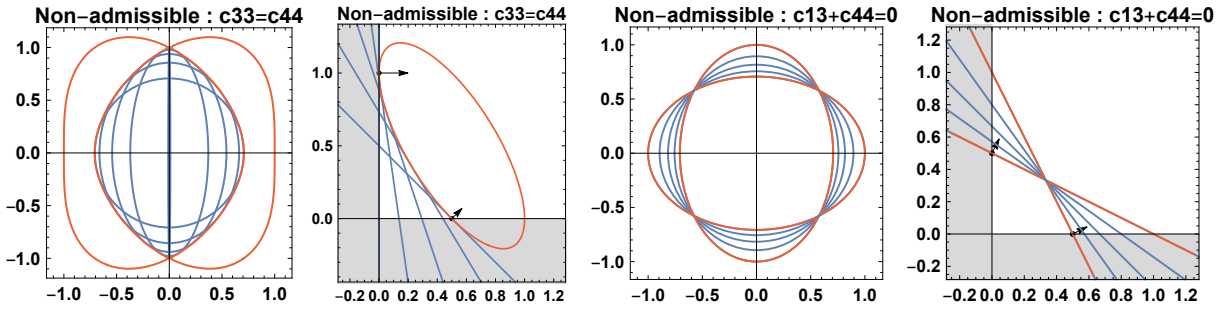

Figure 40: Examples where coefficients $(c_{11}, c_{13}, c_{33}, c_{44})$ *fail* the admissibility conditions in such way that the Hooke elasticity tensor remains positive, but is not separable. As a result the inner slowness surface, associated with pressure waves, is non-smooth and intersects the outer slowness surface, associated with shear waves. Coefficients : $(2, 0, 1, 1)$ for subfig. (i,ii), and $(2, -1, 2, 1)$ for subfig. (iii,iv). Subfigures (i,iii): slowness surfaces (red) and tangent ellipsoids (blue). Subfigures (ii,iv): root domain with the conic $\mathcal{C}$ (red), its tangent lines (blue), and normal vectors of Lemma 8.25 (black arrows).

of $\mathcal{Q}$. A portion of this conic bounds the domain $\mathcal{A}$, which is the image of the set $\mathcal{B}'$ of Proposition 8.21 by a square root transformation.

Our first lemma describes two extremal points of the set $\mathcal{A}$, lying on the coordinate axes.

Lemma 8.25. *Define* $p_* := (1/c_{11}, 0)$ *and* $p^* := (0, 1/c_{33})$. *Then* $p_*, p^* \in \mathcal{A} \cap \mathcal{C}$ *and*

$$\nabla\mathcal{Q}(p_*) = \big(c_{11}(c_{11} - c_{44}),\ (c_{13} + c_{44})^2 + (c_{11} - c_{44})c_{44}\big)\,/c_{11},$$
$$\nabla\mathcal{Q}(p^*) = \big((c_{13} + c_{44})^2 + (c_{33} - c_{44})c_{44},\ c_{33}(c_{33} - c_{44})\big)\,/c_{33}.$$

If follows that $\nabla\mathcal{Q}(p_*), \nabla\mathcal{Q}(p^*) \in]0, \infty[^2$.

Proof. By symmetry and w.l.o.g., we limit our attention to p_*. The polynomial $1 - \mathcal{Q}(x, 0) = c_{11}c_{44}x^2 - (c_{11}+c_{44})x + 1$ admits the two roots $1/c_{11}$ and $1/c_{44}$. Since $c_{11} > c_{44}$, by admissibility, one has $\mathcal{Q}(x, 0) \le 1$ iff $x \in]-\infty, 1/c_{11}] \cup [1/c_{44}, \infty[$ and thus $(1/c_{11}, 0) \in \mathcal{A}$. A direct computation yields the announced expression of $\nabla\mathcal{Q}(p_*)$, and the positivity of its components follows again from the admissibility conditions $c_{11} > c_{44}$, $c_{33} > c_{44}$ and $c_{44} > 0$. □

Corollary 8.26. *If* $\mathcal{I} \ge 0$ *then* $\mathcal{A}$ *is convex, and if* $\mathcal{I} \le 0$ *then* $\mathbb{R}^2_+ \setminus \mathcal{A}$ *is convex, where*

$$\mathcal{I} := c_{11}c_{33} - c_{13}^2 - c_{11}c_{44} - 2c_{13}c_{44} - c_{33}c_{44}, \qquad \mathcal{I} \propto \det(\nabla\mathcal{Q}(p_*), \nabla\mathcal{Q}(p^*)). \tag{157}$$

Proof. By Proposition 8.18 one has $\mathcal{A} = \{r(\cos\theta, \sin\theta);\ 0 \le \theta \le \pi/2,\ 0 \le r \le R(\theta)\}$. Therefore the boundary $\partial\mathcal{A}$ is the union of the two segments $[(0, 0), p_*]$ and $[(0, 0), p^*]$, and of the portion of conic $\mathcal{A} \cap \mathcal{C} = \{R(\theta)(\cos\theta, \sin\theta);\ 0 \le \theta \le \pi/2\}$. Likewise $\partial(\mathbb{R}^2_+ \setminus \mathcal{A}) = [p_*, (\infty, 0)[\cup [p^*, (0, \infty)[\cup (\mathcal{A} \cap \mathcal{C})$ is the union of two half lines (with the obvious notation) and of the same portion of conic.

A direct computation yields the determinant of $\nabla\mathcal{Q}(p_*)$ and $\nabla\mathcal{Q}(p^*)$, which are normal vectors to $\mathcal{C}$ oriented outwards of $\mathcal{A}$, at the endpoints p_* and p^* of $\mathcal{A} \cap \mathcal{C}$. More precisely

$$\det(\nabla\mathcal{Q}(p_*), \nabla\mathcal{Q}(p^*)) = \mathcal{I}\mathcal{J}, \quad \text{where } \mathcal{J} := \big((c_{13} + c_{44})^2 + c_{11}c_{33} - c_{44}^2\big)\,/(c_{11}c_{33}) > 0.$$

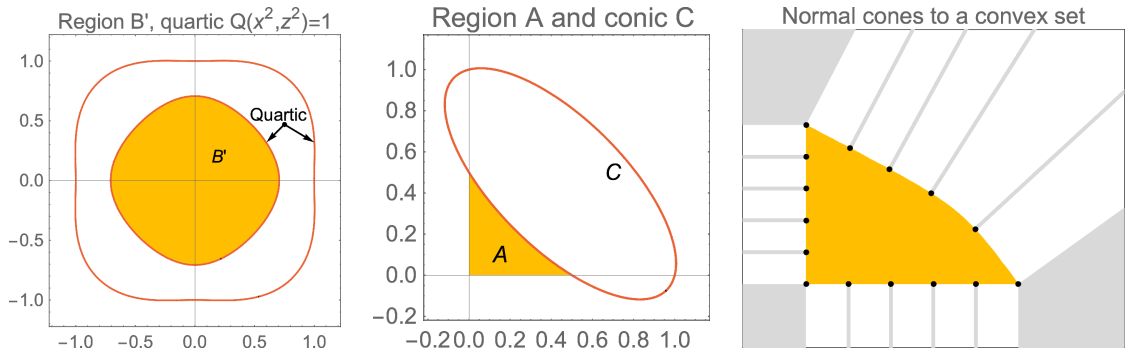

Figure 41: (Left) Quartic slowness surfaces, and xz-slice $\mathcal{B}'$ of the anelliptic ball $\mathcal{B}$. (Center) Region $\mathcal{A}$ and curve $\mathcal{C}$ in the root domain. (Right) Normal cones, shown gray, to a convex set.

In order to establish the announced convexity properties, we distinguish two cases:

- Case of a degenerate conic $\mathcal{C}$ (the union of two lines). Then $\mathcal{A} \cap \mathcal{C} = [p_*, p^*]$ is a straight segment. Indeed, by Proposition 8.18, either the two lines are parallel, or their intersection lies outside $[0, \infty[^2$. As a result $\mathcal{I} = 0$ (since the normal along a line is constant) and both $\mathcal{A}$ and $\mathbb{R}^2_+ \setminus \mathcal{A}$ are convex, as announced.

- Case of a non-degenerate conic (ellipse, hyperbola, parabola). Since a conic is a curve of degree two, it has no inflection point. Therefore the sign of the curvature is constant along $\mathcal{C}$, and thus either $\mathcal{A}$ or $\mathbb{R}^2_+ \setminus \mathcal{A}$ is convex, recalling the endpoint normals $\nabla\mathcal{Q}(p_*), \nabla\mathcal{Q}(p^*) \in]0, \infty[^2$. Since the normal vectors along the boundary of a convex set are ordered trigonometrically in clockwise order, we obtain that $\det(\nabla\mathcal{Q}(p_*), \nabla\mathcal{Q}(p^*)) \geq 0$ if $\mathcal{A}$ is convex (resp. $\det(\nabla\mathcal{Q}(p^*), \nabla\mathcal{Q}(p_*)) \leq 0$ if $\mathbb{R}^2 \setminus \mathcal{A}$ is convex), which concludes the proof. □

8.2.3 Properties and computation of $\mu(\alpha)$

We establish Theorem 8.3 which describes the set $\mathcal{A}$ as an intersection or a union of triangles, and thus $\mathcal{B}$ as an intersection or a union of ellipses see Figure 37, whose size is determined by a function $\mu : [\alpha_*, \alpha^*] \to]0, \infty[$. We also prove that μ is either convex or concave. The argument relies on Proposition 8.27 which is an elementary result on the support function of a convex set, see [BL10] for more detail on this rich subject. In the second part of this subsection, we establish that μ is smooth and provide expressions of μ, α_*, α^* suitable for numerical implementation.

Proposition 8.27. *The support function $\mu_K : \mathbb{R}^d \to]-\infty, \infty]$, of a closed and convex set $K \subset \mathbb{R}^d$, is defined for all $v \in \mathbb{R}^d$ as*

$$\mu_K(v) := \sup_{p \in K} \langle v, p \rangle.$$

This function is convex and lower semi-continuous (l.s.c.), and furthermore

$$K = \{p \in \mathbb{R}^d;\ \forall v \in V,\ \langle v, p \rangle \leq \mu_K(v)\}, \tag{158}$$

provided the set $V \subset \mathbb{R}^d$ contains a generator of each extreme ray of each normal cone to K.

The proof of Proposition 8.27 is postponed to the end of this section. We obtain in Corollary 8.28 and Corollary 8.29 two descriptions of the set $\mathcal{A}$, announced in Theorem 8.3, concluding its proof except for the smoothness of the function μ which follows from the explicit expression (167) below. They are deduced from the description (158) of convex sets as half-space intersections, and from the fact established in Corollary 8.26 that $\mathcal{A}$ is either convex or the complement of a convex set. The endpoints p_*, p^* of the portion of conic $\mathcal{A} \cap \mathcal{C}$ are defined in Lemma 8.25. We denote by $\mathrm{Cone}(E) = \{\sum_{i=1}^I \lambda_i e_i;\ I \geq 0,\ \lambda_1, \cdots, \lambda_I \geq 0,\ e_1, \cdots, e_I \in E\}$ the convex cone generated by non-negative linear combinations within a set of vectors E.

Corollary 8.28. *Assume that $\mathcal{A}$ is convex, which corresponds to the case (max). Define $0 < \alpha_* \leq \alpha_* < 1$ by $\nabla \mathcal{Q}(p_*) \propto (1 - \alpha_*, \alpha_*)$ and $\nabla \mathcal{Q}(p^*) \propto (1 - \alpha^*, \alpha^*)$, and let $\mu(\alpha) := \mu_{\mathcal{A}}(1 - \alpha, \alpha)$. Then μ is convex and $\mathcal{A} = \{p \in \mathbb{R}^2_+;\ \forall \alpha \in [\alpha_*, \alpha^*], \langle (1 - \alpha, \alpha), p \rangle \leq \mu(\alpha)\}$.*

Proof. Denote $\nabla \mathcal{Q}_*(p_*) = (v_1, v_2)$, and note that $v_1, v_2 > 0$ by Lemma 8.25. Then the positive proportionality relation $\nabla \mathcal{Q}_*(p_*) \propto (1 - \alpha_*, \alpha_*)$ admits the unique solution $\alpha_* := v_2/(v_1 + v_2) \in]0, 1[$. Likewise $\alpha^* \in]0, 1[$, and furthermore by Corollary 8.26 we obtain as announced

$$0 \leq \det\big(\nabla \mathcal{Q}(p_*), \nabla \mathcal{Q}(p^*)\big) \propto \big[(1 - \alpha_*)\alpha^* - \alpha_*(1 - \alpha^*)\big] = \alpha^* - \alpha_*. \tag{159}$$

The function $\mu(\alpha) := \mu_{\mathcal{A}}(1 - \alpha, \alpha)$ is convex since it is the composition of $\mu_{\mathcal{A}}$, which is convex by Proposition 8.27, with an affine mapping. Observing that $\mathcal{A}$ is closed and convex, we obtain

$$\begin{gathered} \mathcal{A} = \{p \in \mathbb{R}^2;\ \forall v \in V, \langle v, p \rangle \leq \mu_{\mathcal{A}}(v)\}, \\ \text{with } V := \{(-1, 0), (0, -1)\} \cup \{(\alpha, 1 - \alpha); \alpha_* \leq \alpha \leq \alpha^*\}. \end{gathered}$$

by Proposition 8.27, which implies the announced expression of $\mathcal{A}$. To show that V obeys the assumption of Proposition 8.27, we describe the normal cones to $\mathcal{A}$, illustrated on Figure 41, for all points of the boundary $\partial\mathcal{A} = [(0, 0), p_*] \cup [(0, 0), p^*] \cup (\mathcal{C} \cap \mathcal{A})$. At the corners one has $N_{\mathcal{A}}(0, 0) = \mathbb{R}^2_- = \mathrm{Cone}\{(-1, 0), (0, -1)\}$, $N_{\mathcal{A}}(p_*) = \mathrm{Cone}\{(-1, 0), (1 - \alpha_*, \alpha_*)\}$, and $N_{\mathcal{A}}(p^*) = \mathrm{Cone}\{(-1, 0), (1 - \alpha^*, \alpha^*)\}$. On the straight segments $N_{\mathcal{A}}(p) = \mathrm{Cone}\{(-1, 0)\}$ for all $p \in](0, 0), p_*[$, and $N_{\mathcal{A}}(p) = \mathrm{Cone}\{(0, -1)\}$ for all $p \in](0, 0), p^*[$. Finally for $p \in (\mathcal{A} \cap \mathcal{C}) \setminus \{p_*, p^*\}$ one has $N_{\mathcal{A}}(p) = \mathrm{Cone}\{\nabla \mathcal{Q}(p)\} = \mathrm{Cone}\{(1 - \alpha, \alpha)\}$ for some $\alpha_* \leq \alpha \leq \alpha^*$, since $(1, 0) \preceq \nabla \mathcal{Q}(p_*) \preceq \nabla \mathcal{Q}(p) \preceq \nabla \mathcal{Q}(p^*) \preceq (0, 1)$ in the circular trigonometric ordering of vectors, by convexity of $\mathcal{A}$. The result follows. □

Corollary 8.29. *Assume that $\mathbb{R}^2_+ \setminus \mathcal{A}$ is convex, which corresponds to case (min). Define $0 < \alpha_* \leq \alpha^* < 1$ by $\nabla \mathcal{Q}(p^*) \propto (1 - \alpha_*, \alpha_*)$ and $\nabla \mathcal{Q}(p_*) \propto (1 - \alpha^*, \alpha^*)$, and let $\mu(\alpha) := -\mu_{\mathcal{A}^c}(-(1 - \alpha, \alpha))$ where $\mathcal{A}^c := \overline{\mathbb{R}^2_+ \setminus \mathcal{A}}$. Then μ is concave and $\mathcal{A} = \{p \in \mathbb{R}^2_+;\ \exists \alpha \in [\alpha_*, \alpha^*], \langle (1 - \alpha, \alpha), p \rangle \leq \mu(\alpha)\}$.*

Proof. Similarly to the proof of Corollary 8.28, one has $\alpha_*, \alpha^* \in]0, 1[$ by positivity of the gradient coordinates, see Lemma 8.25, and $0 \geq \det(\nabla \mathcal{Q}(p_*), \nabla \mathcal{Q}(p^*)) \propto (\alpha_* - \alpha^*)$. The

function μ is concave since it is the opposite of $\mu_{\mathcal{A}^c}$, which is convex by Proposition 8.27, composed with an affine mapping. In addition

$$\mathcal{A}^c = \{p \in \mathbb{R}^2;\, \forall v \in V, \langle v, p\rangle \leq \mu_{\mathcal{A}^c}(v)\}, \tag{160}$$

$$\mathbb{R}^2 \setminus \mathcal{A}^c = \{p \in \mathbb{R}^2;\, \exists v \in V, \langle v, p\rangle > \mu_{\mathcal{A}^c}(v)\}, \tag{161}$$

with $V = \{(-1,0),(0,-1)\} \cup \{-(1-\alpha,\alpha);\, \alpha_* \leq \alpha \leq \alpha^*\}$. Equation (160) follows from Proposition 8.27, where the assumption on V is checked as in Corollary 8.28 since $\partial\mathcal{A}^c = [p_*, (\infty,0)[\cup[p^*,(0,\infty)[\cup(\mathcal{A}\cap\mathcal{C})$. Equation (161) is obtained by taking the complement. Noting that $\mathcal{A} = \mathbb{R}^2_+ \cap (\overline{\mathbb{R}^2 \setminus \mathcal{A}^c})$ we conclude the proof. □

Explicit formulas for implementation. In the rest of this section, we obtain explicit formulas for the function μ and the bounds α_* and α^*, suitable for implementing our numerical scheme, and announced below Theorem 8.3. For that purpose, we rewrite the quadratic function $\mathcal{Q}$ defining the eikonal equation (126) in the following form:

$$\mathcal{Q}(p) = \langle l, p\rangle + \tfrac{1}{2}\langle p, Qp\rangle, \qquad \text{hence } \nabla\mathcal{Q}(p) = l + Qp, \tag{162}$$

where $l \in \mathbb{R}^2$ and $Q \in S_2$ is a symmetric matrix. Indeed, our numerical implementation relies on the linear and quadratic forms defined by l and Q, rather than the coefficients $(c_{11}, c_{13}, c_{33}, c_{44})$ which lead to expressions more complicated and of higher algebraic degree, and also restrict the generality. We nevertheless assume that all the guarantees derived previously apply.

By computing the smallest root x_* of the quadratic equation $1 = \mathcal{Q}(x,0) = l_1 x + Q_{11}x^2/2$, we obtain the endpoint $p_* = (x_*, 0)$ of $\mathcal{A}\cap\mathcal{C}$. Likewise we obtain the second endpoint p^*, and then $\nabla\mathcal{Q}(p_*)$ and $\nabla\mathcal{Q}(p^*)$ by (162, right). The cases (max) or (min) of Theorem 8.3 can be distinguished by computing the sign of

$$\det(\nabla\mathcal{Q}(p_*), \nabla\mathcal{Q}(p^*)). \tag{163}$$

Also α_* and α^* are trivially obtained from $\nabla\mathcal{Q}(p_*)$ and $\nabla\mathcal{Q}(p^*)$, as in Corollary 8.28 and Corollary 8.29.

In the rest of this section, we fix $\alpha \in]\alpha_*, \alpha^*[$ and we denote $\boldsymbol{\alpha} := (1-\alpha, \alpha)$. Then one has

$$\mu(\alpha) = \max_{p\in\mathcal{A}\cap\mathcal{C}} \langle \boldsymbol{\alpha}, p\rangle \qquad \Big(\text{resp. } \mu(\alpha) = \min_{p\in\mathcal{A}\cap\mathcal{C}} \langle \boldsymbol{\alpha}, p\rangle\Big), \tag{164}$$

in case (max) (resp. (min)). Indeed, the formula (164) is equivalent to the definitions presented in Corollary 8.28 and Corollary 8.29 involving the support function of the set $\mathcal{A}$ (resp. $\overline{\mathbb{R}^2 \setminus \mathcal{A}}$), because $\boldsymbol{\alpha}$ (resp. $-\boldsymbol{\alpha}$) is proportional their exterior normal at some point of the boundary $\mathcal{A}\cap\mathcal{C}$. In addition, since $\alpha \in]\alpha_*, \alpha^*[$, the extremum (164) is attained at a point $p \in \mathcal{A}\cap\mathcal{C}$ distinct from the endpoints p_* and p^*, hence to which the method of Lagrange multipliers is applicable (as opposed to the more complex KKT relations). As a result, and recalling that $\mathcal{C} = \{\mathcal{Q} = 1\}$, there exists a Lagrange multiplier $\lambda \in \mathbb{R}$ such that

$$\nabla\mathcal{Q}(p) = \lambda\boldsymbol{\alpha}, \qquad \mathcal{Q}(p) = 1. \tag{165}$$

Rewriting (165, left) as as $l + Qp = \lambda \boldsymbol{\alpha}$, yields

$$p = Q^{-1}(\lambda \boldsymbol{\alpha} - l). \tag{166}$$

We assume here that Q is an invertible matrix, which actually is not ensured by our admissibility assumptions (131). Further discussion of the case where Q is not invertible is postponed to the end of this section. Rewriting (132, right) as $1 = \langle l, p\rangle + \frac{1}{2}\langle p, Qp\rangle$ and inserting (166) yields

$$1 = \langle l, Q^{-1}(\lambda \boldsymbol{\alpha} - l)\rangle + \tfrac{1}{2}\langle (\lambda \boldsymbol{\alpha} - l), Q^{-1}(\lambda \boldsymbol{\alpha} - l)\rangle = \tfrac{1}{2}\lambda^2 \langle \boldsymbol{\alpha}, Q^{-1}\boldsymbol{\alpha}\rangle - \tfrac{1}{2}\langle l, Q^{-1} l\rangle.$$

Therefore,

$$\lambda^2 = \frac{2 + \langle l, Q^{-1} l\rangle}{\langle \boldsymbol{\alpha}, Q^{-1}\boldsymbol{\alpha}\rangle}.$$

By (165, left), the scalar λ is the proportionality coefficient between the gradient $\nabla \mathcal{Q}(p)$, which has positive components for all $p \in \mathcal{A} \cap \mathcal{C}$, and the vector $\boldsymbol{\alpha} = (1 - \alpha, \alpha)$ which is likewise positive. Thus $\lambda > 0$, and the Lagrange multiplier λ is fully determined. Therefore

$$\mu(\alpha) := \langle \boldsymbol{\alpha}, p\rangle = \lambda \langle \boldsymbol{\alpha}, Q^{-1}\boldsymbol{\alpha}\rangle - \langle \boldsymbol{\alpha}, Q^{-1} l\rangle = \varepsilon \sqrt{\langle \boldsymbol{\alpha}, Q^{-1}\boldsymbol{\alpha}\rangle (2 + \langle l, Q^{-1} l\rangle)} - \langle \boldsymbol{\alpha}, Q^{-1} l\rangle. \tag{167}$$

We denoted by $\varepsilon \in \{-1, 1\}$ the sign of $\langle \boldsymbol{\alpha}, Q^{-1}\boldsymbol{\alpha}\rangle$, which by (165) is also the sign of the expression

$$\langle \nabla \mathcal{Q}(p), Q^{-1} \nabla \mathcal{Q}(p)\rangle = \langle Qp + l, Q^{-1}(Qp + l)\rangle = \langle p, Qp\rangle + 2\langle p, l\rangle + \langle l, Q^{-1} l\rangle = 2 + \langle l, Q^{-1} l\rangle.$$

In particular, ε is independent of α, and can be determined in advance from the coefficients Q and l of the PDE. From (167) we obtain that μ has C^∞ regularity, as announced in Theorem 8.3, and is computable in a straightforward manner. We also recover the fact that it must be convex or concave, by an immediate application of the following lemma to the polynomial $P(\alpha) := \langle \boldsymbol{\alpha}, Q^{-1}\boldsymbol{\alpha}\rangle$.

Lemma 8.30. *Let $P(t) := at^2 + bt + c$ be a second degree polynomial, with discriminant $\Delta := b^2 - 4ac$. If $\Delta \geq 0$ (resp. $\Delta \leq 0$) then $\sqrt{P}$ is concave (resp. convex) on each connected component of $\{t \in \mathbb{R}; P(t) > 0\}$.*

Proof. This follows from a direct computation: assuming $P(t) > 0$ one obtains

$$\frac{d^2}{dt^2}\sqrt{P(t)} = \frac{2P''(t)P(t) - P'(t)^2}{4P(t)^{\frac{3}{2}}} = \frac{2 \times 2a(at^2 + bt + c) - (2at + b)^2}{4P(t)^{\frac{3}{2}}} = -\frac{\Delta}{4P(t)^{\frac{3}{2}}}. \qquad \square$$

Case of a singular matrix Q. We denote by J the adjugate matrix of Q, and let $\delta := \det Q$ and $\varepsilon' = \operatorname{sign}(2\delta + \langle l, Jl\rangle)$. Note that $Q^{-1} = J/\delta$ if $\delta \neq 0$. We obtain from (167) and after straightforward manipulations (namely, the multiplication by the conjugate root) the following alternative expression of μ. Its numerical evaluation is usually more stable than (167) when Q is singular or almost singular, since it does not involve Q^{-1}

$$\mu(\alpha) = \frac{\det(\boldsymbol{\alpha}, l)^2 + 2\langle v, Jv\rangle}{\varepsilon' \sqrt{\langle \boldsymbol{\alpha}, J\boldsymbol{\alpha}\rangle (2\delta + \langle l, Jl\rangle)} + \langle \boldsymbol{\alpha}, Jl\rangle}. \tag{168}$$

One still needs to handle separately the degenerate case, at the intersection of the (max) and (min) cases, where $\alpha_* = \alpha^*$ (in that case the conic $\mathcal{C}$ is a union of two lines, and the TTI ball $\mathcal{B}$ is an ellipsoid rather than a quartic surface).

Proof of Proposition 8.27. The function μ_K is convex (resp. l.s.c.) since it is defined as the supremum of a family of linear functions, which are convex (resp. continuous hence l.s.c.) by definition. In the following, we denote by $P_K : \mathbb{R}^d \to K$ the orthogonal projection, and by $N_K(p_*) \subset \mathbb{R}^d$ the normal cone at a point $p_* \in K$, illustrated on Figure 41, which admits the following equivalent characterizations:

$$v \in N_K(p_*) \;\Leftrightarrow\; \forall p \in K, \langle v, p_* - p\rangle \geq 0 \;\Leftrightarrow\; \mu_K(v) = \langle v, p_*\rangle \;\Leftrightarrow\; P_K(p_* + v) = p_*. \quad (169)$$

Denote (158, r.h.s.) by $\tilde{K}$, and note that $K \subset \tilde{K}$ by definition of the support function. In the following, we consider $p \notin K$, and denote $p_* := P_K(p)$ and $v := p - p_*$. For any $q \in K$ one has $\langle p - p_*, p_* - q\rangle \geq 0$ by general properties of the orthogonal projection, therefore $\langle v, p\rangle \geq \|v\|^2 + \langle v, q\rangle$ by rearranging terms, and thus by taking the supremum over $q \in K$

$$\langle v, p\rangle \geq \|v\|^2 + \mu_K(v) > \mu_K(v). \quad (170)$$

Since $P_K(p_* + v) = P_K(p) = p_*$, one has $v \in N_K(p_*)$ by (169). By the Krein-Milman theorem, the cone $N_K(p_*)$ is the convex hull of its extreme rays, and thus by assumption there exists $\lambda_1, \cdots, \lambda_N \geq 0$ and $v_1, \cdots, v_N \in V \cap N_K(p_*)$ such that $v = \sum_{n=1}^N \lambda_n v_n$, for some $N \geq 0$. Then, assuming *for contradiction* that $p \in \tilde{K}$ we obtain:

$$\langle v, p\rangle = \sum_{1\leq n\leq N} \lambda_n \langle v_n, p\rangle \leq \sum_{1\leq n\leq N} \lambda_n \mu_K(v_n) = \sum_{1\leq n\leq N} \lambda_n \langle v_n, p_*\rangle = \langle v, p_*\rangle = \mu_K(v). \quad (171)$$

We used successively (i) linearity, (ii) the assumption $p \in \tilde{K}$, (iii) the normal cone characterization (169), (iv) linearity again, and (v) again (169). Noting that (171) contradicts (170), we must have $p \notin \tilde{K}$. We have shown that $p \notin K \Rightarrow p \notin \tilde{K}$, which establishes the reverse inclusion $K \supset \tilde{K}$, and concludes the proof. □

8.3 Quasi-convexity or quasi-concavity of the update operator

We present two constructions of quasi-convex and quasi-concave functions in Section 8.3.1. By an adequate choice of parameters, they encompass the update operator associated to our finite differences discretization of the TTI eikonal PDE, which establishes Theorem 8.11. We study the primal metric associated to a TTI model using a similar strategy in Section 8.3.2, thus establishing Corollary 8.4. Interestingly, the proof differs in the (max) and (min) cases, a discrepancy also encountered in the convergence analysis Section 8.4.

8.3.1 Two constructions of quasi-convex and quasi-concave functions.

We show that quasi-convex and quasi-concave functions, introduced in Definition 8.10, can be obtained as ratios of suitable functions in Lemma 8.31, and as implicit functions in Proposition 8.32.

For that purpose, we fix a convex subset A of a vector space, and recall from Definition 8.10 that a map $f : A \to \mathbb{R}$ is quasi-convex iff $\{x \in A; f(x) \leq \lambda\}$ is a convex set for

all $\lambda \in \mathbb{R}$. Likewise we say that f is quasi-concave if $-f$ is convex, equivalently iff $\{x \in A; f(x) \geq \lambda\}$ is a convex set for all $\lambda \in \mathbb{R}$.

Lemma 8.31. *If $f : A \to [0, \infty[$ is convex, and $g : A \to]0, \infty[$ is concave, then f/g is quasi-convex. Likewise if $f : A \to]0, \infty[$ is convex, and $g : A \to [0, \infty[$ is concave, then g/f is quasi-concave.*

Proof. We only prove the first statement, since the second one is similar. Let $\lambda \in \mathbb{R}$. If $\lambda < 0$, then $\{f/g \leq \lambda\} = \emptyset$ is convex. Otherwise $\lambda \geq 0$ and $\{f/g \leq \lambda\} = \{f - \lambda g \leq 0\}$ is convex since $f - \lambda g$ is convex. □

Proposition 8.32. *Let $F : A \times \mathbb{R} \to \mathbb{R}$ be such that (i) $\alpha \in A \mapsto F(\alpha, \lambda)$ is quasi-convex (resp. quasi-concave) for all $\lambda \in \mathbb{R}$, and that (ii) $\lambda \in \mathbb{R} \mapsto F(\alpha, \lambda)$ is non-decreasing for all $\alpha \in A$. Assume also that (iii) $F(\alpha, \Lambda(\alpha)) = 0$ admits for all $\alpha \in A$ a unique solution $\Lambda(\alpha)$, thus defining a mapping $\Lambda : A \to \mathbb{R}$. Then Λ is quasi-concave (resp. quasi-convex).*

Proof. We limit our attention to the case where $\alpha \mapsto F(\alpha, \lambda)$ is quasi-convex, since the second case is similar. By (ii) and (iii) one obtains $\Lambda(\alpha) \geq \lambda \Leftrightarrow F(\alpha, \lambda) \leq 0$, for any $\alpha \in A$, $\lambda \in \mathbb{R}$. Thus

$$\{\alpha \in A;\, \Lambda(\alpha) \geq \lambda\} = \{\alpha \in A; F(\alpha, \lambda) \leq 0\},$$

for any $\lambda \in \mathbb{R}$. Noting by (i) that the r.h.s. is a convex set, we obtain that Λ is quasi-concave, which concludes the proof. □

Proof of Theorem 8.11, on the update operator quasi-convexity or quasi-concavity. We proceed to apply Proposition 8.32 to a function of the following form, defined in view of the expression (143) of the numerical scheme for the TTI eikonal PDE,

$$F(\alpha, \lambda) = \frac{1}{\mu(\alpha)} \sum_{1 \leq i \leq I} \rho_i(\alpha) \max\{0, \lambda - u_i\}^2. \tag{172}$$

Specifically using the notations of Proposition 8.7 and (141), we fix $q \in h\mathbb{Z}^d$ and $0 \leq k < K$, define $A = [\alpha_k, \alpha_{k+1}]$ which is a segment of $\mathbb{R}$, and let $u_i = \min\{u(q + he_{ik}), u(q - he_{ik})\}$, for all $1 \leq i \leq I$ where $u : h\mathbb{Z}^d \to]-\infty, \infty]$ is the unknown of the finite difference scheme. (Recall that u is finite on Ω_h and extended by $+\infty$ elsewhere.) Then $F(\alpha, \lambda) = \hat{\mathfrak{F}}^\alpha(q, [\lambda - u(r)]_{r \in \overline{X}})$ is the numerical scheme (143) with the base point value $u(q)$ replaced with the unknown λ, consistently with the formulation of the update operator (138).

By Theorem 8.3 the function $\mu : A \to]0, \infty[$ is convex in case (max) (resp. concave in case (min)). By Proposition 8.7 the functions $\rho_i : A \to [0, \infty[$ are affine. For any given $\lambda \in \mathbb{R}$ the sum $f(\alpha) := \sum_{i=1}^I \rho_i(\alpha) \max\{0, \lambda - u_i\}^2$ is thus non-negative and affine w.r.t. $\alpha \in A$, thus simultaneously convex and concave. Lemma 8.31 therefore yields that $\alpha \in A \mapsto F(\alpha, \lambda)$ is quasi-concave in case (max) (resp. quasi-convex in case (min)).

The partial mapping $\lambda \in \mathbb{R} \mapsto F(\alpha, \lambda)$ is non-decreasing, for any $\alpha \in A$, since the weights $\rho_i(\alpha)$ are non-negative and since $\lambda \mapsto \max\{0, \lambda\}^2$ is non-decreasing. As already observed in Section 8.1.2 there is a unique solution $\Lambda : A \to \mathbb{R}$ to the equation $F(\alpha, \Lambda(\alpha)) = 1$ (one has $\Lambda(\alpha) := \Lambda^\alpha u(q)$ with the notations of Theorem 8.11). Applying Proposition 8.32 to $F - 1$, we obtain that the update operator Λ is quasi-convex in case (max) (resp. quasi-concave in case (min)), which concludes the proof of Theorem 8.11.

8.3.2 Expression of the norm value, gradient, and dual.

Given admissible TTI parameters $\sigma \in \mathbb{R}^3$, and a co-vector $p = (p_x, p_y, p_z) \in \mathbb{R}^3 \setminus \{0\}$, we obtain

$$\begin{aligned}
\mathcal{F}^*_\sigma(p) &:= \min\{\nu > 0;\, p/\nu \in \mathcal{B}_\sigma\} \\
&= \min\{\nu > 0;\, (p_x^2 + p_y^2,\, p_z^2)/\nu^2 \in \mathcal{A}_\sigma\} \\
&= \min\{\nu > 0;\, \forall \alpha \in [\alpha_*, \alpha^*],\, (1-\alpha)(p_x^2 + p_y^2) + \alpha p_z^2 \leq \nu^2 \mu(\alpha)\} \qquad (173) \\
&= \max_{\alpha \in [\alpha_*, \alpha^*]} \sqrt{\frac{(1-\alpha)(p_x^2 + p_y^2) + \alpha p_z^2}{\mu(\alpha)}}, \qquad (174)
\end{aligned}$$

assuming case (max) of Theorem 8.3 in (173). We used successively (i) the norm definition (128, right), (ii) the definitions (132) and (127) of the sets $\mathcal{A}_\sigma$ and $\mathcal{B}_\sigma$, (iii) Theorem 8.3 in case (max), and (iv) a direct algebraic computation. Alternatively, in case (min) of Theorem 8.3, the universal quantifier $\forall$ of (173) is replaced with an existential quantifier $\exists$, and as a result the max operator in (174) is replaced with the min operator. The announced expression (134) of $\mathcal{F}^*(p) := \mathcal{F}^*_\sigma(Rp)$ follows. The expression (135, left) of the gradient $\nabla \mathcal{F}^*(p)$ then follows from the envelope theorem [Car01, Theorem 6.1], on the differentiation of functions defined as an extremum.

We turn to the computation of the dual norm (135, right), which is obtained as follows

$$\begin{aligned}
\frac{1}{2}\mathcal{F}(v)^2 &= \max_p \left(\langle p, v \rangle - \frac{1}{2}\mathcal{F}^*(p)^2 \right) \\
&= \max_p \; \overline{\min_{\alpha \in [\alpha_*, \alpha^*]}} \left(\langle p, v \rangle - \frac{1}{2\mu(\alpha)} \|p\|^2_{D(\alpha)} \right) \\
&= \overline{\min_{\alpha \in [\alpha_*, \alpha^*]}} \max_p \left(\langle p, v \rangle - \frac{1}{2\mu(\alpha)} \|p\|^2_{D(\alpha)} \right) \qquad (175) \\
&= \overline{\min_{\alpha \in [\alpha_*, \alpha^*]}} \frac{\mu(\alpha)}{2} \|v\|^2_{D(\alpha)^{-1}}.
\end{aligned}$$

We used successively, (i) Legendre-Fenchel duality, which is a generalization of norm duality, (ii) the explicit expression of $\mathcal{F}^*$, recalling that $\{\min, \overline{\min}\} = \{\min, \max\}$, see (133), (iii) an interversion of the extremum operators max and $\overline{\min}$, discussed in more detail below, and (iv) the known explicit expression of the Legendre-Fenchel dual of the positive quadratic form $p \mapsto \langle p, D(\alpha)p \rangle / \mu(\alpha)$.

As announced, we discuss in more detail the interversion of max and $\overline{\min}$ in (175), and in this occasion we need to distinguish the treatment of the (min) and (max) case, associated with Theorem 8.3. If $\overline{\min} = \max$, then (175) amounts to a maximization over the joint variable $(p, \alpha) \in \mathbb{R}^3 \times [\alpha_*, \alpha^*]$, hence the order of the maximizations is irrelevant and the result is proved. On the other hand, if $\overline{\min} = \min$, then we invoke Sion's minimax theorem [Kom88] to exchange the ordering of the min and max operators, whose assumptions are checked below. Define, for $\alpha \in [\alpha_*, \alpha^*]$ and $p \in \mathbb{R}^3$

$$F(\alpha, p) := \langle p, v \rangle - \frac{1}{2} F_0(\alpha, p), \qquad \text{where } F_0(\alpha, p) := \frac{1}{\mu(\alpha)} \|p\|^2_{D(\alpha)}.$$

Then $p \mapsto F_0(\alpha, p)$ is a positive quadratic form, hence is a convex function. On the other hand $\alpha \mapsto F_0(\alpha, p)$ is quasi-concave by Lemma 8.31, since it is the ratio of the non-negative and affine (hence concave) function $\|p\|^2_{D(\alpha)} = (1-\alpha)\|p\|^2_{D_0} + \alpha\|p\|^2_{D_1}$, divided by $\mu(\alpha)$ which is convex by Theorem 8.3 in case (max) (recall that we assume $\overline{\text{mix}} = \min$ here, and see (133)). The function F thus matches the assumptions of Sion's minimax theorem, as announced: F is quasi-convex w.r.t. α, concave w.r.t. p (hence also quasi-concave), and in addition we note that F is continuous, that $[\alpha_*, \alpha^*]$ is convex and compact, and that $\mathbb{R}^3$ is convex. This completes the proof.

8.4 Convergence analysis

We prove that the solutions to our discretization of the TTI eikonal PDE, obey a Lipschitz regularity property in the (max) case, and a weaker growth estimate in the (min) case, from which we deduce their convergence as the grid scale is refined, see Theorem 8.33. This discrepancy between the (max) and (min) cases illustrates the fact that PDE operators presented in (generalized) Bellman form, i.e. as a maximum of simpler monotone operators, are usually more easily amenable to analysis than those presented as a minimum. Such Bellman forms are at the foundation of multistencil fast marching methods [HF07], and of discretizations of second order PDEs with general coefficients [Kry05], as well as the special cases of the Monge-Ampere [BM21] or Pucci [BBM20] PDEs. In the case of the TTI eikonal PDE, we only obtain a Bellman form in the (max) case, and by Theorem 8.3 this depends on the model coefficients.

We believe that the difference between the (max) and (min) cases, both in the proof technique and in the obtained regularity results (177) and (178), is interesting since it departs from the symmetrical treatment of these two cases in the introduction. Nevertheless, we do establish in both cases the convergence to the viscosity solution of the TTI eikonal PDE as the grid scale is refined, at least in the simplified setting of null Dirichlet boundary conditions on $\partial\Omega$ (as opposed to the point source and outflow boundary conditions often considered in applications), see Theorem 8.33. Let us also mention that, empirically, our numerical experiments in Section 8.5 do not show a difference in behavior between the (max) and (min) cases.

We fix an open and bounded domain $\Omega \subset \mathbb{R}^d$, where $d \in \{2, 3\}$, with a smooth boundary. Given $h > 0$ we let $\Omega_h := \Omega \cap h\mathbb{Z}^d$, and $\partial\Omega_h := h\mathbb{Z}^d \setminus \Omega_h$. The notation $C = C(\Omega, \sigma, R)$ means that the constant C only depends on the specified parameters.

Theorem 8.33. *Consider continuous TTI coefficients $\sigma \in C^0(\overline{\Omega}, \mathbb{R}^5)$ obeying the admissibility conditions* (131) *pointwise, and a continuous field of invertible matrices $R \in C^0(\overline{\Omega}, \mathrm{GL}_d(\mathbb{R}))$. Then there exists a unique solution $u : h\mathbb{Z}^d \to [0, \infty[$, to*

$$\mathfrak{F}u(q) = 1, \ \forall q \in \Omega_h, \qquad\qquad u(q) = 0, \ \forall q \in \partial\Omega_h, \tag{176}$$

where $\mathfrak{F}$ stands for the proposed finite differences discretization of the TTI eikonal PDE (143)*. In addition, there exists a constant $C = C(\Omega, \sigma, R)$ such that for any $h > 0$ sufficiently small and for all $q, r \in h\mathbb{Z}^d$.*

- *If the parameters σ fall in the (max) case of Theorem 8.3 over the whole $\overline{\Omega}$, then*

$$|u(q) - u(r)| \leq C|q - r|. \tag{177}$$

- *For arbitrary parameters σ, possibly mixing of the (max) and (min) cases over $\overline{\Omega}$, one has*

$$u(q) \;\leq\; \max\{u(r');\; r' \in h\mathbb{Z}^d,\; |r'-r| \leq Ch\} + C|q-r|. \tag{178}$$

In both cases, one has $\|u_h - \mathbf{u}\|_{L^\infty(\Omega_h)} \to 0$ as $h \to 0$, where $u_h : \Omega_h \to \mathbb{R}$ denotes the discrete solution to (176), and $\mathbf{u} : \overline{\Omega} \to \mathbb{R}$ denotes the unique viscosity solution of the TTI eikonal PDE

$$\mathcal{F}_q^*(\nabla \mathbf{u}(q)) = 1,\ \forall q \in \Omega, \qquad\qquad \mathbf{u}(q) = 0,\ \forall q \in \partial\Omega, \tag{179}$$

where we denoted $\mathcal{F}_q^*(v) := \mathcal{F}_{\sigma(q)}^*(R(q)v)$, for all $q \in \Omega$.

Before turning to the proof, we recall the definition of a sub-solution or super-solution to a numerical scheme, whereas the corresponding PDE notions are briefly evoked below (194).

Definition 8.34. *We say that $u : h\mathbb{Z}^d \to \mathbb{R}$ is a sub-solution to a scheme $\mathfrak{F}$ on Ω_h if*

$$\mathfrak{F}u(q) \leq 1,\ \forall q \in \Omega_h, \qquad\qquad u(q) = 0,\ \forall q \in \partial\Omega_h. \tag{180}$$

Likewise, we define the notions of strict sub-solution, solution, super-solution, and strict super-solution, by replacing the comparison operator with $<$, $=$, $\geq$, $>$, in (180, left) respectively.

Notations. Selling's decomposition is denoted $D = \sum_{e \in \mathbb{Z}^d} \rho(e;D) e e^\top$, consistently with Section 8.B and (201); this notation avoids introducing an arbitrary indexing $(e_i)_{i=1}^I$ of the active offsets $\{e \in \mathbb{Z}^d;\ \rho(e;D) > 0\}$, and is thus more convenient for discussing the regularity of the weights $D \mapsto \rho(e;D)$, see Proposition 8.47. Throughout this section, the quantities associated in Theorem 8.3 to admissible TTI parameters define the following functions pointwise on $\overline{\Omega}$:

$$\alpha_*, \alpha^* \in C^0(\overline{\Omega},]0,1[), \qquad\qquad \mu \in C^0(\mathcal{A},]0,\infty[),$$

where $\mathcal{A} := \{(\alpha, q) \in]0,1[\times \overline{\Omega};\ \alpha_*(q) \leq \alpha \leq \alpha^*(q)\}$. For all $\alpha \in]0,1[$ and $q \in \overline{\Omega}$ we let

$$D(q,\alpha) := R(q)^\top \begin{pmatrix} 1-\alpha & & \\ & 1-\alpha & \\ & & \alpha \end{pmatrix} R(q). \tag{181}$$

8.4.1 Lipchitz property in case (max)

The main result of this subsection is a Lipschitz regularity property for sub-solutions to the numerical scheme $\mathfrak{F}^D$ discretizing the Riemannian eikonal PDE (141), see Proposition 8.36. The Lipschitz estimate (177) in case (max) of Theorem 8.33 is then deduced. Note that the weaker growth estimate (178), valid in all cases, suffices for the proof of convergence in Section 8.4.3. We nevertheless present the Lipschitz estimate since it is simple, expected, and since the proof exploits a number of properties of Selling's matrix decomposition, gathered in Proposition 8.47, which is central in the method. A similar

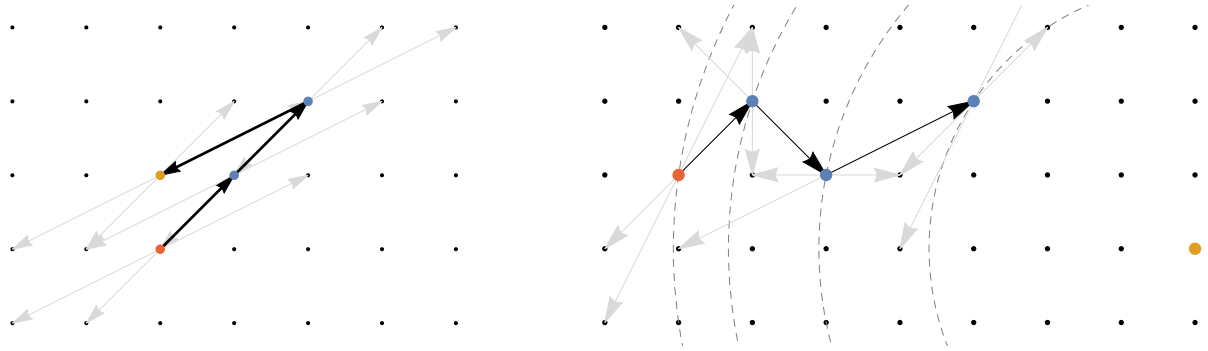

Figure 42: Left: In the (max) case, one can exploit consistency in the scheme stencils, and use the fact that they span $\mathbb{Z}^d$, so as to join two neighbor points using a chain of offsets. Right: In the (min) case, the active stencils at the different points may be uncorrelated. Their offsets can be used to move toward a given target, up to some radius, but not to reach it exactly in general.

Lipschitz regularity result is proved in [Mir14a, Lemma 2.7] for a different discretization of the Riemannian eikonal PDE.

A preliminary technical lemma defines, between any two neighbor points on the grid, a chain whose length is bounded above, and such that successive points are connected by offsets of Selling's decomposition of a given matrix field, see Figure 42 (left).

Lemma 8.35. *Given a field of symmetric positive definite matrices $\mathcal{D} \in C^0(\overline{\Omega}, S_d^{++})$, there exists $h_0 > 0$, $\rho_0 > 0$ and N_0 such that the following holds. Let $0 < h < h_0$, and let $q_*, q^* \in h\mathbb{Z}^d$ be such that $|q_* - q^*| = h$. Then there exists a chain $q_0, \cdots, q_N \in h\mathbb{Z}^d$ of length $N \leq N_0$, whose endpoints are $q_0 = q_*$ and $q_N = q^*$, and signs $\varepsilon_1, \cdots, \varepsilon_n \in \{-1, 1\}$, such that*

$$\rho\big((q_{n+1} - q_n)\varepsilon_n/h;\, \mathcal{D}(q_n)\big) \geq \rho_0, \text{ for any } 0 \leq n < N \text{ such that } q_n \in \Omega_h. \tag{182}$$

Proof. Since the matrix field $\mathcal{D}$ is pointwise positive definite and continuous, it is bounded over the compact set $\overline{\Omega}$, as well as its inverse and condition number, which fits the assumptions of Proposition 8.47 on the properties of Selling's decomposition. We can assume $q_* \in \Omega_h$, since otherwise the condition (182) is empty for $n = 0$, and the trivial chain of length $N = 1$ complies.

By Proposition 8.47 (spanning property), there exists a direct basis $e_1, \cdots, e_d$ of $\mathbb{Z}^d$ such that $\rho(e_i; \mathcal{D}(q_*)) \geq 2\rho_0$ for all $1 \leq i \leq d$, where $\rho_0 = \rho_0(\mathcal{D})$. By Proposition 8.47 (Lipschitz weights), the functions $q \in \overline{\Omega} \mapsto \rho(e_i; \mathcal{D}(q))$ are continuous, hence there exists $r_0 = r_0(\mathcal{D}) > 0$ such that:

$$\rho(e_i; \mathcal{D}(q)) \geq \rho_0, \qquad \text{for all } 1 \leq i \leq d, \text{ and all } q \in \overline{\Omega} \text{ s.t. } |q - q_*| \leq r_0. \tag{183}$$

By Proposition 8.47 (bounded offsets), one has $|e_i| \leq R_0$ for all $1 \leq i \leq d$, where $R_0 = R_0(\mathcal{D})$. Defining the matrix $G := [e_1, \cdots, e_d]$, and noting that $\det(G) = \det(e_1, \cdots, e_d) = 1$, we obtain that G^{-1} has integer coefficients bounded in absolute value by $R_1 = R_1(\mathcal{D})$. Denote $e := (q^* - q_*)/h$, recall that this vector or its opposite belongs to the canonical basis

of $\mathbb{R}^d$ by assumption, and let $(\lambda_1, \cdots, \lambda_d) = G^{-1}e$, in such way that $e = \lambda_1 e_1 + \cdots \lambda_d e_d$, $|\lambda_1|, \cdots, |\lambda_d| \leq R_1$ and $\lambda_1, \cdots, \lambda_d \in \mathbb{Z}$. Assuming w.l.o.g. that $\lambda_1, \cdots, \lambda_d \geq 0$ we define for all $n \leq N := \lambda_1 + \cdots + \lambda_d$:

$$q_n = q_* + h(\lambda_1 e_1 + \cdots + \lambda_r e_r + \lambda e_{r+1}), \qquad \text{where } n = \lambda_1 + \cdots + \lambda_r + \lambda,$$

with λ integer such that $1 \leq \lambda \leq \lambda_{r+1}$. Observe that $N \leq N_0$ where $N_0 = N_0(\mathcal{D}) := dR_1$, and that $|q_n - q_*| \leq hN_0R_0$ is smaller than r_0 provided $h \leq h_0$ where $h_0 = h_0(\mathcal{D}) = r_0/(N_0R_0)$, for any $0 \leq n \leq N$. This construction of $q_0, \cdots, q_N$ satisfies in view of (183) the announced properties, which concludes the proof. □

Proposition 8.36. *Let $\mathcal{D} \in C^0(\overline{\Omega}, S_d^{++})$, and let $u : h\mathbb{Z}^d \to [0, \infty[$ obey*

$$\mathfrak{F}^{\mathcal{D}(q)}u(q) \leq 1, \ \forall q \in \Omega_h, \qquad u(q) = 0, \ \forall q \in \partial\Omega_h. \tag{184}$$

Then $|u(q) - u(r)| \leq C|q - r|$ for all $q, r \in h\mathbb{Z}^d$, where $h > 0$ is small enough and $C = C(\mathcal{D})$.

Proof. It suffices to prove that $|u(q) - u(r)| \leq C|q - r|$ when $|q - r| = h$ are neighbors on the grid $h\mathbb{Z}^d$ (up to multiplying C by $\sqrt{d}$). It also suffices to prove the one sided inequality $u(q) \leq u(r) + C|q - r|$, by symmetry.

Assumption (184, left) at a point $q \in \Omega_h$ can be rewritten as

$$\sum_{e \in \mathcal{Z}} \rho\big(e; \mathcal{D}(q)\big) \max\{0, \, u(q) - u(q - he), \, u(q) - u(q + he)\}^2 \leq h^2,$$

in view of the Riemannian scheme definition (141). Therefore $u(q) \leq u(q + he) + h\rho(\varepsilon e; \mathcal{D}(q))^{-\frac{1}{2}}$ for any $e \in \mathbb{Z}^d$ and any sign $\varepsilon \in \{-1, 1\}$, with the convention $0^{-\frac{1}{2}} = \infty$. Let $q = q_0, \cdots, q_N = r$ be a chain as described in Lemma 8.35, joining the points of interest. Then

$$u(q_n) \leq u(q_{n+1}) + h\rho_0^{-\frac{1}{2}}, \tag{185}$$

for all $0 \leq n < N$. Indeed this follows from (182) and the previous estimate when $q_n \in \Omega_h$, and otherwise $u(q_n) = 0$ by the boundary condition satisfies the bound since u is non-negative. Accumulating these inequalities we obtain $u(q) \leq u(r) + hN_0\rho_0^{-\frac{1}{2}}$, as announced. □

Proof of (177) *in Theorem 8.33, using Proposition 8.36.* Assume that $\mathfrak{F}u(q) \leq 1$ for all $q \in \Omega_h$, and $u(q) = 0$ for all $u \in \partial\Omega_h$. Define $\mathcal{D}(q) := D(q, \alpha_*(q))/\mu(q, \alpha_*(q))$, and note that $\mathcal{D} \in C^0(\overline{\Omega}, S_d^{++})$. Since the TTI scheme (143) is defined as a maximum of Riemannian schemes parameterized by $\alpha \in [\alpha_*(q), \alpha^*(q)]$, we obtain that $1 \geq \mathfrak{F}u(q) \geq \mathfrak{F}^{\mathcal{D}(q)}u(q)$. Thus Proposition 8.36 applies and (177) is proved. □

8.4.2 Growth estimate in case (min)

The main result of this subsection is a growth estimate for sub-solutions to a minimum of discretized eikonal PDEs, established in Proposition 8.38. The growth estimate (178) is then deduced. The proof strategy differs from Section 8.4.1 since we cannot exploit

any local consistency between the active stencils of close discretization points. Instead, Lemma 8.37 below shows that we can use Selling offsets whose weights are positive to move in the general direction (186) of a given point, assumed to be far enough, see Figure 42 (right).

Lemma 8.37. *Given a compact set $\mathfrak{D} \subset S_d^{++}$, there exists $\rho_0 > 0$, $r_0 > 0$ and R_0 such that the following holds. For any $D \in \mathfrak{D}$ and any $v \in \mathbb{R}^d$ with $|v| \geq R_0$, there exists an offset $e \in \mathbb{Z}^d$ and a sign $\varepsilon \in \{-1, 1\}$ such that*

$$\rho(\varepsilon e; D) \geq \rho_0 \text{ and } |v - e| \leq |v| - r_0. \tag{186}$$

Proof. Since $\mathfrak{D}$ is compact, its elements are bounded, and likewise their inverses and condition numbers. Let $D \in \mathfrak{D}$ and $v \in \mathbb{R}^d$, then

$$\lambda_0 |v|^2 \leq \|v\|_D^2 = \sum_{e \in \mathbb{Z}^d} \rho(e; D) \langle v, e \rangle^2 \leq I \max_{e \in \mathbb{Z}^d} \rho(e; D) \langle v, e \rangle^2, \tag{187}$$

where we used successively (i) a lower bound $\lambda_0 = \lambda_0(\mathfrak{D})$ on the eigenvalues of $D \in \mathfrak{D}$, (ii) Selling's formula (198), and (iii) the fact that Selling's decomposition involves at most $I := d(d+1)/2$ positive weights. Thus there exists $e \in \mathbb{Z}^d$, a maximizer of (187, right), such that $\rho(e; D)\langle v, e \rangle^2 \geq |v|^2 \lambda_0 / I$. Observing that $|e| \leq R_1 = R_1(\mathcal{D})$ by Proposition 8.47 (bounded offsets), we obtain $\rho(e; D) \geq \rho_0$ where $\rho_0 = \rho_0(\mathfrak{D}) := \lambda_0 / (R_1^2 I)$.

On the other hand $\rho(e; D) \leq \rho_1$ where $\rho_1 = \rho_1(\mathfrak{D}) := \max\{\mathrm{Tr}(D); D \in \mathfrak{D}\}$, since $\rho(D) \leq \rho(D)\|e\|^2 \leq \sum_{e \in \mathbb{Z}^d} \rho(e; D)\|e\|^2 = \mathrm{Tr}(D)$. Therefore $|\langle v, e \rangle| \geq 2r_0 |v|$ with $r_0 = r_0(\mathfrak{D}) = \frac{1}{2}\sqrt{\lambda_0 / I \rho_1}$. Then, assuming w.l.o.g. that $\langle v, e \rangle \geq 0$, we obtain

$$\|v\| - \|v - e\| \geq \frac{\langle e, v \rangle}{|v|} - \frac{|e|^2}{2|v|} \geq 2r_0 - \frac{R_1^2}{2R_0},$$

using successively (i) [Mir19, Lemma 2.10], and (ii) the upper bounds on $\langle v, e \rangle$ and $|e|$, and the lower bound on $|v|$. Defining $R_0 = R_0(\mathfrak{D}) := R_1^2 / (2r_0)$ we conclude the proof. □

Proposition 8.38. *Let $\mathfrak{D} \subset S_d^{++}$ be a compact set, and let $u : h\mathbb{Z}^d \to [0, \infty[$ obey*

$$\min_{D \in \mathfrak{D}} \mathfrak{F}^D u(q) \leq 1, \forall q \in \Omega_h, \qquad u(q) = 0, \forall q \in \partial\Omega_h. \tag{188}$$

Then for any $q, r \in h\mathbb{Z}^d$, one has with $R = R(\mathfrak{D})$ and $C = C(\mathfrak{D})$

$$u(q) \;\leq\; \max\{u(r'); |r' - r| \leq Rh\} + C|q - r|. \tag{189}$$

Proof. We claim that the announced result holds with the constants $R = R_0$ and $C = r_0^{-1} \rho_0^{-\frac{1}{2}}$, where R_0, r_0 and ρ_0 are from Lemma 8.37. For that purpose, we fix the point $r \in h\mathbb{Z}^d$, and prove the announced result for all $q \in h\mathbb{Z}^d$, by induction on $|(q - r)/h|^2$ which is a positive integer. Note that if $q \notin \Omega_h$, then $u(q) = 0$ by the boundary condition, and (189) holds. Also (189) clearly holds if $|q - r| \leq Rh$.

The assumption (188, left) at $q \in \Omega_h$ such that $|q - r| \geq Rh$, can be rewritten as

$$\sum_{e \in \mathcal{Z}} \rho(e; D) \max\{0, u(q) - u(q - he), u(q) - u(q + he)\}^2 \leq h^2, \tag{190}$$

for some $D \in \mathfrak{D}$, in view of the Riemannian scheme definition (141). Denoting $v := (q-r)/h$, and noting that $|v| \geq R$, we find by Lemma 8.37 an offset $e \in \mathbb{Z}^d$ such that $q' := q - he$ satisfies,

$$u(q) \leq u(q') + h\rho_0^{-\frac{1}{2}}, \qquad\qquad |q'-r| \leq |q-r| - hr_0,$$

using (190) for the first estimate. The announced result follows by induction. □

Proof of the growth estimate (178). Define the set of positive definite matrices

$$\mathfrak{D} := \big\{ D(q,\alpha)/\mu(q,\alpha);\ q \in \overline{\Omega},\ \alpha \in [\alpha_*(q), \alpha^*(q)] \big\}, \tag{191}$$

which is compact since the functions $D, \mu, \alpha_*, \alpha^*$ are continuous over a compact domain. Then (176, left) implies (188, left), and the announced growth estimate (178) is established in (189). □

8.4.3 Proof of convergence

We follow a standard proof strategy [BR06] to establish the uniform convergence of the solutions to the proposed discretization scheme of the TTI eikonal PDE, henceforth denoted $\mathfrak{F}_h$ where $h > 0$ is the grid scale, towards the continuous solution as $h \to 0$. Note that alternative proof strategies exist which may yield stronger quantitative results, including convergence rates, see Remark 8.40.

As a first step, in Lemma 8.39, we establish the existence of a sub-solution and of a super-solution to $\mathfrak{F}_h$, which are bounded independently of h, as well as an approximation property of super-solutions by strict super-solutions.

Lemma 8.39. *The proposed discretization scheme $\mathfrak{F}_h$ of the TTI eikonal PDE* (143) *satisfies:*

- *(Explicit sub-solution) The null function $\overline{u} = 0$ satisfies $\mathfrak{F}_h \overline{u} = 0$ identically on $h\mathbb{Z}^d$.*
- *(Explicit super-solution) Let $a \in \mathbb{R}$, $b \in \mathbb{R}^d$ be such that $a + \langle b, q\rangle > 0$ for all $q \in \overline{\Omega}$, and $\mathcal{F}_q^*(b) \geq 1$ for all $q \in \overline{\Omega}$. Let $\underline{u}(q) := a + \langle b, q\rangle$ on Ω, and $\underline{u} = 0$ on $\mathbb{R}^d \setminus \Omega$. Then $\mathfrak{F}_h \underline{u} \geq 1$ on Ω_h, for any sufficiently small $h > 0$.*
- *(Approximation of super-solutions) One has $\mathfrak{F}_h[\lambda u(q)] = \lambda^2 \mathfrak{F}_h u(q)$ for any $u : h\mathbb{Z}^d \to \mathbb{R}$, $\lambda \geq 0$, and $q \in \Omega_h$. In particular, if u is a super-solution of $\mathfrak{F}_h$, then $(1+\varepsilon)u$ is a strict super-solution for any $\varepsilon > 0$, converging to u as $\varepsilon \to 0$.*

Proof. The 2-homogeneity property, announced in the last point, is obvious in view of the definition of the Riemannian (141) and TTI (143) schemes. The points (Explicit sub-solution) and (Perturbation of sub-solution) follow; for instance if $\mathfrak{F}_h u \geq 1$ then $\mathfrak{F}_h[(1+\varepsilon)u] = (1+\varepsilon)^2 \mathfrak{F}_h u > 1$ for any $\varepsilon > 0$.

The constants $a \in \mathbb{R}$, $b \in \mathbb{R}^d$, of the second point exist by compactness of $\overline{\Omega}$ and continuity and definiteness of the the norms $\mathcal{F}_q^*$, $q \in \overline{\Omega}$. Define $v(q) = a + \langle b, q\rangle$ on $\mathbb{R}^d$, and note that $v(q+he) \geq 0$ for any point $q \in \Omega$ and offset $\|e\| \leq R_0$, where R_0 is a bound on the scheme stencil radius, by continuity and provided the discretization scale h

is small enough. It follows under these conditions that $\underline{u}(q + he) \leq v(q + he)$. Then for any $q \in \Omega_h$

$$1 \leq \mathcal{F}_q^*(b) = \mathfrak{F}_h v(q) \leq \mathfrak{F}_h \underline{u}(q), \tag{192}$$

using successively (i) the assumption on b, (ii) the scheme consistency (144), and (iii) the DDE property of the scheme, see Definition 8.6, and the observation that $\underline{u}(q + he) \leq v(q + he)$ for all offsets e of the scheme stencil, whereas $\underline{u}(q) = v(q)$ since $q \in \Omega$. The result follows. □

By [Mir19, Theorem 2.3] there exists a unique solution $u_h : h\mathbb{Z}^d \to \mathbb{R}$ to the scheme $\mathfrak{F}_h$. For context, uniqueness is established using the *comparison principle*, whereas existence is proved by *Perron's method* (maximal sub-solution). Both of these classical techniques require a scheme obeying the DDE property, see Definition 8.6. Since in addition the scheme $\mathfrak{F}_h$ is causal, see again Definition 8.6, its solution may be computed using the single pass FMM on Ω_h, see Algorithm 5.

The scheme solution $u_h : \Omega_h \to \mathbb{R}$ is bounded above and below by any super- and sub-solution, hence choosing those of Lemma 8.39 we obtain the bounds $\overline{u} \leq u_h \leq \underline{u}$ on Ω_h which are independent of h. This allows us to consider the lower and upper limits $\overline{\mathbf{u}}, \underline{\mathbf{u}} : \overline{\Omega} \to \mathbb{R}$, defined for all $q \in \overline{\Omega}$ as

$$\overline{\mathbf{u}}(q) := \liminf_{h \to 0,\, q_h \to q} u_h(q_h), \qquad \underline{\mathbf{u}}(q) := \limsup_{h \to 0,\, q_h \to q} u_h(q_h), \tag{193}$$

where implicitly $q_h \in h\mathbb{Z}^d$. By construction $0 = \overline{u} \leq \overline{\mathbf{u}} \leq \underline{\mathbf{u}} \leq \underline{u}$ on $\overline{\Omega}$, and $\overline{\mathbf{u}}$ is lower semi-continuous whereas $\underline{\mathbf{u}}$ is upper semi-continuous. By the growth estimate (178) one has $0 \leq \overline{\mathbf{u}}(q) \leq \underline{\mathbf{u}}(q) \leq C \operatorname{dist}(q, \partial\Omega)$, hence $\overline{\mathbf{u}}$ and $\underline{\mathbf{u}}$ obey the null Dirichlet boundary condition on $\partial\Omega$. By the DDE property of the scheme $\mathfrak{F}_h$, and by consistency (144), passing to the limit one obtains that in the sense of viscosity solutions[12] [BCD08]

$$\mathcal{F}_q^*(\nabla \overline{\mathbf{u}}(q)) \geq 1, \qquad \mathcal{F}_q^*(\nabla \underline{\mathbf{u}}(q)) \leq 1, \tag{194}$$

for all $q \in \Omega$, with the notation $\mathcal{F}_q^*$ of (179). By the continuous comparison principle [BCD08, Theorem 5.9], we obtain that $\underline{\mathbf{u}} \leq \overline{\mathbf{u}}$. Hence $\mathbf{u} = \overline{\mathbf{u}} = \underline{\mathbf{u}}$ is a viscosity solution to (179), and one has $u_h \to \mathbf{u}$ uniformly as $h \to 0$ by (193). This concludes the proof of Theorem 8.33.

Remark 8.40. *A quantitative convergence rate $\|u_h - \mathbf{u}\|_{L^\infty(\Omega_h)} = \mathcal{O}(\sqrt{h})$ as $h \to 0$, improving on the uniform convergence result of Theorem 8.33, can likely be established by following the same scheme of proof as [Mir19, §2.1], and by assuming Lipschitz regularity for the TTI parameters σ and R. However this would introduce a number of technicalities, such as the doubling of variables argument [Eva10], that we have chosen to avoid here since they are not specifically related to the models of interest.*

[12] In the sense of viscosity solutions, (194) should be understood as follows: let $\varphi \in C^2(\Omega)$ be arbitrary. If $\overline{u} - \varphi$ attains its minimum at $q \in \Omega$, then $\mathcal{F}_q^*(\nabla\varphi(q)) \geq 1$. If $\varphi - \underline{u}$ attains its minimum at $q \in \Omega$, then $\mathcal{F}_q^*(\nabla\varphi(q)) \leq 1$.

8.5 Numerical experiments

In this section, we present numerical experiments on three-dimensional test cases so as to evaluate the cost and accuracy of our TTI eikonal solver.

First, we consider a TTI medium with a semi-analytical solution to determine the convergence order and computation time of our numerical scheme. In the numerical experiments, we compare the two versions of our scheme to solve the underlying 1D-optimization problem: quasi-convex optimization (with CPU implementation), and optimization by grid search (with GPU implementation), see Section 8.1.3. In the latter case, we also study the influence of the sampling rate, denoted K in (146), on the solution accuracy. We find that the GPU implementation is fifty times faster than the CPU implementation in this test-case.

We then consider two alternative eikonal solvers, able to handle speed propagation profiles either (i) *less* or (ii) *more* general than TTI anisotropy. (i) A standard isotropic fast marching solver [Set96], enhanced with source factorization and second order finite differences, addresses isotropic (spherical) speed profiles. (ii) The state of the art CPU eikonal solver [DCC$^+$21], referred to as the "general scheme", handles anisotropy associated with a full Hooke tensor, of which TTI anisotropy is a special case, see Section 8.A. The comparison is done on a medium with orthorhombic anisotropy, with an analytical solution to the eikonal equation, and its projections to the closest TTI medium and isotropic medium. This experiment allows to quantify and compare the *discretization error*, associated to the grid scale, with the *consistency error*, related to the approximation of the anisotropic speed propagation profile. We also compare the computation time and accuracy of the different schemes.

Last, we consider an application to a realistic synthetic test-case which comes from the homogenization of an isotropic medium. The resulting anisotropy is fully general, but is expected to be close to TTI anisotropy because of the sedimentary structure of the medium. We verify this assumption by considering a projection of the general medium to a TTI medium and to an isotropic medium, and compare the results of the general scheme, TTI scheme and isotropic scheme. We also compare the solution to the eikonal equation with the solution to the elastic wave equation, to verify that the solution to the eikonal equation indeed is consistent with the first-arrival traveltimes of the wave propagation.

All the computations presented here have been performed on the Univ. Grenoble Alpes HPC perform. For the numerical scheme in the CPU case as well as for the general scheme, the computation has been performed on one Intel node equipped with a Xeon Skylake Gold processor, with a core clocked 2.1 GHz and 192 GO of RAM. For the numerical scheme in the GPU case as well as for the isotropic scheme, the computation has been performed using an Nvidia Tesla V100 with 5120 CUDA cores and 96 GO of RAM.

The fast marching method only uses a single CPU core due to its intrinsic sequential nature. The massively parallel solver presently uses a single GPU card, but its multi-GPU extension is an opportunity for future work, possibly along the lines of [HJ16] in the isotropic setting.

8.5.1 Numerical application on a synthetic case obtained from the conformal transformation of a TTI metric

We consider a semi-analytical test case, so as to investigate the convergence rate and numerical error of our numerical scheme. It is a non-trivial heterogeneous TTI metric, obtained from a conformal diffeomorphic transformation of a homogeneous TI metric. By design, the exact solution to the eikonal equation is known and easily evaluated numerically to machine precision.

Conformal transformations are helpful to create non-trivial media with known solutions: indeed, the Jacobian of a conformal transformation $\phi : \Omega \subset \mathbb{R}^3 \to \mathbb{R}^3$ is a scaled rotation, namely $\operatorname{Jac}\phi(x) = \alpha(x)R(x)$ with scaling parameter $\alpha(x) > 0$ and rotation matrix $R(x)$. Our test medium is obtained as the pull-back of a homogeneous TI medium by ϕ, so that the associated TTI eikonal PDE (124) parameters take the form $(\frac{a}{\alpha(x)^4}, \frac{b}{\alpha(x)^4}, \frac{c}{\alpha(x)^4}, \frac{d}{\alpha(x)^2}, \frac{e}{\alpha(x)^2})$ and $R_0R(x)$ pointwise, where (a, b, c, d, e) and R_0 are fixed. One can likewise define the pull back by a conformal transformation of an eikonal equation which is isotropic, or whose anisotropy is defined by a general Hooke tensor [DCC+21, Appendix A].

Three dimensional conformal transformations include dilations, translations, rotations, the inversion $x \in \mathbb{R}^3 \setminus \{0\} \mapsto x/\|x\|^2$, and compositions of these. In our experiments we use a "special conformal transformation", defined by: $\phi(x) := \frac{x-b\|x\|^2}{1-2\langle b,x\rangle+\|b\|^2\|x\|^2}$. It is smooth except for a singularity at $b/\|b\|^2$, where $b \in \mathbb{R}^3$ is a parameter. It is obtained as the composition of an inversion, a translation by $-b$, and another inversion. We choose $b := (1/6, 1/9, 1/18)$ and let $\widetilde{\Omega} :=]-1, 1[^3$ with seed at the origin, so that the singular point $b/\|b\|^2 \notin \widetilde{\Omega}$, and the image domain $\Omega := \phi(\tilde{\Omega})$ is star shaped with respect to the origin, see Figure 43.

We consider the homogeneous TI metric from the mica medium [BC91], defined by:

$$\begin{pmatrix} 178 & 42.4 & 14.5 & 0 & 0 & 0 \\ 42.4 & 178 & 14.5 & 0 & 0 & 0 \\ 14.5 & 14.5 & 54.9 & 0 & 0 & 0 \\ 0 & 0 & 0 & 12.2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 12.2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 12.2 \end{pmatrix}, \ \rho = 2.79\text{g/cm}^3,$$

which we rotate by $3\pi/5$ with Euler axis $(2, 1, 3)$, in such way that the transverse isotropy plane is in a generic position rather than axis aligned (the latter may unfairly advantage eikonal solvers based on a Cartesian discretization grid such as ours).

The TTI metric from the mica corresponds to a maximization case, see (125). We also consider the same setting with different materials, such as the stishovite medium [BC91] which corresponds to a minimization case. Remarkably, in our numerical experiments, the two different cases yield completely similar computation time and error convergence, despite the difference in the formulation of the numerical scheme and in the mathematical proof of convergence Section 8.4. As the results are very close, we only illustrate the case of the mica medium. A cross-section of the solution to the eikonal equation is presented in Fig 44, which shows how the solution u relative to the constant metric on a transformed

domain (right figure) translates to a solution $\tilde{u}$ relative to a non-trivial metric on the regular cube domain (left figure) with the conformal transformation.

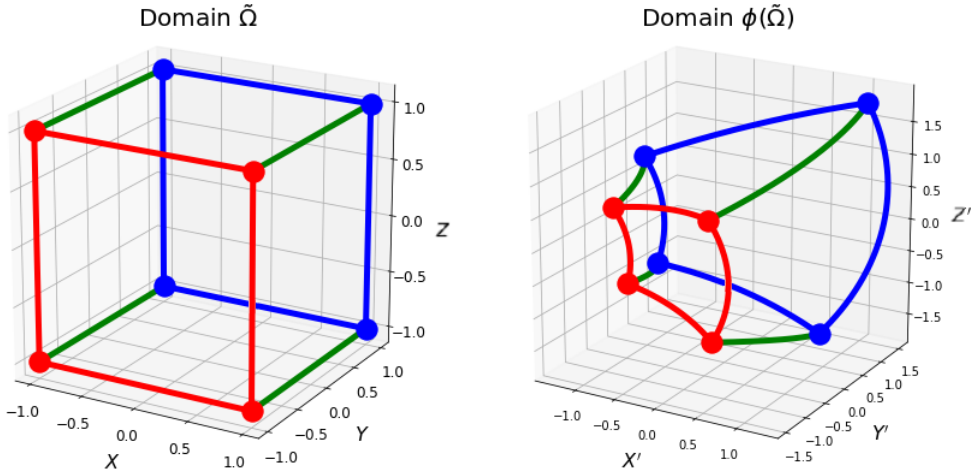

Figure 43: Edges of the domain $\tilde{\Omega} =]-1,1[^3$ (a cube) and of its image $\Omega = \phi(\tilde{\Omega})$ by a special conformal transformation.

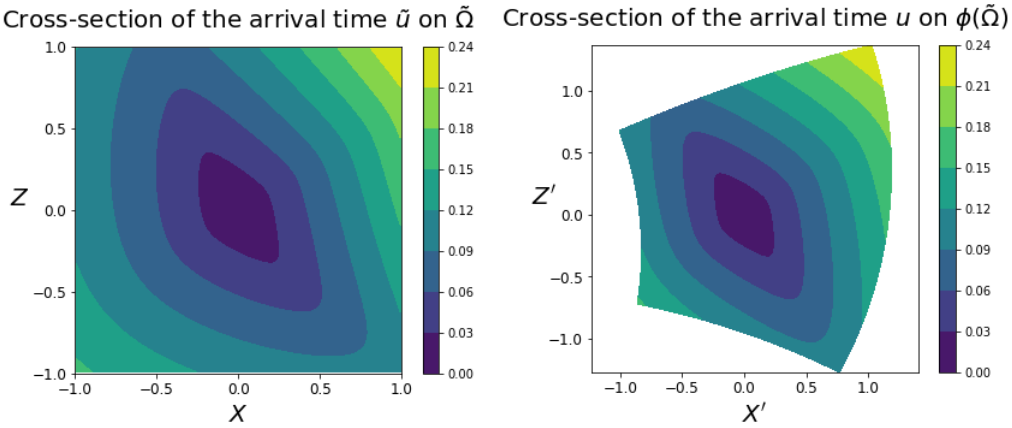

Figure 44: Cross-section at $Y = 0$ of the solution to the eikonal equation for a non-trivial TTI metric on $\tilde{\Omega}$ (left), which corresponds to a constant metric on the transformed domain $\Omega = \phi(\tilde{\Omega})$ (right).

In the CPU implementation of our TTI eikonal solver, the 1D optimization problem underlying our scheme is solved up to machine precision, by taking advantage of its quasi-convexity, see Theorem 8.11. As a result, the numerical accuracy is directly related to the scale h of the Cartesian discretization grid, and thus to the number of points $N \approx |\Omega|/h^3$. Second order $\mathcal{O}(h^2)$ convergence rates are observed on Fig. (45, right), as expected since we use second order finite differences, see Section 8.C.

In the GPU implementation, there is an additional source of approximation, related to the sampling parameter K in the optimization by grid search of the 1D optimization

problem underlying our scheme, see (146). As illustrated on Figure 36, this amounts to approximating the slowness surface of the pressure wave with the union or the intersection of $k := K + 1$ ellipses. When k is fixed, the numerical error of the scheme first decreases as the grid scale is refined, until a plateau is reached. This could be expected from the $\mathcal{O}(h^2+k^{-2})$ consistency error of the scheme, with second order finite differences as here, see Proposition 8.9. The plateau occurs for a number of ellipses k approximately proportional to the inverse grid scale h^{-1}. This scheme exhibits second-order convergence for small domain sizes, provided k is large enough, but the convergence rate then slightly degrades for the finest grid scales h; from the theoretical standpoint, convergence is guaranteed by Theorem 8.33, but not a specific rate.

In our numerical experiment, we consider from 10 to 26 ellipses. The size of the medium also goes from $39 \times 39 \times 39$ to $217 \times 217 \times 217$. This upper limit on the size of the domain comes from a memory limit on the GPU, rather than a limit on computation time. The computation time is quasi-linear w.r.t the total number of points for both the CPU and GPU eikonal solvers, and increases with the number k of ellipses in the latter case. The CPU implementation is about twice more accurate than the GPU implementation, for the examples considered in Figure 45, but comes at the cost of a computation time more than 10 times larger.

Remark 8.41 (Elliptic approximation)**.** *For computational efficiency purposes, on can be tempted to approximate the TTI eikonal equation with a Riemannian eikonal equation, and thus the algebraic P-wave slowness surface with a single ellipsoid. This corresponds to a special case of our scheme, with a single ellipse ($k = 1$), and thus a trivial grid search. For the test case considered here, the numerical error almost immediately reaches a plateau, even with the use of ten ellipses ($k = 10$) as considered in Figure 45: the scheme converges towards an erroneous solution, which shows the importance of properly taking into account a TTI anisotropy compared with elliptic anisotropy.*

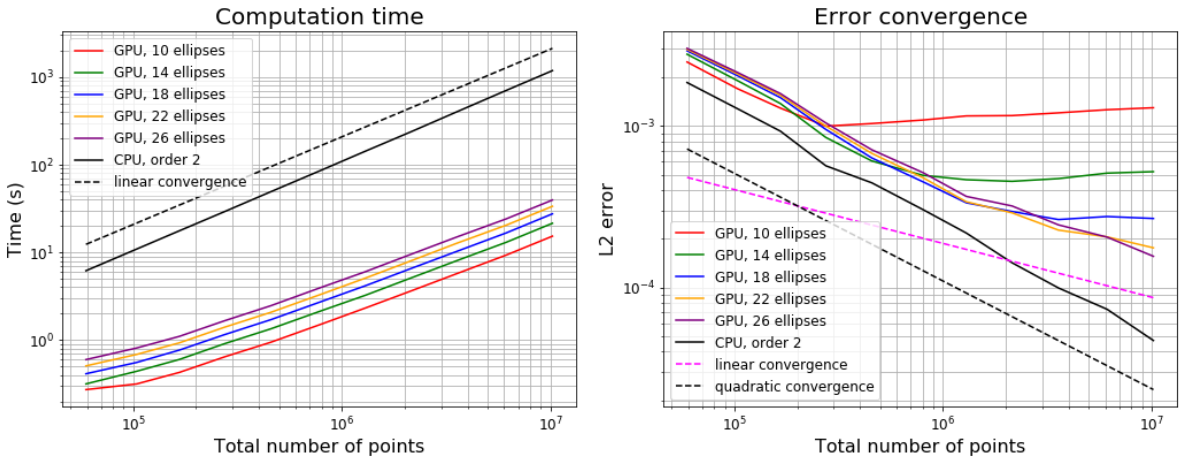

Figure 45: Computation time and error of the numerical scheme, depending on the number of ellipses for the GPU implementation. For visual interpretation, dotted lines illustrate linear convergence $\mathcal{O}(h)$ and quadratic convergence $\mathcal{O}(h^2)$.

8.5.2 Projection error for the orthorhombic anisotropy

Orthorhombic anisotropy is a more general type of anisotropy compared with the TTI anisotropy, and can be found in some crystalline structures. An orthorhombic medium does not exhibit the rotational symmetry that is found in TTI media, and corresponds to a Hooke tensor with nine independent elastic parameters, compared with five for the TTI anisotropy. The eikonal equation with orthorhombic anisotropy can be solved by the anisotropic variant of the Fast Marching method presented in [DCC+21], which we refer to as the "general scheme".

We consider the projection of an orthorhombic Hooke tensor to the closest Hooke tensor with TTI anisotropy. We have two goals in mind: first, we want to compare the computation time and the accuracy of our numerical scheme with the general scheme, and second, we want to quantify the projection error caused by the approximation in the anisotropy. Likewise, we consider the projection from the orthorhombic to an isotropic Hooke tensor, which we solve with an isotropic Fast Marching method implemented on GPU.

We consider the orthorhombic anisotropy defined from the olivine medium [BC91]:

$$\begin{pmatrix} 323.7 & 66.4 & 71.6 & 0 & 0 & 0 \\ 66.4 & 197.6 & 75.6 & 0 & 0 & 0 \\ 71.6 & 75.6 & 235.1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 64.6 & 0 & 0 \\ 0 & 0 & 0 & 0 & 78.7 & 0 \\ 0 & 0 & 0 & 0 & 0 & 79.0 \end{pmatrix}, \ \rho = 3.311\text{g/cm}^3,$$

Similarly to the previous subsection, we use a conformal transformation to create a non-trivial heterogeneous metric with a known solution. The projection of the Hooke tensor and the pull-back of the metric by the conformal transformation can be done in any order, so the TTI and isotropic settings also correspond to a conformal transformation of a homogeneous metric, with known solution.

We illustrate on Fig 46 the slowness surfaces related to the olivine (orthorhombic medium) and its TI and isotropic projections. Contrarily to TI and isotropic metrics, the orthorhombic anisotropy does not possess a rotational symmetry, and so we show two cross-sections of the corresponding slowness surfaces. For the eikonal equation, we are only interested in the inner surface, related to the fastest speed.

We show in Table 5 the results of the different numerical schemes on the different media. The domain size is $77 \times 77 \times 77$. For the GPU case, we consider 10 ellipses for the optimization by grid search.

We observe that the general scheme and the TTI scheme with CPU have a similar computation time, and that the TTI scheme with GPU is approximately fifty times faster. The isotropic scheme with GPU is also fifty times faster compared with the TTI scheme with GPU. The error due to the numerical scheme is much smaller than the error due to the approximation in the anisotropy of the medium: indeed, we observe that the numerical L^2-error and the exact L^2-error are almost identical. Besides, the L^2-error is more than two times bigger when comparing the TTI projection to the isotropic projection, going from an error of 2.27% to 5.31%.

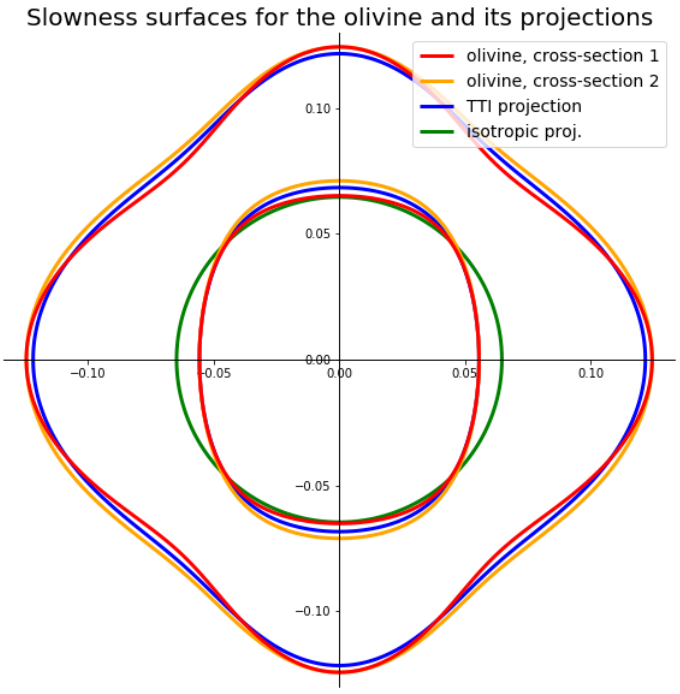

Figure 46: Slowness surfaces for the olivine and its TTI and isotropic projections. The TTI and isotropic projections have a rotational symmetry along the vertical axis, but not the olivine which is an orthorhombic medium, so we present two cross-sections of the slowness surface for the olivine along two vertical plans. Only the inner surfaces are of interest for the eikonal equation, related to the fastest speed.

With this example, we see that we need to make a choice between the accuracy of the numerical scheme, the accuracy of the anisotropy model, and the computation time required to solve the corresponding eikonal equation. It can be interesting to consider media with fully general Hooke tensors to better encompass the anisotropy of a geophysical medium, especially for orthorhombic media, as we see that the TTI projection leads to errors. However, for use-cases in seismic imaging which do not deviate too much from TTI anisotropy, such as the one presented in the next section, the TTI scheme is a better choice: the computation time is greatly improved as it is approximately fifty times faster than the general scheme, which can open the door to efficient applications in seismic imaging.

8.5.3 Anisotropic media coming from the homogenization of an isotropic medium

In the case of a medium with a sedimentary structure, the local invariance by rotation around the normal axis to the layers is expected, which leads to TTI anisotropy. Even if the materials of the medium are intrinsically isotropic, the medium can be represented with TTI anisotropy if the typical wavelength of seismic waves is larger than the typical size of the layers caused by the sedimentation, and this can be done through the homogenization process.

	Computation time (s)	L^2-error (numeric)	L^2-error (exact)
General scheme on orth. med.	22.6	0.000109	0
TTI scheme (CPU) on TTI med.	10.4	0.0227	0.0226
TTI scheme (GPU) on TTI med.	0.545	0.0227	0.0226
Isotr. scheme (GPU) on isotr. med.	0.0125	0.0531	0.0546

Table 5: Computation on a synthetic test-case of size $77 \times 77 \times 77$, with orthorhombic anisotropy and its projection to TTI anisotropy and isotropy. The L^2-error (num.) corresponds to the difference between the numerical solution on the corresponding medium and the exact solution on the orthorhombic medium. The L^2-error (exact) corresponds to the difference between the exact solution on the corresponding medium and the exact solution on the orthorhombic medium. All the L^2-errors are normalized by the L^2 norm of the exact solution on the orthorhombic medium. The general scheme is the numerical scheme from [DCC+21], used with second-order accuracy and "cut-cube" setting. The TTI scheme with GPU is used with 10 ellipses. The isotropic scheme is a Fast Marching scheme with second-order accuracy using GPU.

In this subsection, we study the application of our numerical scheme on a realistic dataset, which comes from the homogenization of an isotropic medium into a model with fully general anisotropy: the Hooke tensor has 21 independent elastic parameters. The isotropic model is the SEG/EAGE overthrust model, see [ABK97] and Figure 47, and information on the homogenized model can be found on [CMA+20]. We consider the projection of the general medium into a TTI medium, with the projection being made on the Hooke tensor at each point of the domain, see [CMA+20]. We then study the relevance of this TTI projection by comparing the solution in the medium with general anisotropy to the solution in the medium with TTI anisotropy, by using the scheme from [DCC+21] which can handle general anisotropy.

We also consider the projection of the homogenized medium to an isotropic medium, which is a way to study the strength of the anisotropy coming from the homogenization process in this medium. We use the Fast Marching isotropic scheme with GPU to solve the corresponding eikonal equation.

The medium is discretized on a $107 \times 534 \times 534$ grid, corresponding to a real medium of dimensions 4 km $\times$ 20 km $\times$ 20 km. The source point is placed on the point $(0, 267, 267)$, which corresponds to the middle of the medium on the surface. In order to use the multi-scale source factorization described in Section 8.C, we need a finer discretization near the source. For that purpose, we interpolate the value of the general metric near the source with a trilinear interpolation on each coefficients of the Hooke tensor, and then we use the projection to TTI metric again.

First, we use the scheme from [DCC+21], which can solve the eikonal equation with a Hooke tensor of general anisotropy, and use it with second-order precision. With this scheme, we get the solution to the eikonal equation on both the general metric and the TTI metric, and we can consider these two results as the closest we have to the exact solutions. We also compute the solution on the TTI metric with our numerical scheme, with both CPU and GPU implementations. For the GPU case, we use 10 ellipses for the

optimization by grid search, and checked that using a higher number of ellipses does not significantly change the result.

On Figure 47, we show a superposition of the solution to the eikonal equation with the solution to the wave equation in the same medium. The elastic wave propagation problem is solved using the spectral-element based modeling and inversion code SEM46 [TBM+19b, CBM20]. We observe that the isochrones computed from our eikonal solver properly follows the wavefront of the solution to the wave equation at the corresponding time, as expected.

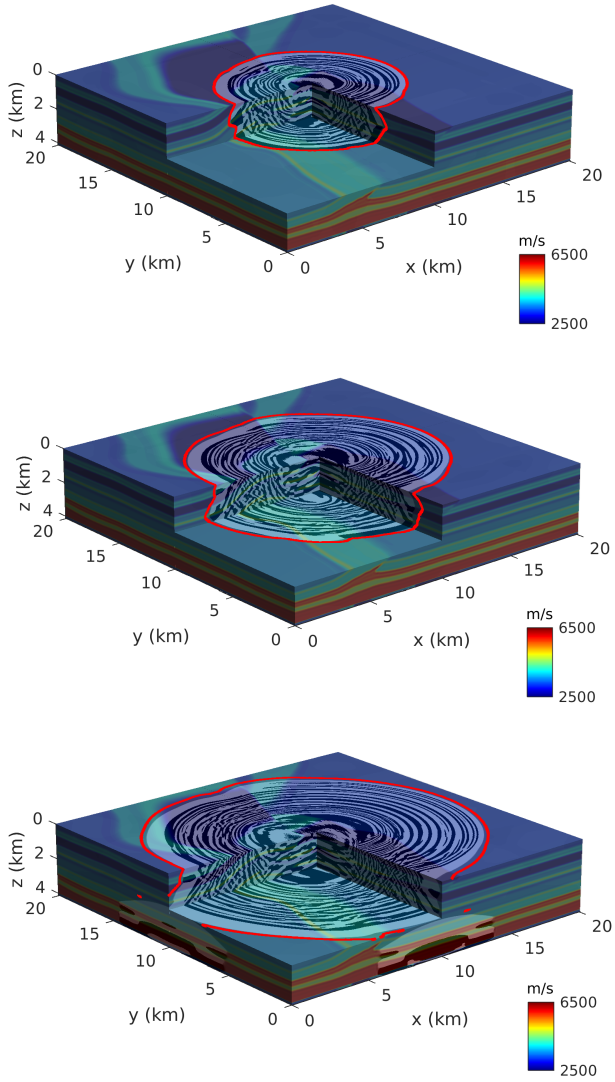

Figure 47: Elastic wavefield (black and white) computed in the 3D TTI medium coming from the homogenization of the SEG/EAGE overthrust model. The background corresponds to the P-wave velocity of this model. The red contour corresponds to the isochrone computing through our fast marching eikonal solver (with GPU implementation). The different snapshots are obtained at $t = 1.5$ s (top), $t = 2$ s (middle) and $t = 2.5$ s (bottom).

	Computation time (s)	L^2-error
General scheme on general medium	1571	-
General scheme on TTI medium	1569	0.00062
TTI scheme with CPU on TTI medium	478	0.0068
TTI scheme with GPU on TTI medium	13	0.0056
Isotr. scheme with GPU on isotr. med.	0.28	0.011

Table 6: Computation time on a realistic synthetic test-case. The L^2-error corresponds to the difference between the numerical solution and the solution computed by the general scheme on general medium, normalized by the L^2 norm of the solution computed by the general scheme on general medium. The general scheme is the scheme from [DCC+21], used with second-order accuracy. The TTI scheme with GPU is used with 10 ellipses. The isotropic scheme is a Fast Marching scheme with second-order accuracy using GPU.

The error due to the TTI projection can be estimated by the L^2-error between the general scheme on general medium and the general scheme on TTI medium as shown in Table 6, and is around 0.062%. For comparison, it is more than thirty times less than the projection error in the orthorhombic setting presented Section 8.5.2. The proposed TTI scheme, and the general scheme [DCC+21], rely on completely different discretization principles, and for this reason they produce slightly different numerical solutions even when applied to the *same* TTI eikonal PDE. We observe on Table 6 that the L^2 error between the TTI scheme and the general scheme on a general medium is around 0.68%, which is ten times larger than the error associated with the TTI projection alone, observed with the general scheme on the TTI medium. This validates the assumption that the anisotropy in the homogenized model is very close to TTI anisotropy, and that the associated projection error is well below the discretization error, related to the grid scale and scheme design. Besides, we observe that the computation time is greatly improved by the TTI scheme with GPU implementation compared with the general scheme, with a computation time a hundred times faster. On the other hand, the isotropic projection of the homogenized model leads to an error of around 1.1%, which is seventeen times higher than the projection error due to the TTI scheme, with a computation again fifty times faster.

As a conclusion, the anisotropy in this dataset is close to TTI anisotropy, as is expected from the homogenization process in an isotropic medium with sedimentary structure: the normal axis to the layers is a natural axis of symmetry. In the case of seismics faults or complex interactions between the layers, the anisotropy can become more complex and lose this symmetry axis, but in this realistic instance, the TTI anisotropy seems to be enough to explain the general anisotropy coming from the homogenization process. Contrast this with the Olivine medium considered Section 8.5.2, which is an orthorhombic crystal system with no such rotation invariance, and whose TTI projection error is significant. Therefore, the TTI scheme is adapted to efficiently compute the first arrival traveltimes in a realistic medium with sedimentary structure and no intrinsic anisotropy coming from inner crystal structure.

8.6 Conclusion

We presented a discretization for the eikonal equation with anisotropy coming from a TTI Hooke tensor. The scheme is monotone and causal, hence solvable in a single pass using the fast-marching method, but also has a simple Eulerian structure, hence fits massively parallel architectures as well; using classical enhancements such as source factorization, we achieve second order accuracy. Two implementations have been proposed, one for CPU and one for GPU. The GPU implementation features an additional parameter which must be correctly tuned, namely the number of ellipses whose envelope approximates the P-slowness surface, but performs much faster compared with the CPU implementation. The scheme is more than fifty times faster with a small loss in accuracy compared with the scheme from [DCC+21], which is a state-of-the-art scheme able to handle media with Hooke tensors of general anisotropy. In addition, the scheme from [DCC+21] suffers from a limit on the strength of anisotropy it can tackle (defined as the ratio between the highest velocity and the lowest velocity at a given position) even for TTI media, whereas our present scheme does not exhibit such a restriction on TTI media.

Future research will be devoted to applications to seismic imaging by tomographic inversion. Besides, an extension of the method to orthorhombic Hooke tensors seems possible, since those are TTI in every cross section. The generalization of our present scheme would involve a two dimensional - rather than one dimensional in the present TTI setting - minimization, maximization, or min-max saddle point optimization problem at each grid point.

8.A Thomsen parameters and Hooke tensor symmetry

In this section, we briefly describe how the TTI eikonal PDE (124) is related to classical descriptions of an elastic medium, based either on the Hooke elasticity tensor, or on the Thomsen parameters. A 3D geological medium is described by a fourth-order elasticity tensor, referred to as the Hooke tensor and denoted $\mathfrak{c} = (\mathfrak{c}_{ijkl})$, where $i, j, k, l \in \{1, 2, 3\}$, and by the density ρ of the medium. The Hooke tensor is subject to the symmetry relations $\mathfrak{c}_{ijkl} = \mathfrak{c}_{jikl} = \mathfrak{c}_{jkij}$, allowing it to be represented as a 6×6 matrix $\mathfrak{C}$ using Voigt's notation, see Section 8.2.1 and (148).

Some additional symmetries are often considered for a geological medium. A transversely isotropic medium is a geological medium whose local elasticity properties are invariant by rotation around a specific axis. It is called *vertically transversely isotropic* (VTI) in case of invariance around the vertical axis, and *tilted transversely isotropic* (TTI) otherwise. In the case of VTI symmetry, the Hooke tensor (in Voigt notation) only has 5 independent elastic parameters and can be written as [Tho86]:

$$\mathfrak{C}^{VTI} = \begin{pmatrix} c_{11} & c_{12} & c_{13} & 0 & 0 & 0 \\ c_{12} & c_{11} & c_{13} & 0 & 0 & 0 \\ c_{13} & c_{13} & c_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & c_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & c_{44} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{c_{11}-c_{12}}{2} \end{pmatrix}.$$

A Hooke tensor with TTI symmetry can be obtained from a Hooke tensor with VTI symmetry C^{VTI} and a rotation matrix R, through the usual change of variables formula

$$\mathfrak{c}^{TTI}_{i'j'k'l'} = \sum_{i,j,k,l\in\{1,2,3\}} \mathfrak{c}^{VTI}_{ijkl} R_{ii'} R_{jj'} R_{kk'} R_{ll'}. \tag{195}$$

Conversely, a (non-convex) projection procedure allows to reconstruct a VTI tensor and a rotation R from a given Hooke tensor [CMA+20], up to some accuracy loss if the latter only has approximate TTI symmetry.

We now present the eikonal equation related to a TTI Hooke tensor, starting from the Christoffel equation (obtained as a high-frequency approximation of the elastic wave equation) [Sla03]:

$$\det\Big[\sum_{j,l\in\{1,2,3\}} \mathfrak{c}_{ijkl} \frac{\partial u}{\partial x_j} \frac{\partial u}{\partial x_l} - \rho\delta_{ik} \Big] = 0. \tag{196}$$

Between brackets is a 3×3 matrix, of indices i, k, obtained as the difference of (i) a contraction of the Hooke tensor by the partial derivatives of the arrival time function u, and (ii) the identity matrix scaled by the density ρ of the medium. For simplicity we assume in the following that $\rho = 1$, up to considering the reduced tensor $\mathfrak{c}/\rho$.

Define the *slowness vector* $(p_x, p_y, p_z) := R\nabla u$, and let $p_r^2 := p_x^2 + p_y^2$. Then the Christoffel equation (196) for a TTI symmetry can be algebraically factored as follows:

$$\begin{aligned} 0 = &(\frac{c_{11} - c_{12}}{2} p_r^2 + c_{44} p_z^2 - 1) \times \\ &(c_{11}c_{44}p_r^4 + c_{33}c_{44}p_z^4 - (2c_{13}c_{44} + c_{13}^2 - c_{11}c_{33})p_r^2 p_z^2 - (c_{11} + c_{44})p_r^2 - (c_{33} + c_{44})p_z^2 + 1). \end{aligned}$$

The first factor of this equation characterizes the arrival time of the SH (horizontal shear wave) propagation. This factor defines a Riemannian eikonal equation, which can be solved numerically [Mir14a, Mir19], but is of not interest for the computation of the first travel time. The second factor corresponds to the coupling P-SV, between the qP (quasi-pure pressure wave) and the qSV (quasi-pure vertical shear wave), and is the factor we need to consider for the first-arrival time. The P-SV equation for a TTI symmetry is a non-Riemannian anisotropic eikonal equation of degree four, mathematically more complex than the SH equation, which is reproduced in (129) and studied in this paper. Interestingly, the parameter c_{12} only appears in the SH equation, and the four relevant parameters $(c_{11}, c_{13}, c_{33}, c_{44})$ for the P-SV equation can be organized in a 2-dimensional Hooke tensor (151).

The P-SV equation, henceforth referred to as the TTI eikonal equation, is summarized as

$$a p_r^4 + b p_z^4 + c p_r^2 p_z^2 + d p_r^2 + e p_z^2 = 1, \quad \text{where } (p_x, p_x, p_z) = R\nabla u \text{ and } p_r^2 := p_x^2 + p_y^2, \tag{197}$$

with coefficients (a, b, c, d, e) derived from the Hooke tensor as above. For the Hooke tensors considered in geophysics, the TTI equation has two distinct solutions. In the (p_x, p_y, p_z) coordinate system, these solutions are called *slowness surfaces*, and are invariant by the rotation R. The inner surface corresponds to the slowness of the P wave (that is, the inverse of its velocity), whereas the outer surface corresponds to the slowness of the S wave. They are illustrated on Figure 36 and Figure 37, and studied in detail in Section 8.2.

Remark 8.42. *Thomsen's elastic parameters* $(V_p, V_s, \epsilon, \delta)$ *define another approach to obtain the TTI eikonal equation* (197)*, with the conversion formula [Tho86]:*

$$V_p = \sqrt{\frac{c_{33}}{\rho}}, \qquad V_s = \sqrt{\frac{c_{44}}{\rho}}, \qquad \varepsilon = \frac{c_{11} - c_{33}}{2c_{33}}, \qquad \delta = \frac{(c_{13} + c_{44})^2 - (c_{33} - c_{44})^2}{2c_{33}(c_{33} - c_{44})}.$$

The Thomsen parameters have physical interpretations in a weakly anisotropic setting: in particular, V_p *approximates the speed of the P-wave, and* V_s *of the S-wave. Nevertheless this is only an approximation in a special asymptotic setting, and in general both the P and S slowness surfaces depend on the four Thomsen parameters. For this reason we do not use here the convention* $V_s = 0$*, which has sometimes been considered to simplify the PDE* (197) *when one is only interested in the first travel time computation, corresponding to the P-wave.*

Remark 8.43. *In this paper, and in our numerical method, we only require the matrix* R *to be invertible in the definition* (197) *of the eikonal equation. This may be surprising since the TTI formalism* (195) *and* (196)*, based on physical considerations, makes the stronger assumption that* R *is a rotation. Our motivation for allowing non-rotations is that the computational domain is often the image of the physical domain by a diffeomorphism, e.g. to take into account the topography of the surface. In that case the equivalent eikonal PDE in the computational domain involves a matrix* R *defined as the product of the original rotation* R_0*, associated with the TTI model, and of the Jacobian of the diffeomorphism, which is usually not a rotation.*

Finally, we want to create a criterion based on the coefficients of a TTI metric, to quantify its anellipticity. We suggest the criterion: $c_{anel} := \alpha^* - \alpha_*$, where $0 < \alpha_* \leq \alpha^* < 1$ are defined in Theorem 8.3. It characterizes the difference between the two most extreme ellipses when doing the enveloppe of the TTI metric, see Figure 48. In the case of an elliptic metric, we have: $c_{anel} = 0$.

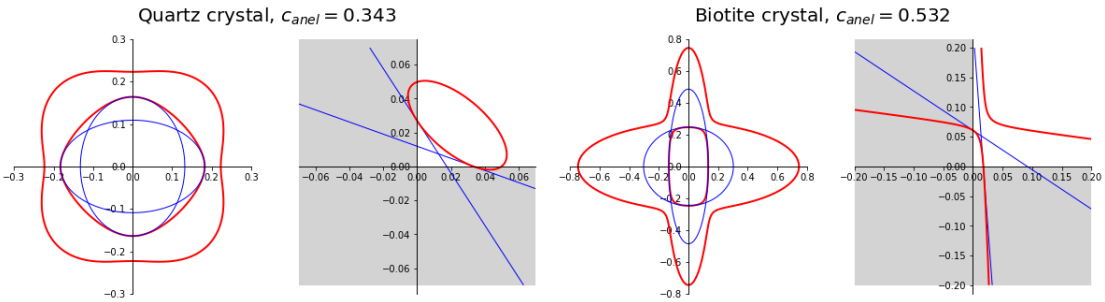

Figure 48: Example of two TTI media from the article [Tho86], represented as in 37, and with corresponding c_{anel}. In blue are shown the two most extreme ellipses in the optimization problem, which we use to define c_{anel}.

In order to evaluate the update operator Λ with the optimization by fine sampling, we consider a sampling of an interval over $K + 1$ elements, with K chosen by the user: the parameter K could reasonably be chosen depending on the criterion c_{anel}, as the accuracy of our approach should depend on how far the TTI metric is from an elliptic metric.

For illustrative purposes, we consider the article [Tho86], in which there are 58 examples of TTI metrics, corresponding to real and hypothetical materials, and we show the histogram of the corresponding c_{anel} in Figure 49. The two media with the highest c_{anel} correspond to crystallographic media, which are not usual in geophysics.

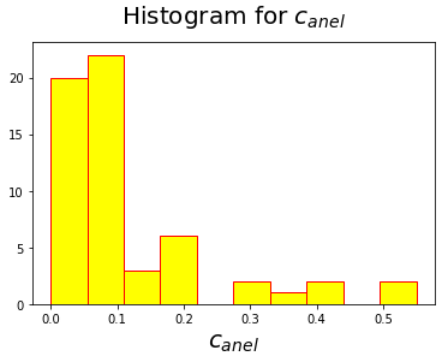

Figure 49: Histogram of c_{anel} for TTI examples from the article [Tho86].

From this study, we can reasonably assume that the usual TTI metrics do not exhibit a very strong anellipticity, apart from some crystallographic media. Therefore, for most real test-cases, the optimization by fine sampling should give a good result even if we use a small amount of ellipses in the sampling.

8.B Selling's decomposition

This section is devoted to Selling's decomposition of positive definite matrices, and to the proof of Proposition 8.7. Selling's decomposition is a tool originating from the field of lattice geometry [Sel74, CS92], which has recently found a number of applications in the design of anisotropic PDE discretizations, see [FM14, Mir19, BBM20, BM21] and Section 8.1.2. In the following, we present its definition, basic properties, and practical construction. For that purpose we introduce the concept of superbase, which is a slightly redundant coordinates system in $\mathbb{Z}^d$.

Definition 8.44. *A superbase of* $\mathbb{Z}^d$ *is a family* $(v_0, ..., v_d) \in (\mathbb{Z}^d)^{d+1}$ *such that* $v_0+...+v_d = 0$ *and* $|\det(v_1, ..., v_d)| = 1$.

Given a superbase $(v_0, \cdots, v_d)$ of $\mathbb{Z}^d$ and a matrix $D \in S_d^{++}$, we have Selling's formula

$$D = -\sum_{0 \le i < j \le d} \langle v_i, Dv_j \rangle e_{ij} e_{ij}^\top, \tag{198}$$

where the offsets are defined by the linear relations $\langle e_{ij}, v_k \rangle = \delta_{ij} - \delta_{ik}$ for all $0 \le i < j \le d$ and $0 \le k \le d$. This decomposition involves $I = \binom{d+1}{2} = d(d+1)/2$ terms, and the offsets have integer coordinates by construction. They admit simple expressions in small dimension: if $d = 2$ and $\{i, j, k\} = \{0, 1, 2\}$ then $v_{ij} = \pm e_k^\perp$, and if $d = 3$ and $\{i, j, k, l\} = \{0, 1, 2, 3\}$ then $v_{ij} = \pm v_k \times v_l$ (where $\times$ denotes the cross product). For a proof of Selling's formula (198), see [BBM21a, Lemma B.2].

Selling's decomposition of $D \in S_d^{++}$, $d \in \{2,3\}$, is defined as Selling's formula (198) applied to a D-obtuse superbase, defined below, in such way that the weights $-\langle v_i, Dv_j\rangle$ are non-negative.

Definition 8.45. *A superbase $b = (v_0, \cdots, v_d)$ is said D-obtuse, where $D \in S_d^{++}$, if $\langle v_i, Dv_j\rangle \leq 0$ for all $0 \leq i < j \leq d$. We let $S_b := \{D \in S_d^{++};\ b \text{ is } D\text{-obtuse}\}$.*

Using Pauli matrices in dimension $d = 2$, one obtains a linear parametrization

$$D(x,y) = \begin{pmatrix} 1+x & y \\ y & 1-x \end{pmatrix}, \qquad x^2 + y^2 < 1, \tag{199}$$

of the set of symmetric positive definite matrices of trace two, by the Euclidean unit ball. Figure 38 (left) illustrates the anisotropy defined by $D(a,b)$. The domains S_b associated to superbases b of $\mathbb{Z}^2$ appear as triangles in this parametrization, and together they define an infinite triangulation of the open unit ball $\{x^2 + y^2 < 1\}$, see Figure 38 (center).

In order to conclude the proof of the first part of Proposition 8.7, one needs to show that a D-obtuse superbase exists, which is the purpose of Selling's algorithm.

Proposition 8.46 (Selling's algorithm, [Sel74] or [BBM21a, Proposition B.3])**.** *Let $b = (v_0, \cdots, v_d)$ be a superbase of $\mathbb{Z}^d$, where $d \in \{2,3\}$, and let $D \in S_d^{++}$. If b is not D-obtuse, permute it so that $\langle v_0, Dv_1\rangle > 0$ and update it as follows*

$$b \leftarrow (-v_0, v_1, v_0 - v_1) \text{ if } d = 2, \qquad b \leftarrow (-v_0, v_1, v_2 + v_0, v_3 + v_0) \text{ if } d = 3. \tag{200}$$

Repeating this operation yields a D-obtuse superbase in finitely many steps.

In order to establish the second part of Proposition 8.7, we normalize Selling's decomposition as in [BBM21a, Appendix A], up to replacing some offsets with their opposites:

$$D = \sum_{e \in \mathcal{Z}^d} \rho(e;D)\, ee^\top, \qquad \text{where } \mathcal{Z}^d := \{e \in \mathbb{Z}^d;\ e \succ_{\text{lex}} 0\}, \tag{201}$$

where $\succ_{\text{lex}}$ stands for the lexicographic ordering. (Note that exactly one of $e \succ_{\text{lex}} 0$ or $-e \succ_{\text{lex}} 0$ holds for each $e \in \mathbb{Z}^d \setminus \{0\}$, and that $ee^\top = (-e)(-e)^\top$.) The weights $[\rho(e;D)]_{e \in \mathcal{Z}^d}$ are known as Selling parameters [CS92], and depend on D but *not* on the choice of D-obtuse superbase, see e.g. [BBM21b, Remark 2.13] for a proof. In view of Selling's formula (198), there exists at most $d(d+1)/2$ offsets $e \in \mathcal{Z}^d$ such that $\rho(e;D) \neq 0$, for any given $D \in S_d^{++}$. With these notations, we summarize in the next result some properties of Selling's decomposition: the offsets are bounded, the weights are locally Lipschitz, and a basis of $\mathbb{Z}^d$ can be extracted. We denote by $\mu(D) := \sqrt{\|D\|\|D^{-1}\|}$ the square root of the condition number of a matrix $D \in S_d^{++}$.

Proposition 8.47 (Propositions B.4, B.5 and B.8 in [BBM21a])**.** *The following holds in dimension $d \in \{2,3\}$, denoting $C = 2$ if $d = 2$ (resp. $C = 2\sqrt{3}$ if $d = 3$), and for some absolute constant $c > 0$:*

- *(Offset boundedness) For any $e \in \mathcal{Z}^d$, $D \in S_d^{++}$ s.t. $\rho(e;D) \neq 0$, one has $\|e\| \leq 2C\mu(D)$.*

- *(Lipschitz weights) For any $e \in \mathcal{Z}^d$, the mapping $D \in S_d^{++} \mapsto \rho(e;D)$ is locally Lipschitz with constant $C^2\mu(D)^2$.*
- *(Spanning property) For any $D \in S_d^{++}$, there exists $e_1, \cdots, e_d \in \mathcal{Z}^d$ such that*

$$\det(e_1, \cdots, e_d) = 1, \qquad \min_{1 \le i \le d} \rho(e_i; D) \ge c\|D^{-1}\|^{-1}.$$

For each superbase $b = (v_0, \cdots, v_d)$ of $\mathbb{Z}^d$, and each offset $e \in \mathcal{Z}$, the mapping $D \in S_b \mapsto \rho(e;D)$ is *linear*, where S_b is defined in Definition 8.45. Indeed, in view of Selling's formula (198), either $e = \pm e_{ij}$ for some $0 \le i < j \le d$ and thus $\rho(e;D) = -\langle v_i, Dv_j\rangle$ is a linear function of D, or $\rho(e;D) = 0$ identically on S_b. This linearity property, already used in the design of the PDE schemes [BBM20, BM21], allows here to conclude the proof of Proposition 8.7.

Proof of the second part of Proposition 8.7. For concreteness and w.l.o.g. we can assume that $\alpha_* = 0$, $\alpha^* = 1$, and thus $D(\alpha) := (1-\alpha)D_0 + \alpha D_1$, $\alpha \in [0,1]$, for some given $D_0, D_1 \in S_d^{++}$. Note that $\mu(D(\alpha)) \le \max\{\mu(D_0), \mu(D_1)\}$ for all $\alpha \in [0,1]$.

Denote by B be the collection of all superbases of $\mathbb{Z}^d$ whose elements are bounded by $2C\max\{\mu(D_0), \mu(D_1)\}$, where C is from Proposition 8.47 (offset boundedness). Then any $D(\alpha)$-obtuse superbase belongs to B, for any $\alpha \in [0,1]$.

Given a superbase $b = (v_0, \cdots, v_d) \in B$ the set $I_b = \{\alpha \in [0,1];\ b \text{ is } D(\alpha)\text{-obtuse}\}$ is defined by linear inequalities: $\langle v_i, [(1-\alpha)D_0 + \alpha D_1]v_j\rangle \le 0$ for all $0 \le i < j \le d$. Therefore I_b is closed and convex, hence either $I_b = \emptyset$, or $I_b = [\alpha_b^-, \alpha_b^+]$ is a segment with $0 \le \alpha_b^- \le \alpha_b^+ \le 1$. The weights of Selling's decomposition (198) are affine functions of the parameter $\alpha \in I_b$, with the general form $\alpha \mapsto -\langle v_i, [(1-\alpha)D_0 + \alpha D_1]v_j\rangle$, whereas the offsets e_{ij} are constant over I_b, as announced. Noting that $[0,1] = \cup_{b\in B} I_b$ is a finite union of such segments, we establish that Selling's decomposition is piecewise affine (140) which concludes the proof of Proposition 8.7. □

For concreteness and implementation purposes, we present in Algorithm 6 (without proof) a variant of Selling's algorithm (Proposition 8.46), which can be regarded as a constructive implementation of the above proof of the piecewise affine nature of Selling's decomposition. For notational simplicity and w.l.o.g. we assume again that $D(\alpha) := (1-\alpha)D_0 + \alpha D_1$ is parametrized over the interval $[\alpha_*, \alpha^*] = [0,1]$. This algorithm produces some breakpoints $0 = \alpha_0 < \cdots \le \alpha_K = 1$, and corresponding superbases $b_0, \cdots, b_{K-1}$ of $\mathbb{Z}^d$, such that b_k is $D(\alpha)$-obtuse for all $\alpha \in [\alpha_k, \alpha_{k+1}]$, $0 \le k < K$. Thus Selling's decomposition (198) of $D(\alpha)$ is affine on each of these intervals, as required.

8.C Scheme enhancements for higher accuracy

We describe in this subsection some algorithmic enhancements to the finite differences discretization (125) of the TTI eikonal PDE, meant to improve its accuracy, and we discuss of their relevance and applicability to the CPU and GPU implementations. The improvements are validated by a consistency analysis and by numerical experiments in Section 8.5.1, but not by a formal convergence analysis. The proposed scheme variants are adapted from the literature, hence are not original in themselves: various approaches

Algorithm 6 A modification of Selling's algorithm, producing $0 = \alpha_0 \leq \cdots \leq \alpha_K = 1$ and $b^0, \cdots, b^{K-1}$ superbases, such that b_k is $[(1-\alpha)D_0 + \alpha D_1]$-obtuse $\forall \alpha \in [\alpha_k, \alpha_{k+1}]$, $0 \leq k < K$.

Input : $D_0, D_1 \in S_d^{++}$.
Initialization : set $\alpha_0 = 0$, $k = 0$, and compute a D_0-obtuse superbase b^0using Proposition 8.46.
Repeat :
Denote $b^k = (v_0, \cdots, v_d)$, and let $\alpha_{k+1} \in [\alpha_k, \infty]$ be the smallest α such that
$\exists 0 \leq i < j \leq d$, $\langle v_i, (D_1 - D_0)v_j \rangle > 0$ and $\langle v_i, [(1-\alpha)D_0 + \alpha D_1]v_j \rangle = 0$.
(Up to permuting b^k, we assume that $i = 0$ and $j = 1$ are the active indices.)
If $\alpha_{k+1} \geq 1$ **then** set $K := k + 1$ and **exit**.
Define b^{k+1} by applying Selling's update (200) to b^k.
Set $k \leftarrow k + 1$ and proceed to the next iteration.

to source factorization are presented in [LQ12], second order accurate fast marching is introduced in [Set99], and the use of several discretization scales and coordinate systems depending on the distance from the source is documented in [WFNBZ20].

The discussion applies to any finite differences scheme of the following form, including the discretizations of the Riemannian (141) and TTI (143) eikonal PDEs:

$$\mathfrak{F}u(q) = \hat{\mathfrak{F}}(q, [\delta_h^e u(q)]_{e \in E}), \qquad \delta_h^e u(q) := \frac{u(q) - u(q - he)}{h}, \tag{202}$$

where the scheme unknown u is defined over a subset of the Cartesian grid $h\mathbb{Z}^d \ni q$ of scale $h > 0$, and where $E \subset \mathbb{Z}^d$ denotes the scheme stencil. We assume that the scheme $\mathfrak{F}$ obeys (i) the DDE (discrete degenerate ellipticity) property and (ii) the causality property, see Definition 8.6. In other words $\mathfrak{F}u(q)$ is (i) a non-decreasing function of the finite difference $\delta_h^e u(q)$, and (ii) only depends on the positive part $\max\{0, \delta_h^e u(q)\}$, for any $e \in E$. Since some considered scheme modifications may unfortunately break these properties, we discuss beforehand the extent to which our numerical eikonal solvers need them.

- The fast marching method, our CPU eikonal solver, requires a causal and DDE scheme. However a higher order non-DDE scheme can be used in the optional post-processing step of Algorithm 5, line 3., so as to improve the accuracy of the accepted value $u(q)$ before it is frozen. Following the principles of the high accuracy fast marching method (HAFMM) [Set99], this modification is only applied if it is small enough. Such post-processing is guaranteed not to degrade the convergence order of the method, by the comparison principle see [DCC+21, Proposition D.5], and in practice appears to improve it.

- The iterative GPU eikonal solver, only requires a DDE scheme. On the positive side, the causality property is not needed. On the negative side there is no clear opportunity to introduce a higher order non-DDE scheme, without the risk to create instabilities and to compromise the convergence of the solver. A basic fix to restore

monotony along the solver iterations, without guarantee but with often good empirical results see Section 8.5.1, is to accept an update value only if it is smaller than the previous one.

The consistency of a finite differences scheme (202) with the corresponding eikonal PDE, such as (125) with the TTI equation (128), is ultimately based on the finite differences approximation:

$$\left|\delta_h^e u(q) - \langle \nabla u(q), e\rangle\right| \leq \tfrac{1}{2}\|\nabla^2 u(r)\| h, \tag{203}$$

for sufficiently smooth u and where $r \in [q, q+he]$. This follows from a Taylor expansion, at a point $q \in \Omega$, in a direction $e \in \mathbb{R}^d$ of differentiation with $\|e\| = \mathcal{O}(1)$, and with finite difference scale $h > 0$. Each of the scheme variants discussed below is based on introducing in (202, left) a modified finite difference operator $\tilde{\delta}_h^e$ whose consistency with $\langle \nabla u(q), e\rangle$ is improved. These scheme variants are easily combined, and all three together are required to achieve second order convergence rates in Section 8.5.1.

Source factorization. The solution to the eikonal equation (128) has a singularity at the source point q_0, which degrades the accuracy of the finite difference approximation (203) since $\|\nabla^2 u(q)\| = \mathcal{O}(1/\|q-q_0\|)$ explodes as $q \to q_0$. Source factorization techniques [LQ12] rely on the computation of an equivalent u_* of the solution near the singularity, which can be easily evaluated and differentiated to machine precision. Typically one uses $u_*(q) = \mathcal{F}_{q_0}(q - q_0)$, which is the exact solution in the case of a constant metric, see Remark 8.5. As a result $\|\nabla^2 u(q) - \nabla^2 u_*(q)\| = \mathcal{O}(1)$ as $q \to q_0$, for a metric of suitable regularity, leading to a corresponding improvement in the scheme consistency (203). Following the principle of additive source factorization [LQ12], we introduce the modified finite differences operator

$$\tilde{\delta}_h^e u(q) := \frac{u(q) - u(q-he)}{h} + \omega_h^e, \quad \text{where } \omega_h^e := \Big(\langle \nabla u_*(q), e\rangle - \frac{u_*(q) - u_*(q-he)}{h}\Big). \tag{204}$$

The additional corrective term $\omega_h^e = \mathcal{O}(h^2/\|q-q_0\|)$ preserves the DDE property, since $\tilde{\delta}_h^e u(q) = \delta_h^e u(q) + \omega_h^e$ is a non-decreasing function of $\delta_h^e u(q)$, but breaks the causality property, since $\max\{0, \tilde{\delta}_h^e u(q)\} = \max\{0, \delta_h^e u(q) + \omega_h^e\}$ is not a function of $\max\{0, \delta_h^e u(q)\}$ when $\omega_h^e > 0$. As a result, this modification fits well in the iterative GPU solver, but introduces slight errors in the FMM CPU solver. (Note that there exists a stronger quantitative variant of the causality property, referred to as δ_1-causality where $\delta_1 > 0$, which is preserved under source factorization, see [DCC+21, Proposition D.4]. However the scheme proposed in this paper is not δ_1-causal.)

Multiscale computation. This technique features a preliminary run of the eikonal solver in a neighborhood $\Omega^1 \subset \Omega$ of the source point q_0 [WFNBZ20], using a smaller grid size $h_1 = h/k_1$ where $k_1 \geq 2$ is an integer. In essence, the pre-computation uses the modified finite difference operator

$$\tilde{\delta}_h^e := \frac{u(q) - u(q-h_1 e)}{h_1} = \frac{u(q) - u(q-(h/k_1)e)}{h/k_1},$$

which is more accurate by virtue of the smaller grid scale, reducing the consistency error (203) by an approximate factor k_1. The dimensions of Ω^1 are chosen (at least) k_1 times smaller than Ω, in such way that the refined computational domain around the source $\Omega_h^1 = \Omega^1 \cap h_1\mathbb{Z}^d$ has comparably many points as the global computational domain $\Omega_h = \Omega \cap h\mathbb{Z}^d$. In principle, this approach can be implemented recursively using an increasing sequence of subdomains $q_0 \in \Omega^N \subset \cdots \subset \Omega^1 \subset \Omega$ and of grid scales $h_N| \cdots |h_1|h$, but a basic two scales approach with $h_1 = h/4$ was found to be appropriate in our experiments, see Section 8.5.1.

Second order finite differences. The modified finite difference operator

$$\tilde{\delta}_h^e u(x) = \frac{u(q) - u(q - he)}{h} + \frac{u(q) - 2u(q - he) + u(q - 2he)}{2h},$$

is upwind, second order accurate, and is often used in the post-processing step of fast marching methods [Set99, Mir19], see Algorithm 5, line 3. The consistency error $\mathcal{O}(\|\nabla^3 u(q)\|h^2)$ is more favorable than (203) in regions where the solution is smooth. This variant is therefore mostly useful far from the source point, in contrast with the previous two modifications. The DDE property fails however, because $\tilde{\delta}_h^e u(q) = 2\delta_h^e u(q) - \frac{1}{2}\delta_h^{2e} u(q)$ is *not* a non-decreasing function of $\delta_h^e u(q)$ and $\delta_h^{2e} u(q)$, and for safety the second order correction is thus rejected if it is too large.

9 Massively parallel computation of globally optimal shortest paths with curvature penalization [MGB+21]

This section corresponds to the paper, currently submitted and under reviewing:

- Jean-Marie Mirebeau, Lionel Gayraud, Rémi Barrère, Da Chen, and Francois Desquilbet. Massively parallel computation of globally optimal shortest paths with curvature penalization. 2021

Abstract

We address the computation of paths globally minimizing an energy involving their curvature, with given endpoints and tangents at these endpoints, according to models known as the Reeds-Shepp car (reversible or forward only), the Euler-Mumford elasticae, and the Dubins car. For that purpose, we numerically solve degenerate variants of the eikonal equation, on a three dimensional domain, in a massively manner on a graphical processing unit. Due to the high anisotropy and non-linearity of the addressed PDE, the discretization stencil is rather wide, has numerous elements, and is costly to generate, which leads to subtle compromises between computational cost, memory usage, and cache coherency. Accelerations by a factor 30 to 120 are obtained w.r.t a sequential implementation. The efficiency and robustness of the method is illustrated in various contexts, ranging from motion planning to vessel segmentation and radar configuration.

9.1 Introduction

The eikonal Partial Differential Equation (PDE) characterizes the minimal travel time of an omni-directional vehicle, from a fixed source point to an arbitrary target point, and allows to backtrack the corresponding globally optimal shortest path. The numerical solution of the eikonal PDE is at the foundation of numerous applications ranging from path planning to image processing or seismic tomography [Set99]. Real vehicles however are usually not omni-directional, but are subject to maneuverability constraints: cars cannot perform side motions, planes cannot stop, etc. In this paper we focus on the Reeds-Shepp, Euler-Mumford and Dubins vehicle models, which account for these constraints by increasing the cost of highly curved path sections, or even forbidding them. The variants of the eikonal PDE corresponding to these models are *non-holonomic* (a degenerate form of anisotropy) and are posed on the three dimensional state space $\mathbb{R}^2 \times \mathbb{S}^1$, which makes their numerical solution challenging. A dedicated variant of the fast marching method is presented in [Mir18, MP19], and together with earlier prototypes it has found applications in medical image segmentation [CMC16, CMC17, DMMP18] as well as the configuration of surveillance systems [MD17, DDBM19]. However, a weakness of the fast marching algorithm is its sequential nature: the points of the discretized domain are *accepted* one by one in a specific order, namely by ascending values of the front arrival times, which imposes the use of a single CPU thread managing a priority queue.

In this paper, we present a massively parallel solver of the non-holonomic eikonal PDEs associated with the Reeds-Shepp, Euler-Mumford and Dubins models of curvature penalized shortest paths. We use the same finite difference discretization as [Mir18, MP19], on

a Cartesian discretization grid, but solve the resulting coupled system of equations using an iterative method implemented on a massively parallel computational architecture, namely a Graphics Processing Unit (GPU), following [WDB+08, JW08, FKW13, GHZ18]. Our numerical schemes involve finite difference offsets which are often numerous (30 for Euler-Mumford), rather wide (up to 7 pixels), and whose construction requires non-trivial techniques from lattice geometry [Mir18]. This is in sharp contrast with the standard isotropic eikonal equation addressed by existing GPU solvers, which only requires few and small finite difference offsets when it is discretized on Cartesian grids [WDB+08, JW08], and depends on unrelated geometric data when the domain is an unstructured mesh [FKW13, GHZ18]. Due to these differences, the compromises needed to achieve optimal efficiency - a delicate balance between the cost of computations and of memory accesses - strongly differ between previous works and ours, and even between the different models considered in this paper.

Our study provides the opportunity to inspect these compromises as the stencil of the finite difference scheme grows in width, number of elements and complexity, from 2 offsets of width 1 pixel (isotropic model in 2D), to 30 offsets of width up to 7 pixels (elastica model). Specifically, our observations regarding the models with *wider* finite difference stencils are the following: (i) They work best, somewhat counter intuitively, with a more finely grained parallelization, in our case obtained with smaller tiles and a smaller number of fixed point iterations within them, see Section 9.2.1 and Table 7. (ii) Precomputing and storing the stencil weights and offsets offers a significant speedup, up to 40% in our case, but the memory cost is prohibitive unless one can take advantage of symmetries in the equation to share this data between grid points, see Section 9.2.2. (iii) The scheme update operation involves a sort of the solution values fetched at the neighbors defined by the stencil, whose cost becomes dominant in the wide stencil case unless implemented in a GPU friendly manner, see Section 9.2.3. We expect our findings to transfer to other *wide stencil finite difference methods*, a class of numerical schemes commonly used to address Hamilton-Jacobi-Bellman PDEs arising in various applications, including deterministic (as here) and stochastic optimal control, optimal transport and optics design via the Monge-Ampere equation [Obe08], etc. Eventually, our GPU accelerated eikonal solver is 30× to 120× faster than the CPU fast marching method from [Mir18], see Table 8. In the numerical experiments Section 9.3, which include applications to medical image segmentation, boat routing and radar configuration, computations times on typical problem instances are often reduced from 30 seconds to less than one, enabling convenient user interaction.

Outline. We describe Section 9.1.1 the curvature penalized path optimization problems addressed, and Section 9.1.2 the eikonal equation formalism and the corresponding finite difference scheme. Our numerical solver is presented Section 9.2, distinguishing routines acting at the grid scale Section 9.2.1, the tile scale Section 9.2.2, and the pixel scale Section 9.2.3, see also Algorithms 7 to 9. Numerical experiments Section 9.3 illustrate the method's efficiency in various applications, corresponding to the best case scenario Section 9.3.1 or to various difficulties such as obstacles Section 9.3.2, strongly inhomogeneous cost functions Section 9.3.3, asymmetric perturbations of the curvature penalization Section 9.3.4, and optimization problems Section 9.3.5.

Remark 9.1 (Intellectual property)**.** *The numerical methods presented in this paper are available as a public and open source library[13], licensed under the Apache License 2.0, and whose development is led by J.-M. Mirebeau. Accelerations of the same order were first obtained with an earlier independent GPU implementation of the HFM [MP19] method (limited to the Dubins model) developed by L. Gayraud with the support of R. Barrere, and in informal collaboration with J.-M. Mirebeau. The two libraries are written in different languages (Python/CUDA versus C++/OpenCL), do not share a single line of code, use different implementation tricks, and offer distinct functionality.*

9.1.1 Curvature penalized path models

Throughout this paper we fix a bounded and closed domain $\Omega \subset \mathbb{R}^2$, and a continuous and positive cost function $\rho : \overline{\Omega} \times \mathbb{S}^1 \to]0, \infty[$, where $\mathbb{S}^1 := [0, 2\pi[$ with periodic boundary conditions. The objective of this paper is to compute paths $(\boldsymbol{x}, \boldsymbol{\theta}) : [0, L] \to \overline{\Omega} \times \mathbb{S}^1$ in the position-orientation state space, which globally minimize the energy

$$\mathcal{E}(\boldsymbol{x}, \boldsymbol{\theta}) := \int_0^L \rho(\boldsymbol{x}, \boldsymbol{\theta})\, \mathcal{C}(\dot{\boldsymbol{\theta}})\, \mathrm{d}l, \qquad \text{subject to } \dot{\boldsymbol{x}} = \mathbb{e}_{\boldsymbol{\theta}}, \tag{205}$$

where we denoted $\mathbb{e}_\theta := (\cos\theta, \sin\theta)$ and $\dot{\boldsymbol{\theta}} := \frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}l}$ and $\dot{\boldsymbol{x}} := \frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}l}$. An additional constraint to (205) is that the initial and final configurations $\boldsymbol{x}(0)$, $\boldsymbol{\theta}(0)$ and $\boldsymbol{x}(L)$, $\boldsymbol{\theta}(L)$ are imposed, in other words the endpoints of the physical path and the tangents at these endpoints. The path is parametrized by Euclidean length in the physical space Ω, and the total length L is a free optimization parameter. The constraint (205, right) requires that the path physical velocity $\dot{\boldsymbol{x}}(l)$ matches the direction defined by the angular coordinate $\mathbb{e}_{\boldsymbol{\theta}(l)} := (\cos\boldsymbol{\theta}(l), \sin\boldsymbol{\theta}(l))$, for all $l \in [0, L]$. This constraint is said *non-holonomic* because it binds together the some of the first order derivatives of the path $(\dot{\boldsymbol{x}}, \dot{\boldsymbol{\theta}})$.

The choice of curvature penalty function $\mathcal{C}(\kappa)$, where $\kappa := \dot{\boldsymbol{\theta}}$ is the derivative of the path direction in Eq. (205), is limited to three possibilities in our approach, in contrast with the state dependent penalty ρ which is essentially arbitrary. The considered curvature penalties are defined by the following expressions, which correspond to the Reeds-Shepp[14], Euler-Mumford, and Dubins models respectively: we define $\mathcal{C}(\kappa)$, for all $\kappa \in \mathbb{R}$, as either

$$\sqrt{1 + \kappa^2}, \qquad 1 + \kappa^2, \qquad 1 + \infty_{|\kappa|>1}, \tag{206}$$

where ∞_{cond} stands for $+\infty$ where *cond* holds, and 0 elsewhere. The Reeds-Shepp model penalizes curvature in a roughly linear manner, which allows in-place rotations[15]. The quadratic curvature penalty of the Euler-Mumford model corresponds to the energy of an elastic bar, hence minimal paths follow the rest position of those objects. Finally the

[13] `www.github.com/Mirebeau/AdaptiveGridDiscretizations`

[14] The following description applies to the *forward* only variant of the Reeds-Shepp model, see Remark 9.3 for a discussion of the *reversible* variant.

[15] In full rigor, a parametrization by Euclidean length in the full state space (both physical and angular), or an arbitrary Lipschitz parametrization, is necessary to ensure the existence of a minimizer of (205) for the Reeds-Shepp forward model. Indeed, in-place rotations are path sections where the the physical velocity vanishes, but the angular velocity does not. See [Mir18] for a discussion of well posedness.

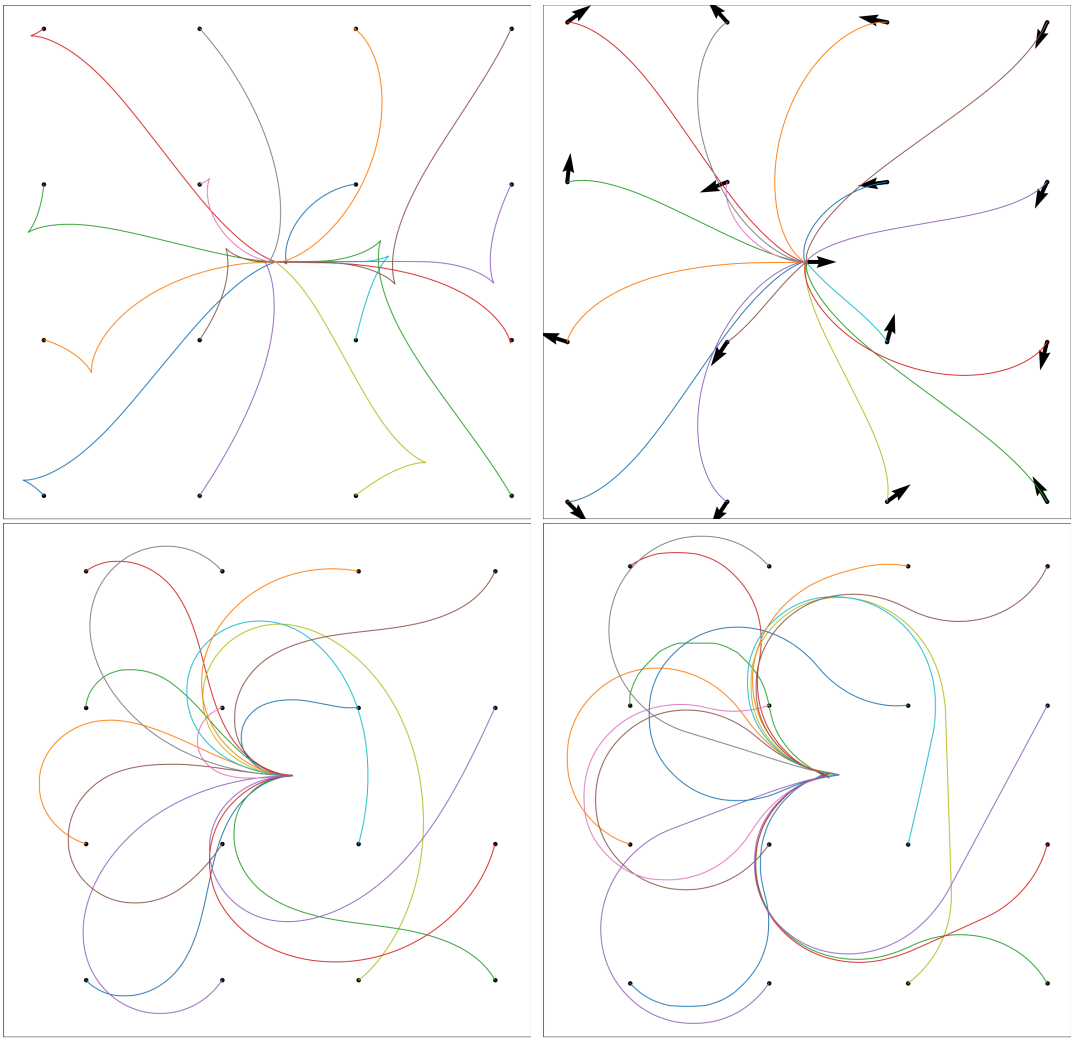

Figure 50: Planar projections of minimal geodesics for the Reeds-Shepp, Reeds-Shepp forward, Elastica and Dubins models (left to right). Seed point $(0, 0)$ with horizontal tangent, regularly spaced tip point with random tangent (but identical for all models).

Dubins model forbids any path section whose curvature exceeds that of the unit disk, by assigning to it the cost $+\infty$. Minimal paths for these models are qualitatively distinct, as illustrated on Fig. 50. The curvature penalty may also be scaled and shifted, so as to control its strength and symmetry, see Remark 9.2 and Section 9.3.4.

In the following, we fix a seed point $(x_*, \theta_*) \in \Omega \times \mathbb{S}^1$ in the state space, and denote by $u(x, \theta)$ the minimal cost of a path from this seed to an arbitrary target $(x, \theta) \in \Omega \times \mathbb{S}^1$:

$$u(x,\theta) := \inf\{\mathcal{E}(\boldsymbol{x},\boldsymbol{\theta}); L \geq 0,\ (\boldsymbol{x},\boldsymbol{\theta}) : [0,L] \to \Omega \times \mathbb{S}^1,\ \dot{\boldsymbol{x}} = \mathrm{e}_{\boldsymbol{\theta}},\\ \boldsymbol{x}(0) = x_*,\ \boldsymbol{\theta}(0) = \theta_*,\ \boldsymbol{x}(L) = x,\ \boldsymbol{\theta}(L) = \theta\}. \tag{207}$$

Once the map $u : \Omega \times \mathbb{S}^1 \to \mathbb{R}$ is numerically computed, as described in Section 9.1.2, a standard backtracking technique [Mir18] allows to extract the path $(\boldsymbol{x}, \boldsymbol{\theta}) : [0, L] \to \overline{\Omega} \times \mathbb{S}^1$ globally minimizing (205), from the seed state (x_*, θ_*) to any given target $(x^*, \theta^*) \in \Omega \times \mathbb{S}^1$.

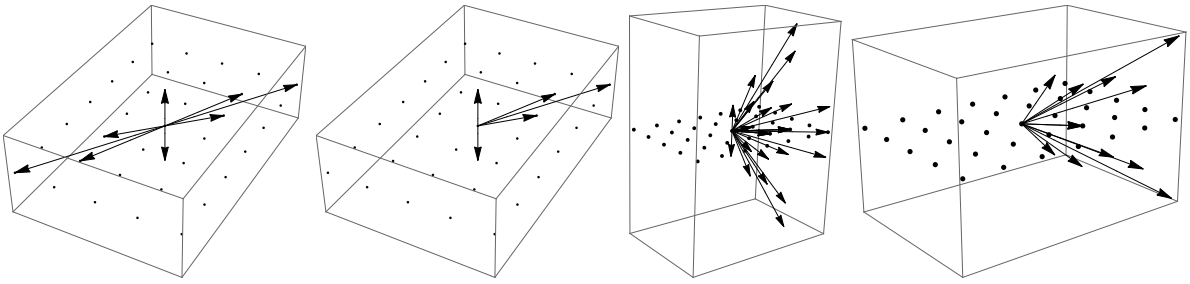

Figure 51: Discretization stencils used for the Reeds-Shepp reversible, Reeds-Shepp forward, Euler-Mumford, and Dubins models. Note the sparseness and anisotropy of the stencils. Model parameters: $\theta = \pi/3$, $\xi = 0.2$, $\varepsilon = 0.1$.

Remark 9.2 (Scaling and shifting the curvature penalty)**.** *The curvature penalty $\mathcal{C}(\dot{\boldsymbol{\theta}})$ appearing in our path models* (205) *can be generalized into $\mathcal{C}(\xi(\dot{\boldsymbol{\theta}} - \varphi))$. The parameter $\xi > 0$ dictates the intensity of curvature penalization, whereas $\varphi \in \mathbb{R}$ can introduce asymmetric penalty. Optionally, $\xi = \xi(x, \theta)$ and $\varphi = \varphi(x, \theta)$ may depend on the current state $(x, \theta) \in \Omega \times \mathbb{S}^1$.*

9.1.2 Non-holonomic eikonal equations, and their discretization

The minimal travel cost (207), from a given source point to an arbitrary target, is the value function of a deterministic optimal control problem. As such, it obeys a first order static non-linear PDE, a variant of the eikonal equation, of the generic form

$$\mathcal{F}u(x,\theta) = \rho(x,\theta) \qquad \text{where} \quad \mathcal{F}u(x,\theta) = \mathfrak{F}(x, \theta, \nabla_x u(x,\theta), \partial_\theta u(x,\theta)),$$

where $\nabla_x u(x,\theta) \in \mathbb{R}^2$ and $\partial_\theta u(x,\theta) \in \mathbb{R}$ denote the partial derivatives of the unknown $u : \Omega \times \mathbb{S}^1 \to \mathbb{R}$ w.r.t the physical position x and angular coordinate θ. This PDE holds in $\Omega \times \mathbb{S}^1 \setminus \{(x_*, \theta_*)\}$, while the constraint $u(x_*, \theta_*) = 0$ is imposed at the seed point (x_*, θ_*), and outflow boundary conditions are applied on $\partial\Omega$. The detailed arguments and adequate concepts of optimal control, Hamilton-Jacobi-Bellman equations, and discontinuous viscosity solutions, are non-trivial and unrelated to the object of this paper (which is GPU acceleration), hence we simply refer the interested reader to [BCD08, Mir18]. For comparison, the standard isotropic eikonal equation [RT92, Set99] on $\mathbb{R}^d$, which corresponds to an omni-directional vehicle not subject to maneuverability constraints or curvature penalization, is defined by the operator $\mathcal{F}u = \|\nabla u\|$.

The considered variants of the eikonal PDE involve the following non-linear and anisotropic operators $\mathcal{F}u(x,\theta)$, see [Mir18].

$$\sqrt{\langle \nabla_x u, \mathbb{e}_\theta \rangle^2 + |\partial_\theta u|^2}, \qquad \sqrt{\max\{0, \langle \nabla_x u, \mathbb{e}_\theta \rangle\}^2 + |\partial_\theta u|^2}, \tag{208}$$

$$\frac{1}{2}(\langle \nabla_x u, \mathbb{e}_\theta \rangle + \sqrt{\langle \nabla_x u, \mathbb{e}_\theta \rangle^2 + |\partial_\theta u|^2}), \qquad \langle \nabla_x u, \mathbb{e}_\theta \rangle + |\partial_\theta u|. \tag{209}$$

They respectively correspond to the Reeds-Shepp reversible (208, left), Reeds-Shepp forward (208, right), Euler-Mumford (209, left) and Dubins (209, right) models.

We rely on a finite differences discretization Fu of the operator $\mathcal{F}u$, on the Cartesian grid

$$X_h := (\Omega \times \mathbb{S}^1) \cap h\mathbb{Z}^3, \tag{210}$$

where the physical domain is usually rectangular $\Omega = [a, b] \times [c, d]$ (or padded as such), and where the grid scale $h > 0$ is such that $2\pi/h \in \mathbb{N}$ so that the samping of $\mathbb{S}^1 := [0, 2\pi[$ is compatible with the periodic boundary conditions. By convention, the value function u is extended by $+\infty$ outside Ω, thus implementing the desired outflow boundary conditions on $\partial\Omega$. For any discretization point $p = (x, \theta) \in X_h$, the finite differences operator $Fu(p)$ is defined as the square root of the following expression, considered in [MP19]

$$\max_{1\le k\le K} \Big(\sum_{1\le i\le I} \alpha_{ik} \max\big\{0, \frac{u(p) - u(p + he_{ik})}{h}\big\}^2 + \sum_{1\le j\le J} \beta_{jk} \max_{\sigma=\pm 1} \big\{0, \frac{u(p) - u(p + \sigma h f_{jk})}{h}\big\}^2 \Big), \tag{211}$$

where I, J, K are fixed integers, $\alpha_{ik}, \beta_{jk} \ge 0$ are non-negative weights, and $e_{ik}, f_{jk} \in \mathbb{Z}^3$ are finite difference offsets, for all $1 \le i \le I$, $1 \le j \le J$, $1 \le k \le K$. The weights and offsets may depend on the current point p. This framework, which can address the variants (208) and (209), is a generalization of the standard discretization [RT92] of the isotropic eikonal equation ($\mathcal{F}u = \|\nabla u\|$), obtained with meta-parameters $J = d$ (and $I = 0$, $K = 1$), choosing unit weights $w_{j1} = 1$, $1 \le j \le d$, and letting $(f_{j1})_{i=1}^d$ be the canonical basis of $\mathbb{R}^d$. Riemannian eikonal PDEs can also be addressed in this framework, with $J = d(d + 1)/2$ (and $I = 0$, $K = 1$) and using weights and offsets defined by an appropriate decomposition of the inverse metric tensor, see [Mir19, MP19]. The anisotropy of the Riemannian metric is not bounded a-priori, but strong anisotropy leads to large stencils : specifically $\|f_{j1}\| \le C\sqrt{\|M\|\|M^{-1}\|}$ for all $1 \le j \le J$ in dimension $d \le 3$, where M denotes the Riemannian metric tensor, see [Mir19, Proposition 1.1]. Excessively large stencils in turn lead to longer execution time due to cache misses, slower convergence of the iterative method, and less precise boundary conditions.

In the curvature penalized case, the weights and offsets in (211) implicitly depend on the base point $p = (x, \theta)$, at least through the angular coordinate θ in view of the continuous PDE (208) and (209). They may depend on the physical position x as well if the strength or symmetry of the curvature penalty varies from point to point, see Remark 9.2. We refer to [Mir18, MP19] for details on the construction of the weights and offsets, which involves a relaxation parameter $\varepsilon > 0$ for the non-holonomic constraint (205, right), and simply report here the meta-parameters for the Reeds-Shepp forward ($I = 3$, $J = 1$, $K = 1$), Euler-Mumford ($I = 30$, $J = 0$, $K = 1$), and Dubins ($I = 6$, $J = 0$, $K = 2$) models, see Figure 51.

A fundamental property of discretization schemes of the form Eq. (211) is that they can be solved in a single pass over the domain, using a generalization of the fast-marching algorithm [Mir18, MP19, Mir19]. This is highly desirable when implementing CPU solver, but anecdotical for a GPU eikonal solver whose massive parallelism forbids taking advantage of this property. Nevertheless, those schemes are robust and well tested. Alternative approaches offering different compromises and possibly more suited to GPUs will be considered in future works.

Remark 9.3 (Forward and reversible Reeds-Shepp models)**.** *The Reeds-Shepp model comes in two flavors [DMMP18]: the* forward *variant, presented above, and the (more*

standard) reversible *variant, modeling a vehicle equipped with a reverse gear additionally. The latter is obtained by relaxing the constraint (205, right) into* $\dot{\boldsymbol{x}} = \pm \circledcirc_{\boldsymbol{\theta}}$. *In turn the eikonal PDE* (208) *is replaced with* $\sqrt{\langle \nabla_x u, \circledcirc_\theta \rangle^2 + |\nabla_\theta u|^2}$, *whose discretization* (211) *uses the meta-parameters* $I = 0$, $J = 4$, $K = 1$.

Remark 9.4 (Monotony and degenerate ellipticity)**.** *The discrete operator* (211) *is* degenerate elliptic*: $Fu(p)$ is a non-decreasing function of the finite differences $[u(p) - u(q)]_{q \in X_h \setminus \{p\}}$. This property implies a comparison principle, used in the proof of convergence of the numerical method [Mir18]. In addition, degenerate ellipticity implies a monotony property of the local update operator implemented in Algorithm 9 below, see [Mir19, Proposition A.4]. As a result, the sequence of approximate solutions $(u_n)_{n \geq 0}$, $u_n : X_h \to [0, \infty]$, produced along the iterations of our numerical method are pointwise non-increasing. In the implementation Section 9.2, these properties allow to use a single array for reading and writing the solution values, as stability is guaranteed independently of data races.*

9.2 Implementation

We describe the implementation of our massively parallel solver of generalized eikonal PDEs, assumed to be discretized in the form (211). The bulk of the method is split in three procedures, Algorithms 7 to 9, discussed in detail in the corresponding sections.

For simplicity, Algorithms 8 and 9 are written in the special case where the meta parameters of the discretization (211) are $J = 0$ and $K = 1$, whereas I is arbitrary. The case of arbitrary J and K is discussed in §9.2.3. The assignment of a value *val* to a scalar (resp. array) variable *var* is denoted $var \leftarrow val$ (resp. $var \Leftarrow val$).

Algorithm 7 Parallel iterative solver (Python)

Variables:
- $u : X_h \to [0, \infty]$ (The problem unknown)
- $active, next : B_h \to \{0, 1\}$. (Blocks marked for current and next update)

Initialization:
- $u \Leftarrow \infty$; $active, next \Leftarrow 0$.
- $u[p_*] \leftarrow 0$; $active[b_*] \leftarrow 1$. (Set seed point value, and mark its block for update)

While an *active* block remains:
- **For all** *active* blocks b in parallel: (CUDA kernel lauch)
  - **For all** $p \in X_h^b$ in parallel: (Block of threads)
    - BlockUpdate$(u, next, b, p)$
- $active \Leftarrow next$; $next \Leftarrow 0$.

9.2.1 Parallel iterative solver

Massively parallel architectures divide computational tasks into *threads* which, in the case of graphics processing units, are grouped into *blocks* following a common sequence of instructions, and able to take advantage of shared data, see Remark 9.5. Following

Algorithm 8 BlockUpdate$(u, next, b, p)$, where $p \in X_h^b$ (CUDA)

Global variables: $u : X_h \to [0, \infty]$, $next : B_h \to \{0, 1\}$, $\rho : X_h \to \mathbb{R}$ (the r.h.s).
Block shared variable: $u_b : X_h^b \to [0, \infty]$.
Thread variables: $\alpha_i \geq 0$, $e_i \in \mathbb{Z}^d$, $u_i \in \mathbb{R}$, for all $1 \leq i \leq I$.

$u_b(p) \leftarrow u(p)$; __syncthreads() (Load main memory values into shared array)
Load or compute the stencil weights $(\alpha_i)_{i=1}^I$ and offsets $(e_i)_{i=1}^I$.
$u_i \leftarrow u(p + he_i)$, for all $1 \leq i \leq I$ such that $p + he_i \notin X_h^b$. (Load the neighbor values)

For r from 1 to R:
    $u_i \leftarrow u_b(p + he_i)$, for all $1 \leq i \leq I$ such that $p + he_i \in X_h^b$. (Load shared values)
    $u_b(p) \leftarrow \Lambda(\rho(p), \alpha_i, u_i, 1 \leq i \leq I)$ (Update $u_b(p)$, unless p is the seed point)
    __syncthreads() (Sync shared values)

$u(p) \leftarrow u_b(p)$ (Export shared array values to main memory)
If appropriate, $next[b] \leftarrow 1$ and/or $next[b'] \leftarrow 1$ for each neighbor block b' of b. (Thread 0 only)

Algorithm 9 Local update operator $\Lambda(\rho, \alpha_i, u_i, 1 \leq i \leq I)$ (C++)

Variables $a \leftarrow 0$, $b \leftarrow 0$, $c \leftarrow -h^2\rho^2$, $\lambda \leftarrow \infty$.
Sort the indices, so that $u_{i_1} \leq \cdots \leq u_{i_I}$.
For r from 1 to I:
    If $\lambda \leq u_{i_r}$ **then** break.
    $a \leftarrow a + \alpha_{i_r}$; $b \leftarrow b + \alpha_{i_r} u_{i_r}$; $c \leftarrow c + \alpha_{i_r} u_{i_r}^2$
    $\lambda \leftarrow (b + \sqrt{b^2 - ac})/a$
return λ

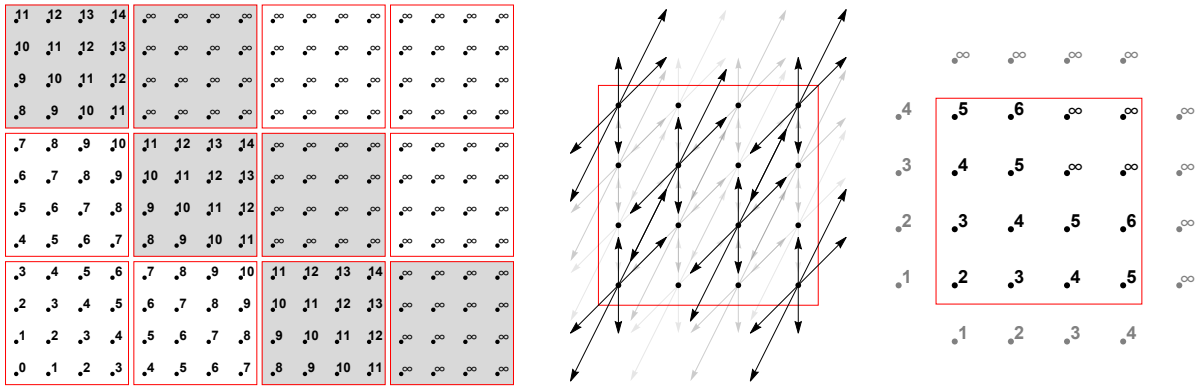

Figure 52: *Left:* Decomposition of the Cartesian grid X_h into tiles X_h^b, with block index $b \in B_h$. Grayed blocks are tagged *active*. *Center:* Updating a block $b \in B_h$ requires loading the unknown values $u : X_h \to \mathbb{R}$, both within X_h^b and at some neighbor points. *Right:* Several local updates are performed within a block (two here).

[WDB+08, JW08, GHZ18], the main loop of our iterative eikonal equation solver is designed to take advantage of this computational architecture, see Algorithm 7. It is written in the Python programming language, which is also used for the pre- and post-processing tasks, and launches Algorithm 8 as a CUDA kernel via the cupy[16] library.

The discretization domain X_h, which is a three dimensional Cartesian grid (210), is split into rectangular tiles X_h^b, indexed by $b \in B_h$, see Figure (52, left). The update of a tile X_h^b is handled by a block of threads, and the tile should therefore contain no less than 32 points in view of Remark 9.5. The best shape of the tiles X_h^b was found to be $4 \times 4 \times 4$ for the Reeds-Shepp models (forward and reversible), and $4 \times 4 \times 2$ for the Euler-Mumford and Dubins models, see Section 9.2.2 and Table 7. One of the findings of our work is indeed that the schemes featuring wider stencils work best with smaller tile sizes, see also the discussion in the second paragraph of Section 9.2.2. Some padding is introduced if the dimensions of the tiles X_h^b do not divide those of the grid X_h.

A boolean table $active : B_h \to \{0, 1\}$ records all tiles tagged for update. Denote by $N_h := \#(B_h)$ the total number of tiles, and by $N_b = \#(X_h^b)$ the number of grid points in a tile, which is independent of $b \in B_h$, so that $\#(X_h) = N_h N_b$ by construction. Let also $N_{act} = \#\{b \in B_h;\ active[b]\}$ be the number of active tiles in a typical iteration of Algorithm 7. Since we are implementing a front propagation in a three dimensional domain, one generally expects that $N_{act} \approx N_h^{2/3}$ (in d-dimensions, $N_{act}^b \approx N_h^{1-1/d}$).

In each iteration of Algorithm 7, the *active* table is checked for emptiness, in which case the program terminates. More importantly, the indices of all non-zero entries of the *active* table are extracted, so as to update only the relevant blocks. The complexity $\mathcal{O}(N_h \ln N_h)$ of this operation is in practice negligible w.r.t the cost of the block updates themselves $\mathcal{O}\big(N_{act} N_b R K (I + J)\big)$ where R is the number of inner loops in Algorithm 8 and I, J, K are the scheme parameters (211). A second boolean table $next : B_h \to \{0, 1\}$, is used to mark the blocks which are to be updated in the subsequent iteration.

A single array $u : X_h \to [0, \infty[$ holds the solution values. Indeed, the block update operator benefits from a monotony property, see Remark 9.4, which guarantees that the values of $(u_n)_{n \geq 0}$ of the approximation solution *decrease* along the iterations of Algorithm 7 toward a limit u_∞. As a result, load/store data races in u between the threads are innocuous.

Remark 9.5 (SIMT architecture)**.** *A block of threads is under the hood handled by a GPU device in a Single Instruction Multiple Threads (SIMT) manner : the same instructions are applied on 32 threads of a same block (also called a warp) simultaneously. For this reason, the number of threads within a block should preferably be a multiple of the width of a warp. For the same reason, thread divergence (threads within a warp going along different execution paths, due to conditional branching statements, implemented by "muting" the threads of the inactive branch) should be avoided for best efficiency.*

Our numerical solver reflects these properties through the choice of the tile size X_h^b, and in the choice of an Eulerian discretization scheme on a Cartesian grid (211) *whose solution by Algorithm 9 involves little branching and yields an even load between threads, as opposed to an unstructured mesh where these properties are by design less ensured [FKW13].*

[16] A NumPy-compatible array library accelerated by CUDA. `https://cupy.dev`

9.2.2 Block update

The BlockUpdate procedure, presented in Algorithm 8, is the most complex part of our numerical method. It is executed in parallel by a block of threads, each handling a given point $p \in X_h^b$ of a tile of the computational grid, where the tile index $b \in B_h$ is fixed.

A array $u_b : X_h^b \to [0, \infty]$ *shared* between the threads of the block is initialized with the values of the unknown $u : X \to [0, \infty]$ at the same positions. Throughout the execution of the BlockUpdate procedure, the values of u_b are updated several times, and then finally stored by in the main array u. If the number R of updates of u_b is sufficiently large, then this procedure amounts to solving a local eikonal equation on X_h^b, with $u_{|X_h \setminus X_h^b}$ treated as boundary conditions. A similar approach is used in [WDB+08, JW08, GHZ18]. We empirically observe that stencil schemes using a wide stencil work best with small number R of iterations, and a small tile size, see Table 7. Our interpretation is the following: using several iterations is meant to propagate the front through the tile X_h^b and stabilize the local solution u_b within the tile [JW08], but this objective loses relevance when the stencil is so wide that the scheme update at a point $x \in X_h^b$ involves fewer values of the local array u_b in X_h^b than of the global array u in $X_h \setminus X_h^b$, which is cached and frozen throughout the iterations in Algorithm 8. In addition, each of the R iterations has a higher cost when the stencil is wide and has numerous elements, see the description of Algorithm 9 in Section 9.2.3.

The finite difference scheme (211) used for curvature penalized fast marching is built using non-trivial tools from lattice geometry [Mir18], whose numerical cost cannot be ignored. Empirical tests show that precomputing the weights and offsets usually reduces overall computation time by 30% to 50%. If the scheme structure only depends on the angular coordinate θ of the point, then the precomputed stencils can be shared across all physical coordinates x and thus have a negligible memory usage, so that these precomputations are a pure benefit. On the other hand, if the scheme stencils depend on all coordinates (x, θ) of the current point, typically for a model whose curvature penalty function depends on the current point as discussed in Remark 9.2 and Section 9.3.4, then the storage cost of the weights and offsets significantly exceeds the problem data. (Stencils are defined by $N = K(I + J)$ scalars and offsets per grid point, see (211), where typically $4 \leq N \leq 30$. In comparison, the problem data u, ρ and optionally ξ, φ consists of 2 to 4 scalars per grid point, see Remark 9.2.) Stencil recomputation is preferred in these cases, in order to avoid crippling the ability of the numerical method to address large scale problems on memory limited GPUs.

The values of the unknown $u : X_h \to \mathbb{R}$ needed for the evaluation of the scheme (211) and lying outside X_h^b are loaded once and for all at the beginning of the BlockUpdate procedure Algorithm 8, and treated as fixed boundary conditions so as to minimize memory bandwidth usage. Contrary to what could be expected, such boundary values are an overwhelming majority in comparison with the values located within the tile X_h^b. For instance the three dimensional isotropic eikonal equation, using standard tiles of $64 = 4 \times 4 \times 4$ points, involves $96 = 6 \times 4 \times 4$ boundary values. Boundary values are even more numerous with the curvature penalization schemes, which involve many wide finite difference offsets, as illustrated on Figure (52, center).

Each thread of a block, associated to a discretization point $p \in X_h^b$ where $b \in B_h$ is

the block index, goes through R iterations of a loop where the local unknown value $u_b(p)$ is updated via Algorithm 9, see §9.2.3. The threads are synchronized at each iteration of this loop, to ensure that the front propagates through the tile X_h^b. Since the values of $u_b : X_h^b \to [0, \infty]$ are decreasing along the iterations, by monotony of the scheme see Remark 9.4, no additional protection of u_b against data races between the threads of the block is required. The number R of iterations is discussed in §9.2.3.

Last but not least, if appropriate, the block b and its immediate neighbors b' need to be tagged for update in the next iteration of the eikonal solver Algorithm 7, via the boolean array $next : B_h \to \{0, 1\}$. This step is *not* fully described in Algorithm 8, and in particular the *neighbors* of a tile and the *appropriate* condition for marking them are not specified. Indeed, a variety of strategies can be plugged in here, and our numerical solver is not tied to any of them. Good results were obtained using Adaptive Gauss Seidel Iteration (AGSI) [BR06, GHZ18] and with the Fast Iterative Method (FIM) [JW08], while other variants were not tested [WDB+08].

Remark 9.6 (Walls and thin obstacles)**.** *Our finite differences scheme involves rather wide stencils, see Figure 51, raising the following issue: the update of a point p may involve neighbor values $u(p + he_i)$ across a thin obstacle. In order to avoid propagating the front through the obstacles, if any are present, an additional* walls *array is introduced in Algorithm 8, as well as and intersection test between the segment $[p, p + he_i]$ and the obstacles. For computational efficiency, the array* walls $: X_h \to \{0, \cdots, 255\}$ *is not boolean, but* walls$[p]$ *instead encodes the Manhattan distance in pixels (capped at 255) from the current point p to the nearest obstacle. If $\|e_i\|_1 <$ walls$[p]$, then $[p, p + he_i]$ does not meet the obstacles, and the intersection test can be bypassed.*

9.2.3 Local update

This section is devoted to the local update operator presented in Algorithm 9. From the mathematical standpoint, one defines $\Lambda u(p)$ as the solution to the equation $Fu(p) = \rho(p)$ w.r.t the variable $u(p)$, regarding all neighbor values as constants, see [Mir19, Appendix A]. We prove in this subsection that Algorithm 9 does compute this value, and comment on its numerical complexity and efficient implementation. A related method is used in the update step of the standard fast marching method for isotropic eikonal equations [Set96], whose discretization is a special case of (211) as mentioned Section 9.1.2.

For simplicity, and consistently with the presentation of Algorithm 9, we first assume a numerical scheme of the form

$$(Fu(x))^2 := h^{-2} \sum_{i=1}^{I} \alpha_i \big(u(x) - u(x + he_i)\big)_+^2, \tag{212}$$

in other words $J = 0$ and $K = 1$ in (211). Denote $u_i := u(x + he_i)$ for all $1 \le i \le I$, and let $\rho := \rho(p)$. The update value $\lambda = \Lambda u(x)$ is by construction the unique root $\lambda \in \mathbb{R}$ of

$$f(\lambda) := \sum_{1 \le i \le I} \alpha_i (\lambda - u_i)_+^2 - h^2 \rho^2, \tag{213}$$

where $a_+ := \max\{0, a\}$. Note that $f(\lambda) = -h^2\rho^2 < 0$ on $]-\infty, \lambda_*]$ where $\lambda_* := \min\{u_i; 1 \leq i \leq I\}$, that f increasing on $[\lambda_*, \infty[$, and that $f(\lambda) \to \infty$ as $\lambda \to \infty$. Thus f does indeed admit a unique root λ, by the intermediate value theorem.

The numerical solution of the non-linear equation (213) takes advantage of its piecewise quadratic structure. For that purpose, introduce a permutation $i_1, \cdots, i_I$ of $\{1, \cdots, I\}$ such that $u_{i_1} \leq \cdots \leq u_{i_I}$. Then for any $1 \leq r \leq I$ on has

$$f(\lambda) = a_r\lambda^2 - 2b_r\lambda + c_r, \qquad \text{for all} \quad \lambda \in [u_{i_r}, u_{i_{r+1}}],$$

with the abuse of notations $[u_{i_I}, u_{i_{I+1}}] := [u_{i_I}, \infty[$. The coefficients of this quadratic function are

$$a_r := \sum_{1\leq s\leq r} \alpha_{i_s}, \qquad b_r := \sum_{1\leq s\leq r} \alpha_{i_s} u_{i_s}, \qquad c_r := \sum_{1\leq s\leq r} \alpha_{i_s} u_{i_s}^2 - h^2\rho^2.$$

Algorithm 9 solves the quadratic equations $a_r\lambda^2 - 2b_r\lambda + c_r = 0$ successively, for increasing values of $1 \leq r \leq I$. Only the largest of the two quadratic roots is relevant, denoted $\lambda_r := (b_r + \sqrt{b_r^2 - a_r c_r})/a_r$, and it is returned if $\lambda_r \in [u_{i_r}, u_{i_{r+1}}]$, at some rank denoted $r = r_*$. By construction, the root λ_r exists and is real for all $1 \leq r \leq r_*$, since $a_r > 0$ and $f(u_{i_r}) < 0$.

From the complexity standpoint, Algorithm 9 begins with a sort of I values, followed by $\mathcal{O}(I)$ floating point operations. In standard isotropic fast-marching, the number of terms is the space dimension $I \in \{2, 3\}$ and the sorting step has a negligible cost, but wide stencil schemes behave differently, especially the Euler-Mumford model for which $I = 30$. In the latter case, a naive bubble sort with complexity $\mathcal{O}(I^2)$ becomes prohibitive, whereas an adequate sorting method cuts overall computation time but more than half. Best results were obtained applying a network sort [Knu98] (an efficient branchless sorting method) to the 15 first (resp. last) values, followed by a merge operation.

If the numerical scheme has the generic form (211), with arbitrary parameters I, J, K, then the update $\lambda = \Lambda u(p)$ is the unique root of $f : \mathbb{R} \to \mathbb{R}$ defined by

$$f(\lambda) := \max_{1\leq k\leq K} f_k(\lambda), \quad \text{where } f_k(\lambda) := \sum_{1\leq i\leq I} \alpha_{ik}(\lambda - u_{ik})_+^2 + \sum_{1\leq j\leq J} \beta_{jk}(\lambda - u'_{jk})_+^2 - \rho^2 h^2,$$

and where $u_{ik} := u(p + he_{ik})$ and $u'_{jk} := \min\{u(p - hf_{jk}),\, u(p + hf_{jk})\}$. For each $1 \leq k \leq K$ two sums defining f_k are grouped into a single one over $1 \leq l \leq I + J$, and the unique root $\lambda^{(k)}$ of f_k is computed similarly to (213). The root of f is $\lambda = \min_{1\leq k\leq K} \lambda^{(k)}$. The complexity of Algorithm 9 is thus $\mathcal{O}(K(I + J))$ overall, up to logarithmic factors due to the sorting step.

9.3 Numerical experiments

We illustrate our numerical solver of curvature penalized shortest paths in variety of contexts ranging from motion planning with obstacles or drift, to image segmentation, and the configuration of radar systems. Some of the test cases are new, whereas others are closely related to previous works [CMC16, CMC17, DDBM19, DMMP18, MD17, Mir18] and meant to illustrate the benefits of the GPU solver over an earlier CPU implementation

Model	N_{fd}	Best tile	Best R
Isotropic (d=2)	2	24×24	48
Isotropic (d=3)	3	$4 \times 4 \times 4$	8
Reeds-Shepp (both)	4	$4 \times 4 \times 4$	6
Dubins	12	$4 \times 4 \times 4$	2
–	–	$4 \times 4 \times 2$	1
Euler-Mumford	30	$4 \times 4 \times 2$	1

Table 7: Number $N_{fd} = K(I + J)$ of finite differences terms in (211) for a variety of path models. Tile shape and number of iterations R in Algorithm 8, producing the smallest running time, found experimentally. Two sets of parameters are reported for Dubins model, since the corresponding running times results are close which and one is fastest depends on the test case. Simple models, whose stencil involves few and short finite differences, work best with large tile sizes and numerous iterations allowing the front to propagate within the tile, whereas complex models involving many wide finite differences and a costly update operator benefit from small tiles and few iterations.

in common use cases. Test data is synthetic except for the medical image segmentation problem Section 9.3.3.

We report in Table 8 the running times of the GPU eikonal solver presented in this paper, and of the CPU solver introduced in [Mir18], as well as the GPU/CPU speedup which varies significantly depending on the experiment. Indeed, the running time of the GPU eikonal solver, which is an iterative method, depends on the presence and layout of obstacles or slow regions in the test case as noted in [WDB+08]. This in contrast with the fast-marching-like method [Mir18] implemented on the CPU, which is guaranteed to update each discretization point at most $N_{\text{neigh}} = K(I + 2J)$ times where I, J, K are the scheme parameters (211) (for this reason, slightly abusively, fast-marching is referred to as a single pass method), and whose complexity $\mathcal{O}(N_{\text{neigh}} N \ln N)$ is independent of the specific test case, where N is the total number of discretization points. However, fast-marching is limited in speed by its sequential nature.

The numerical experiments presented in the following sections are designed to illustrate the following features of the eikonal solver introduced in this paper:

1. *Geodesics in an empty domain.* Illustrates the qualitative properties of the different path models, and the GPU/CPU speedup in the ideal case.

2. *Fastest exit from a building.* Illustrates the implementation of walls and thin obstacles, which is non-trivial with wide stencils as described in Remark 9.6.

3. *Retinal vessel segmentation.* Illustrates a realistic application to image processing, based on the choice of a carefully designed cost function.

4. *Boat routing.* Illustrates a curvature penalty whose strength and asymmetry properties vary over the PDE domain, as described in Remark 9.2.

5. *Radar configuration.* Illustrates the automatic differentiation of the eikonal PDE solution u w.r.t the cost function ρ, for the optimization of a complex objective.

Exp.	model	GPU(s)	CPU(s)	accel
Empty	RS rev	0.28	34.3	120×
	RS fwd	0.25	15.7	62×
	EM	1.53	117	76×
	Dubins	0.44	46.5	105×
Building	RS rev	1.37	50.5	37×
	RS fwd	0.59	29	49×
	EM	3.21	174	54×
	Dubins	1.02	55.4	54×
Boat	Dubins	0.52	30.2	59×
MRI	RS fwd	0.93	30.8	33×
	EM	3.32	275.9	83×
Retina1	RS fwd	0.66	21.1	32×
	EM	2.22	171.3	77×
Retina2	RS fwd	0.98	32.8	33×
	EM	3.21	256.1	80×
Radar	Dubins	0.26	9.57	37×

Table 8: Running time of the CPU and GPU eikonal solver, for the experiments presented Section 9.3.

Remark 9.7 (Computation time and hardware characteristics)**.** *Program runtime is dependent on the hardware characteristics of each machine. The reported CPU and GPU times were obtained on the Blade® Shadow cloud computing service, using the provided Nvidia® GTX 1080 graphics card for the GPU eikonal solver, and an Intel® Xeon E5-2678 v3 for the CPU eikonal solver (a single thread was used, with turbo frequency 3.1Ghz).*

9.3.1 Geodesics in an empty domain

We compute minimal geodesics for the Reeds-Shepp, Reeds-Shepp forward, Euler-Mumford elastica and Dubins model, in the domain $[-1,1]^2 \times \mathbb{S}^1$ without obstacles. The front is propagated from the seed point (x_*, θ_*) placed at the origin $x_* = (0,0)$ and imposing a horizontal initial tangent $\theta_* = 0$. Geodesics are backtracked from several tips (x^*, θ^*) where x^* is placed at 16 regularly spaced points in the domain, whereas θ^* is chosen randomly (but consistently across all models).

This experiment is meant to illustrate the qualitative differences between minimal geodesic paths associated with the four curvature penalized path models, see Fig. 50. The Reeds-Shepp car can move both forward and backward, and reverse gear along its path, as evidenced by the *cusps* along several trajectories. The Reeds-Shepp forward variant cannot move backward, but has the ability to rotate in place (with a cost), and such behavior can often be observed at the endpoints of the trajectories [DMMP18]. The Elastica model produces pleasing smooth curves, which have a physical interpretation as the rest positions of elastic bars. Trajectories of the Dubins model have a bounded radius of curvature, and can be shown to be concatenations of straight segments and of arcs of circles, provided the cost function is constant as here.

The generalized eikonal PDE (208) or (209) is discretized on a $300 \times 300 \times 96$ Cartesian grid, following (211), thus producing a coupled system of equations featuring 8.6 million unknowns[17]. Computation time for the GPU eikonal solver ranges from 0.28s (Reeds-Shepp forward) to 1.54s (Euler-Mumford elastica), reflecting the complexity of the discretization stencil, see Fig. 51. A substantial speedup ranging from $60\times$ to $120\times$ is obtained over the CPU implementation; let us nevertheless acknowledge that, as noticed in [WDB+08], the absence of obstacles and of a position dependent speed function is usually the best case scenario for an iterative eikonal solver such as our GPU implementation.

9.3.2 Fastest exit from a building

We compute minimal paths within a museum map, for the four curvature penalized models under consideration in this paper. Due to the use of rather wide stencils, often 7 pixels long see Fig. 51, some intersection tests are needed to avoid propagating the front through the walls, which are one pixel thick only. A careful implementation, as described in Remark 9.6, allows to bypass most of these intersection tests and limits their impact on computation time. In contrast with [JW08], we do not consider "slightly permeable walls", since they would not be correctly handled with our wide stencils, and since as far as we know they have little relevance in applications. A closely related experiment is presented in [DMMP18] for the Reeds-Shepp models, using a CPU eikonal solver.

The front propagation starts from two seed points located at the exit doors, and a tip is placed in each room for geodesic backtracking, with an arbitrary orientation. The extracted paths are smooth (Euler-Mumford case) or have a bounded curvature radius (Dubins case), but minimize a functional (205) which is unrelated with safety and thus may not be directly suitable for motion planning. Indeed, in many places they are tangent to the obstacles, walls, and doorposts, without any visibility behind, which is a hazardous way to move.

The PDE is discretized on a Cartesian grid of size $705 \times 447 \times 60$, where the first two factors are the museum map dimension, and the third factor is the number of angular orientations, for a total of 19 million unknowns. Computation time on the GPU ranges from 0.59s (Reeds-Shepp forward) to 3.2s (Euler-Mumford elastica), a reduction by approximately $50\times$ over the CPU eikonal solver.

9.3.3 Tubular structure segmentation

A popular approach for segmenting tubular structures in medical images, such as blood vessels on the retinal background in this experiment, is to devise a geometric model whose minimal paths (between suitable endpoints) are the centerlines of the desired structures. For that purpose a key ingredient, not discussed here, is the careful design of a cost function $\rho : \mathbb{R}^2 \times \mathbb{S}^1 \to]0, \infty]$ which is small along the vessels of interest in their tangent direction, and large elsewhere [PKP09]. Curvature penalization, and in particular the Reeds-Shepp forward and Euler-Mumford elastica models [CMC16, CMC17, DMMP18],

[17] For this particularly simple problem (with a constant cost function, without walls), results visually quite similar can be obtained at a fraction of the cost using a smaller discretization grid, eg. of size $100 \times 100 \times 64$.

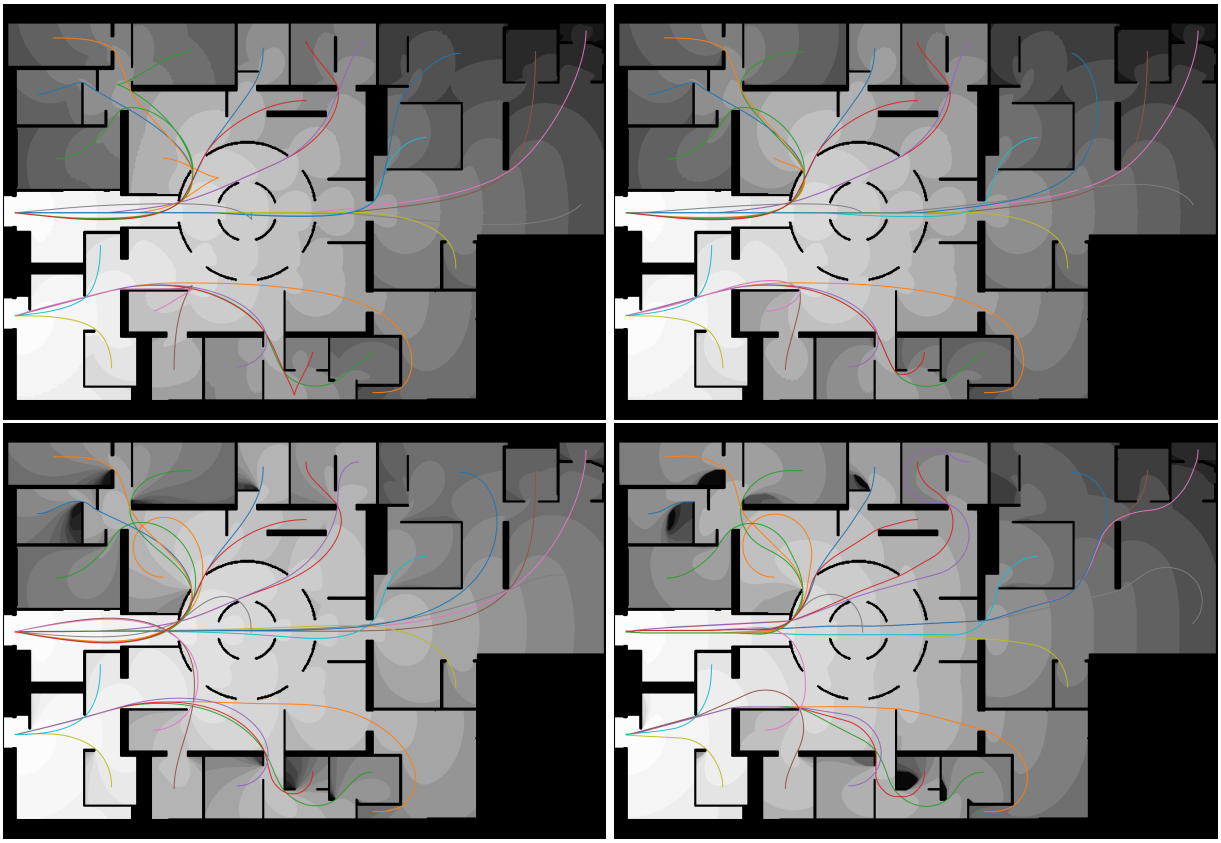

Figure 53: Planar projections of minimal geodesics for the Elastica and Dubins model (left to right). Two seed points at the exits, with horizontal tangents. Geodesics are backtracked from one tip point in each room, with a given but arbitrary tangent.

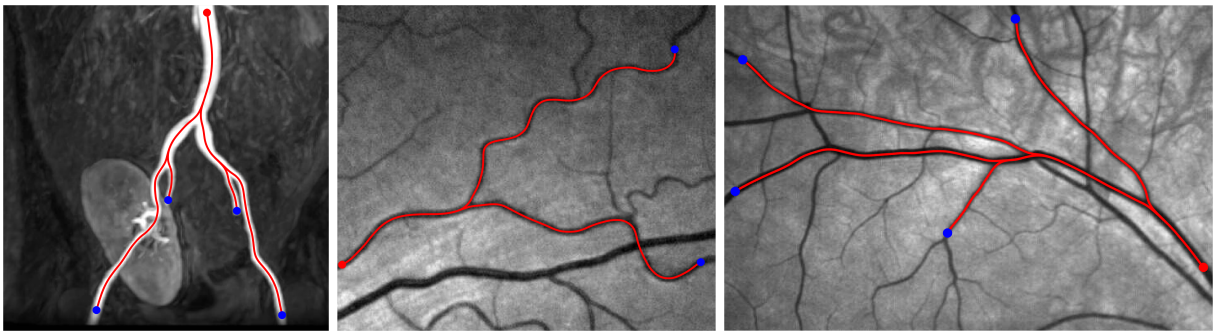

Figure 54: Segmentation of tubular structure centerlines using the Reeds-Shepp forward and Euler-Mumford elastica models, following [CMC17]. Left : Blood vessels in Magnetic Resonance Angiography (MRA) data. Center and right : Blood vessels on an image of the retina.

helps avoid a classical artifact where the minimal paths do not follow a single vessel but jump form one to another at crossings.

The test cases have size $512\times512\times60$, $387\times449\times60$ and $398\times598\times60$ respectively, and the computation time of the GPU eikonal solver ranges from 1s (Reeds-Shepp forward) to 3s (Euler-Mumford elastica) on the GPU. This is compatible with user interaction, in contrast with CPU the run time which is $30\times$ to $80\times$ longer, see Table 8. Note that by construction, the front propagation is fast along the blood vessels, and slower in the rest of the domain. This specificity plays against iterative methods, which are most efficient when velocity is uniform [JW08], yet the speedup achieved by the GPU solver remains very substantial. Computation time could in principle be further reduced, both on the CPU and the GPU, by using advanced stopping criteria and restriction methods [CCV13] to avoid solving the eikonal PDE on the whole domain.

9.3.4 Boat routing with a trailer

The Dubins-Zermelo-Markov model [BT13] describes a vehicle subject to a drift, and whose speed and turning radius *as measured before the drift is applied* are bounded. This problem was introduced to us in the context of maritime seismic prospection, where boats drag long trails of acoustic sensors, and are subject to water currents. Optimal Dubins-Zermelo-Markov trajectories, with drift defined by the water flow, may help avoid entangling and damaging these trails, and reduce the prospection times. In this synthetic experiment we use the drift velocity $V(x) = 0.6\sin(\pi x_0)\sin(\pi x_1)x/\|x\|$ on the domain $[-1,1]^2$. Our vehicle has unit speed, and turning radius $\xi = 0.3$.

From the mathematical standpoint, the Dubins-Zermelo-Markov model can be rephrased in the form of the original Dubins model, but with a curvature penalty which is scaled, shifted (asymmetric), and depends on the current point, as described in Remark 9.2. This does not raise particular issues for discretization, except that the weights and offsets of the numerical scheme (211) depend on the full position $(x,\theta) \in \mathbb{R}^2 \times \mathbb{S}^1$, rather than the orientation $\theta \in S^1$ alone.

The boat routing problem is discretized on a grid of size $151 \times 151 \times 96$. Computation time on the GPU is 0.34s if stencils are pre-computed and stored, and 0.52s if they

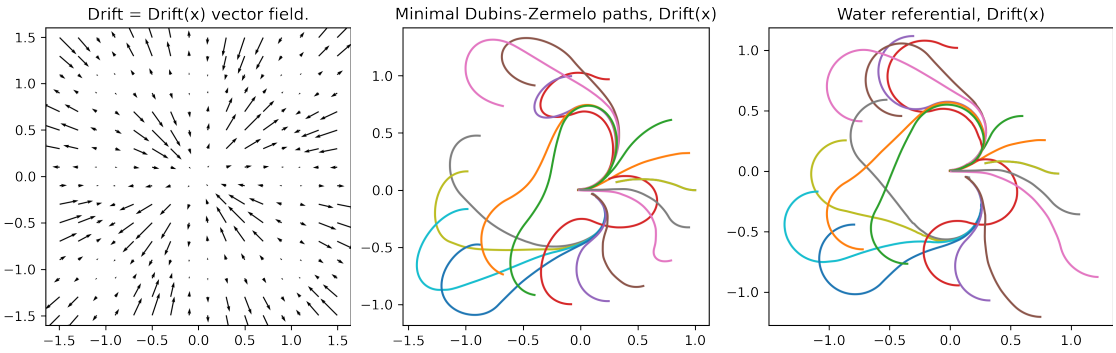

Figure 55: Illustration of the Dubins-Zermelo-Markov problem. Let drift velocity (water current). Center shortest path, between the seed point $(0, 0)$ with horizontal tangent, and other seed points, such that the radius of curvature in the water referential does not exceed the prescribed bound.

are recomputed on the fly when needed. The second approach (recomputation) uses significantly less GPU memory, which is usually a scarce ressource, hence we regard it as default despite the longer runtime, see the discussion Section 9.2.2; it is nevertheless 59× faster than the CPU implementation.

9.3.5 Optimization of a radar configuration

We consider the optimization of a radar system, so as to maximize the probability of detection of an intruder vehicle. The intruder has full knowledge of the radar configuration, and does its best to avoid detection, but is subject to maneuverability constraints as does a fast plane. Following [MD17, DDBM19] the intruder is modeled as a Dubins vehicle, traveling at unit speed with a turning radius of 0.2, whose trajectory starts and ends at a given point $x_* \in \Omega$ and which must visit a target keypoint $x^* \in \Omega$ in between[18]. The problem takes the generic form

$$\sup_{\xi \in \Xi} \inf_{\gamma \in \Gamma} \mathcal{E}(\xi; \gamma), \tag{214}$$

where Ξ is the set of radar configurations, and Γ is the set of admissible trajectories. A trajectory γ escapes detection from a radar configured as ξ with probability $\exp(-\mathcal{E}(\xi; \gamma))$. Following (205), a trajectory is represented as a pair $\gamma = (\boldsymbol{x}, \boldsymbol{\theta}) : [0, L] \to \Omega \times \mathbb{S}^1$, and its cost is defined as

$$\mathcal{E}(\xi; \gamma) = \int_0^L \rho(\boldsymbol{x}, \boldsymbol{\theta}; \xi)\, \mathcal{C}(\dot{\boldsymbol{\theta}})\, \mathrm{d}l$$

where $\mathcal{C}$ denotes the Dubins cost (206, right), and $\rho(x, \theta; \xi)$ is an instantaneous probability of detection depending on the radar configuration ξ, and the intruder position x and orientation θ. We refer to [DDBM19] for a discussion of the detection probability model, and settle for a synthetic and simplified yet already non-trivial construction. The detection probability is the sum of three terms $\rho(x, \theta; \xi) = \sum_{i=1}^{3} \tilde{\rho}(x, \theta; y_i, r_i, v_i)$, corresponding to

[18] This is achieved by concatenating a trajectory $(x_*, \theta_0) \in \Omega \times \mathbb{S}^1$ to (x^*, φ), with a reversed trajectory from (x_*, θ_1) to $(x^*, \varphi + \pi)$, where $\theta_0, \theta_1, \varphi \in \mathbb{S}^1$ are arbitrary, see [MD17].

as many radars, each of the form

$$\tilde{\rho}(x,\theta;\, y,r,v) = \frac{1}{1+2\|x-y\|^2}\,\sigma\Big(\frac{\|x-y\|}{r}\Big)\,\sigma\Big(\frac{\langle \mathbb{e}(\theta), x-y\rangle}{v\|x-y\|}\Big). \tag{215}$$

where y is the radar position, $\sigma(s) = 1 - ((1+\cos(2\pi s))/2)^4$ is a function vanishing periodically, r is the ambiguous distance period, and v is the ambiguous radial velocity period. The ambiguous periods r and v are related to the *pulse repetition interval* and *frequency* used by the radar, and their product is bounded below. In this experiment, we choose to optimize the following configuration parameters, gathered into the abstract variable $\xi \in \Xi$: the position of the first radar x_1 within a disk, the position of the second one x_2 within a line, and the blind distances r_1, r_2, r_3, defining the bling velocities as $v_i = 0.2/r_i$ for $1 \leq i \leq 3$.

Minimization over the parameter $\gamma \in \Gamma$ in (214) is solved numerically using the eikonal solver presented in this paper, thus defining a function $\mathcal{E}(\xi) := \inf\{\mathcal{E}(\xi;\gamma);\, \gamma \in \Gamma\}$ depending on the radar configuration alone $\xi \in \Xi$. We differentiate $\mathcal{E}(\xi)$ in an automatic manner as described in [MD17], and optimize this quantity via gradient ascent. Using these tools, a *local* maximum of $\mathcal{E}(\xi)$ is reached in a dozen iterations approximately. Computation time is dominated by the cost of solving a generalized eikonal equation in each iteration, which takes 0.26s on the GPU and 9.6s on the CPU (Dubins model on a $200\times100\times96$ grid). Since the optimization landscape is highly non-convex, obtaining the global maximum w.r.t ξ would require a non-local optimization method in complement or replacement of local gradient ascent, thus requiring many more iterations and benefitting even more from GPU acceleration.

9.4 Conclusion and perspectives

Geodesics and minimal paths are ubiquitous in mathematics, and their efficient numerical computation has countless applications. In this paper, we present a numerical method for computing paths which globally minimize a variety of energies featuring their curvature, by solving a generalized anisotropic eikonal PDE, and which takes advantage of the massive parallelism offered by GPU hardware for computational efficiency. In comparison with previous CPU implementations, a computation time speed up by $30\times$ to $120\times$ is achieved, which enables convenient user interaction in the context of image processing and segmentation, and reasonable run-times for applications such as radar configuration which solve these problems within an inner optimization loop.

Future work will be devoted to additional applications, to efficient implementations of wide stencil schemes associated with other classes of Hamilton-Jacobi-Bellman PDEs, and to the study of numerical schemes based on different compromises in favor of e.g. allowing grid refinement or using shorter finite different offsets.

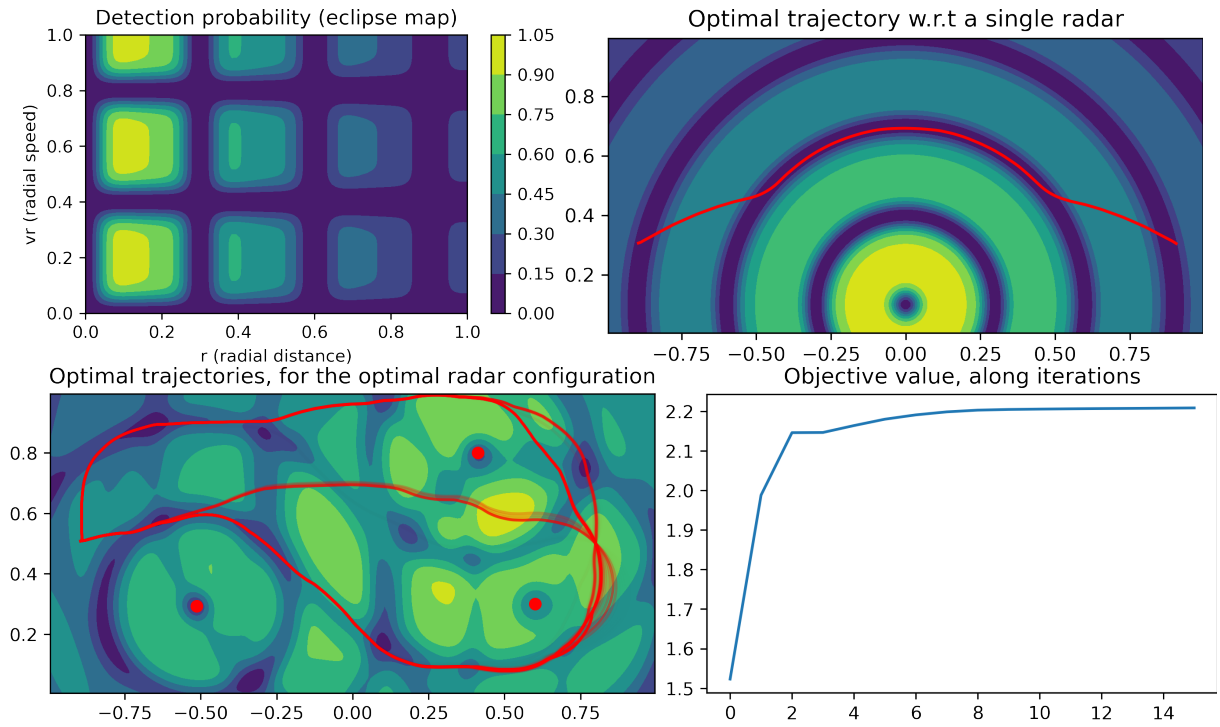

Figure 56: (Top left) Instantaneous detection probability of a vehicle by a radar, depending on the radial distance and radial velocity. Note the blind distance and blind velocity periods. (Top right) Trajectory minimizing the probability of detection between two points in the presence of a single radar. It approximates a concatenation of circles, at a multiple of the blind radial distance, and spirals, corresponding to a multiple of the blind radial velocity. (Bottom left) Configuration of three radars locally optimized, see text, to detect trajectories from the left seed point to the right tip and back. Best adverse trajectories. (Bottom right) Objective value, $\mathcal{E}(\xi)$ see text, along the iterations of gradient ascent.

10

Netted Multi-Function Radars Positioning and Modes Selection by Non-Holonomic Fast Marching Computation of Highest Threatening Trajectories & by CMA-ES Optimization [DDBM19]

This section corresponds to the paper:

- Johann Dreo, François Desquilbet, Frederic Barbaresco, and Jean-Marie Mirebeau. Netted multi-function radars positioning and modes selection by non-holonomic fast marching computation of highest threatening trajectories & by cma-es optimization. In *2019 International Radar Conference (RADAR)*, pages 1–6. IEEE, 2019

Abstract

We aim at designing a radar network which maximizes the detection probability of the worst threatening trajectory which can reach a protected area. In game theory, we represent this problem as a non-cooperative zero-sum game: a first player chooses a setting for the network, and the other player chooses a trajectory from the admissible class of trajectories with full information over the network. The players' objective are respectively to maximize and minimize the path cost which is the detection probability integrated along the trajectory in the network. In comparison with previous works, we added ambiguity maps depending on the distance and the radial speed, which are functions of internal parameters of the radars that can be optimized: PRI (Pulse Repetition Interval) and frequency. We also take into account RCS (Radar Cross Section) and a complex geometry with masks deduced from the DEM (Digital Elevation Map) and Earth curvature. The computation of optimal trajectories is performed by a specialized variant of the Fast-Marching algorithm, devoted to computing curves that globally minimize an energy featuring both a data driven term and a second order curvature penalizing term. The profile of the cost function with regard to the direction of movement is non-convex, which is significant only with a curvature penalization: we chose the Dubins model, in which the curvature radius is bounded. We illustrate results on different Use-Cases.

10.1 Threatening trajectories mitigation for a network of radars

We optimize the configuration of a radar network protecting an area, against an enemy assumed to have unlimited intelligence and computing power, and yet whose vehicle is subject to some maneuverability constraints. The goal is to maximize the probability of detection of the most dangerous trajectory integrated along its path between a given origin and a place to protect, which will take advantage of any hideout in the terrain, blind spot or physical limitation in the radar network. The trajectory is only subject to a lower bound in the turning radius, due to the vehicle high speed. We model this problem as a non-cooperative zero-sum game: a first player chooses a setting ξ for the radar detection network Ξ, and the other player chooses a trajectory γ from the admissible class Γ with

full information over the network. The players' objective is respectively to maximize and minimize the path cost:

$$C(\Xi, \Gamma) = \sup_{\xi \in \Xi} \inf_{\gamma \in \Gamma} \Theta_\xi(\gamma) \tag{216}$$

where $\Theta_\xi(\gamma)$ is the detection probability (integrated along the path) of the trajectory γ in the network ξ. Minimization over $\gamma \in \Gamma$ (given $\xi \in \Xi$) is performed using the fast and reliable techniques of Section 10.2. We rely on the CMA-ES algorithm for the subsequent optimization over $\xi \in \Xi$, which is rather difficult (non-convex, non-differentiable).

In comparison with earlier works [Bar11], we use the curvature bounded Dubins model to reject non-physical enemy trajectories, featuring e.g. angular turns or oscillations in the vehicle direction. We also considerably improve, relative to [MD17], the detection probability model, used to define $\Theta_\xi(\gamma)$, taking into account the three following factors respectively related to the radar, the target, and the terrain.

- The *ambiguity* map accounts for the probability of detection of a generic target by a radar, depending on the distance and the radial speed of the target relatively to the radar, see Figure 57. There are blind speed areas, due to sampling repetition interval and pulse duration causing blind radial distances and blind radial speeds. The positions of the blind areas are periodical and depend on internal parameters of the radar that can be optimized: signal wavelength, and pulse repetition interval.

- The *radar cross section* accounts for the probability of detection of a specific target, depending on its orientation relatively to the radar (Figure 58). For instance, a furtive plane often has a low probability of detection if seen from the front, and a higher one if seen from the side. In our case, we used a simple model with a dependency on frequency, but a more complex model is possible.

- The *elevation map* is used to determine blind regions in the terrain due to obstruction of the radar line of sight. In a mountainous area, a target can take advantage of valleys to move "under the Radar coverage". The Earth curvature is also taken into account.

The profile of the cost function with regard to the direction of movement is typically non-convex, which is significant only in the presence of a curvature penalization. For that, we choose the Dubins model, in which the curvature radius is bounded. We showcase the following three phenomena:

- Trajectories dodging radars through their blind distances (cf. Figure 59). In this picture, only the positional factor in the cost map is shown in greyscale, and not the part of the cost depending on the orientation. The red line represents the optimal trajectory of the target, going from the left to the right of a rectangular domain, with a radar in the center. It features a circle arc, at a precise blind distance from the radar, and two spiral arcs.

- Spiraling threatening trajectories, taking advantage of the blind radial speed (cf. Figure 60). The red line represents the trajectory of the target, going from the left to the center of the domain where the radar is located, maintaining a constant angle

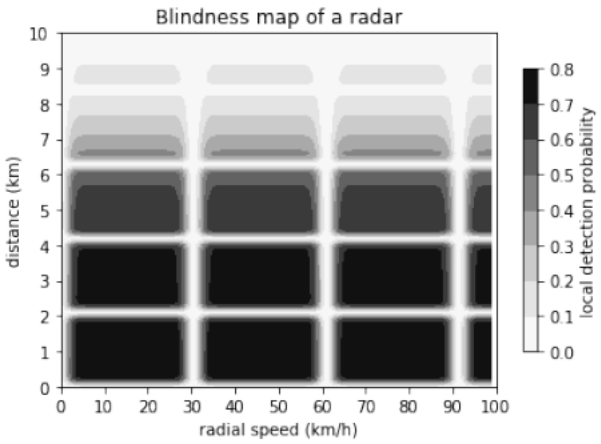

Figure 57: Ambiguity map for a selected waveform waveform

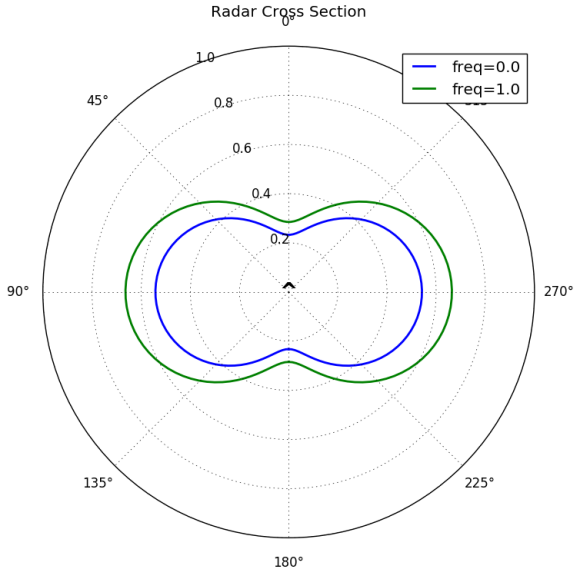

Figure 58: Anisotropy of Radar Cross Section

with the radar in order to minimize visibility, except at the end due to the imposed bound on path curvature. Figure 61 shows the same spiral, along with the zigzag" trajectory that would be optimal if no curvature constraint was taken into account.

- Hiding in valleys. A digital elevation map, of $50km \times 50km$ in a mountain area, is used to construct a probability of detection map (cf. Figure 62). Threatening trajectories tend to concentrate in valleys. The optimized radar positions are close to the target to be defended, and either on high ground or in alignment with long valleys.

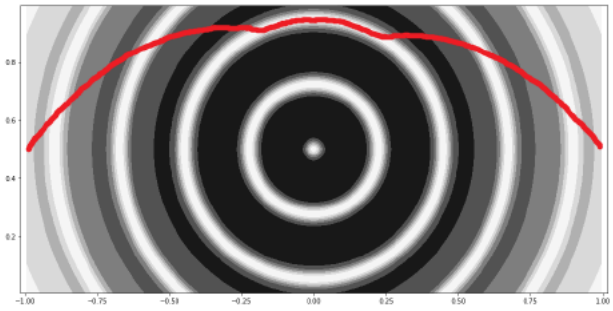

Figure 59: Dodging a radar through a blind distance

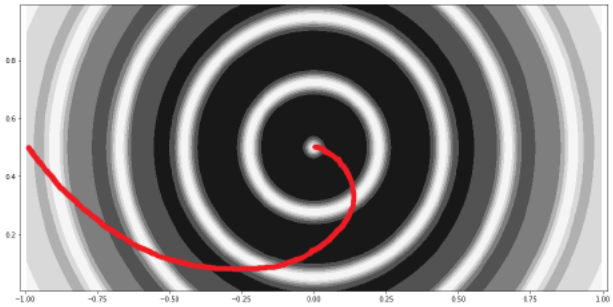

Figure 60: Spiraling threatening trajectory

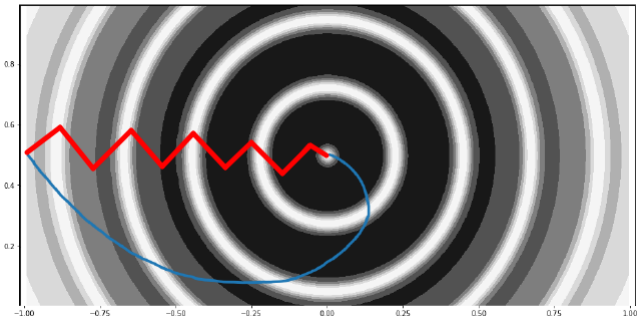

Figure 61: In red, degenerated behaviour with no penalization of curvature, and in blue with curvature constrained solution

10.2 Globally optimal paths with a curvature penalty

This paper deals with planar paths minimizing a specific energy functional, between two given points and with prescribed tangents at these points. The path energy model features a low order data-driven term, and a higher order regularization term. A globally optimal path is found, using optimal control techniques, which involve numerically solving a PDE on the configuration space of positions and orientations.

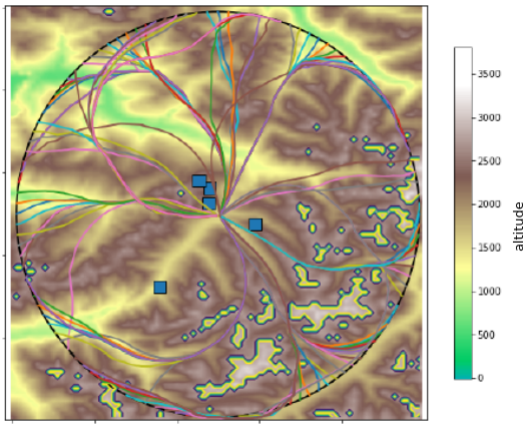

Figure 62: Threatening trajectories, from a circular region towards its center point, with optimized radar positions, and digital elevation map

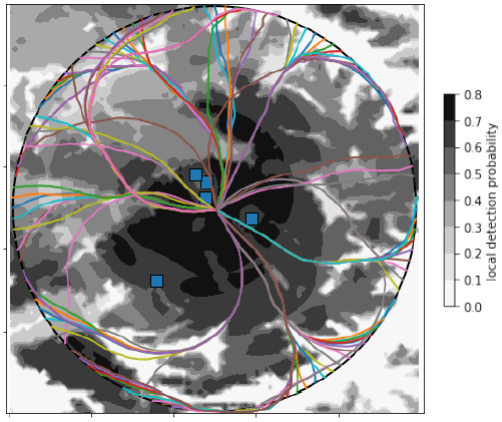

Figure 63: Positional factor in the cost map

10.2.1 Path energy models

In the models of interest to us, the cost of a smooth planar path $x : [0, T] \to \Omega$, parametrized by Euclidean arc length and within a domain $\Omega \subset \mathbb{R}$, takes the following form:

$$\Theta(x) = \int_0^T \rho(x(s), \dot{x}(s)) C(\|\ddot{x}(s)\|) \mathrm{d}s \tag{217}$$

We denoted by $\rho : \overline{\Omega} \times S^1 \to]0, +\infty[$ an arbitrary continuous data-driven term, depending on the path position and direction. An example of definition for ρ has been presented in (215), in a similar setting. The path local curvature $\kappa = \|\ddot{x}(s)\|$ (recall that $\|\dot{x}(s)\| = 1$) is penalized in (216) by a cost function $C(\kappa)$, which may be chosen among the following classical models, here sorted by increasingly stiffness:

- Reeds-Shepp: $\sqrt{1 + \kappa^2}$
- Euler-Mumford: $1 + \kappa^2$

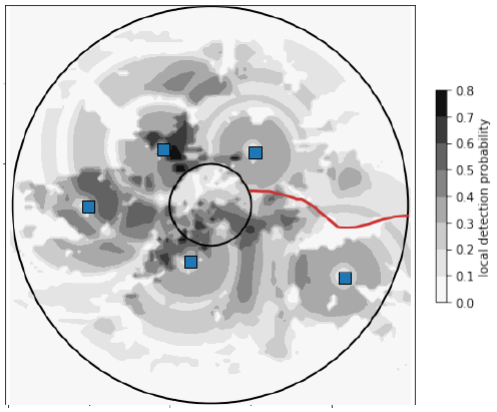

Figure 64: Most threatening trajectories from any point on the limit of coverage to a radial distance from the site to be protected (with cost map in the back ground)

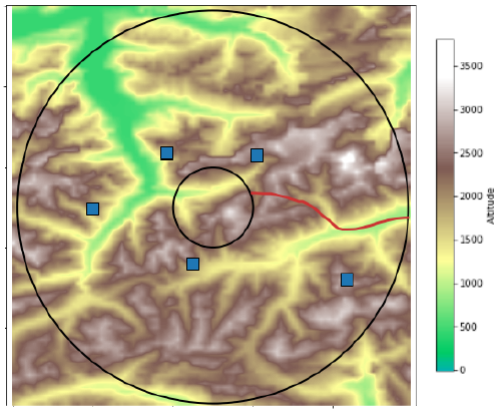

Figure 65: Most threatening trajectories from any point on the limit of coverage to a radial distance from the site to be protected (with DEM in the background)

- Dubins: $\begin{cases} 1 \text{ if } \kappa \leq 1 \\ \infty \text{ else} \end{cases}$

They are respectively representative of (i) a wheelchair-like robot, (ii) the bending energy of an elastic bar, and (iii) a vehicle with a bounded turning radius. In the case of the Reeds-Shepp model, one must further distinguish between the classical model with reverse gear, and the forward only variant.

10.2.2 Viscosity solutions, and the Fast marching algorithm

Data-driven path energies, subject to e.g. fixed endpoints, usually possess many local minima. In order to guarantee that the global minimum is found, path energy minimization must be reformulated as an optimal control problem. The corresponding value function is the unique viscosity solution to a PDE of eikonal type, and the optimal paths can be extracted by backtracking once it is numerically computed.

Only simple first order energies, such as defined by $\int_0^T \rho(x(s))\|\dot{x}(s)\|\mathrm{d}s$ could originally be addressed in the viscosity solution framework, typically using the Fast Marching Method (FMM) which solves the eikonal PDE in a single pass over the domain. Recent progress [Mir18] enabled the *Anisotropic* Fast Marching, in order to solve (217). For that purpose the path is lifted in the configuration space of positions and orientations, defining $\gamma(s) = (x(s), \theta(s))$ subject to the constraint $\dot{x}(s) = (\cos\theta(s), \sin\theta(s))$. This allows to reformulate (217) as a first order energy, since $\|\ddot{x}(s)\| = \|\dot{\theta}(s)\|$. See [Mir18] for details and comparison with alternative approaches.

10.3 Optimization scenario

Digital elevation map is shown in Figure 66, in a domain of $50km \times 50km$. The radar coverage is however complex in the area, due to the high relief from the mountains. The target could fly undetected at a low altitude by hiding in valleys, as in Figure 67. In Figures 67 & 67, the trajectories are computed from a regular sample of the circular boundary of the domain, toward a circular boundary close to the point of interest. The optimized objective function is the smallest probability of detection among all the optimal trajectories reaching the close boundary (cf. Figure 64 and Figure 65), given a configuration of the radar network.

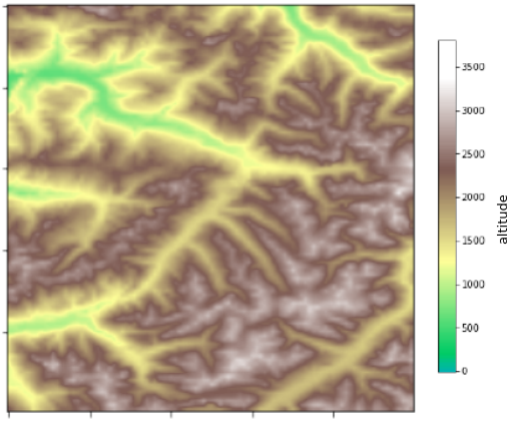

Figure 66: Digital Elevation Map of the Use-Case in Moutain Area

10.4 CMA-ES optimization algorithm

The CMA-ES algorithm is one of the most powerful stochastic numerical optimizers to address difficult black-box problems. Its intrinsic time and space complexity is quadratic limiting its applicability with increasing problem dimensionality. To circumvent this limitation, different large-scale variants of CMA-ES with sub-quadratic complexity have been proposed [VAB+18].

For solving black-box optimization problem of this use-case, we have used the CMA-ES algorithm developed at Paris-Saclay University. The performance of this algorithm on our problem is estimated via ERT-ECDF (expected running time empirical cumulative

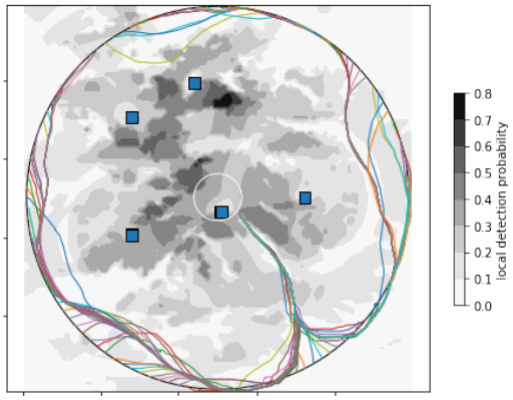

Figure 67: Threatening trajectories in a non-optimized network, with trajectories difficult to detect.

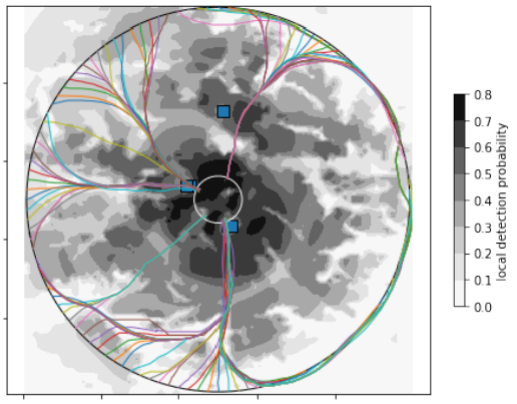

Figure 68: Threatening trajectories in an optimized network, with multiple worst threatening trajectories, but much easier to detect compared with Figure 67

density function): over a large number of runs, calculation of the probability to have a set margin of error, depending on the number of function calls (Figure 69). Difference in threats between a bad configuration and the best one are shown on Figure 67, 68.

For our Use-Case, 20 parameters have been optimized: 4 parameters for each of the radars (position (x,y), Pulse Repetition Interval, wavelength). The trajectory search space has been discretized in the domains: 100 $\times$ 100 for spatial variables, 120 angles for the angular variable. Estimated Computation time for this Use-Case is of a few seconds per function call (largely dominated by the computation of the highest threatening trajectory for a configuration of the radars). The number of function calls needed to find a configuration with a detection probability of 50% in 90% of runs is close to 6 000 (around a day of computation), as shown on Figure 69.

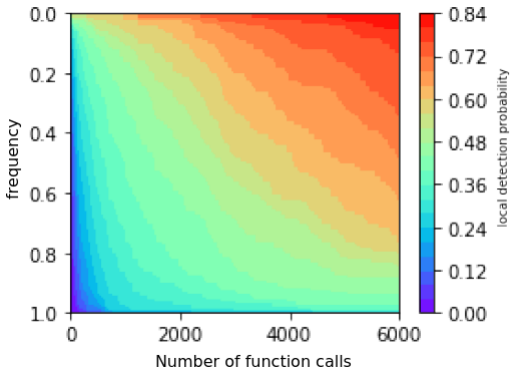

Figure 69: ERT for all scenarios with 150 runs of CMA-ES algorithm.

10.5 Conclusion

We have proved robustness of new algorithms for optimal configuration of netted radars, to reduce integrated probability of detection along the highest threatening trajectory. We have illustrated with realistic modelling of generic anistropic cost functions: ambiguity map, radar cross-section, digital elevation map... AFM (Anisotropic Fast Marching) algorithm can take into account a penalization of curvature in the computation of shortest path, which is critical in that setting. Other uses cases for the AFM algorithm could be elaborated:

- configuration of a radar who lost a target,
- use of mobile radars,
- adding passive radars to cover the defects of the network.

Future works will be devoted to further enhancing the model, taking into account limited knowledge of the ennemy (e.g. due to the use of passive radar receivers), introducing success criteria more complex than mere detection (e.g. requiring detection early enough for interception), and adapting optimization solvers to the problem in order to reach better performances.

References

[ABK97] F. Aminzadeh, J. Brac, and T. Kunz. *3-D Salt and Overthrust models*. SEG/EAGE 3-D Modeling Series No.1, 1997.

[AM12] Ken Alton and Ian M Mitchell. An ordered upwind method with precomputed stencil and monotone node acceptance for solving static convex Hamilton-Jacobi equations. *Journal of Scientific Computing*, 51(2):313–348, 2012.

[Bar11] Frédéric Barbaresco. Computation of most threatening radar trajectories areas and corridors based on fast-marching & level sets. In *2011 IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA)*, pages 51–58. IEEE, 2011.

[BBM20] Joseph Bonnans, Guillaume Bonnet, and Jean-Marie Mirebeau. Monotone and second order consistent scheme for the two dimensional Pucci equation. 2020.

[BBM21a] Frédéric Bonnans, Guillaume Bonnet, and Jean-Marie Mirebeau. A linear finite-difference scheme for approximating Randers distances on Cartesian grids. 2021.

[BBM21b] J Frederic Bonnans, Guillaume Bonnet, and Jean-Marie Mirebeau. Second order monotone finite differences discretization of linear anisotropic differential operators. *Mathematics of computation*, 90(332):2671–2703, 2021.

[BC91] Vladislav Babuska and Michel Cara. *Seismic anisotropy in the Earth*, volume 10. Springer Science and Business Media, 1991.

[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.

[Bey87] G. Beylkin. Mathematical theory for seismic migration and spatial resolution. In M. Bernabini, P. Carrion, G. Jacovitti, F. Rocca, S. Treitel, and M. Worthington, editors, *Deconvolution and inversion.*, pages 291–304. Blackwell scientific publications (Oxford), 1987.

[BL98] F. Billette and G. Lambaré. Velocity macro-model estimation from seismic reflection data by stereotomography. *Geophysical Journal International*, 135(2):671–680, 1998.

[BL10] Jonathan Borwein and Adrian S Lewis. *Convex analysis and nonlinear optimization: theory and examples*. Springer Science and Business Media, 2010.

[Ble87] N. Bleistein. On the imaging of reflectors in the Earth. *Geophysics*, 52(7):931–942, 1987.

[BLP11] A. Bensoussan, J. L. Lions, and G. Papanicolaou. *Asymptotic analysis for periodic structures*, volume 374. American Mathematical Soc., 2011.

[BLZ10] J. D. Benamou, S. Luo, and H. Zhao. A compact upwind second order scheme for the eikonal equation. *Journal of Computational Mathematics*, pages 489–516, 2010.

[BM21] Guillaume Bonnet and Jean-Marie Mirebeau. Monotone discretization of the monge-ampère equation of optimal transport. 2021.

[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.

[Bru95] Heinrich Bruns. *Das eikonal*, volume 35. S. Hirzel, 1895.

[BST83] Iulian Beju, Eugen Soós, and Petre P Teodorescu. *Euclidean tensor calculus with applications*. CRC Press, 1983.

[BT13] Efstathios Bakolas and Panagiotis Tsiotras. Optimal synthesis of the Zermelo–Markov–Dubins problem in a constant drift field. *Journal of Optimization Theory and Applications*, 156(2):469–492, 2013.

[Car01] Michael Carter. *Foundations of mathematical economics*. MIT Press, 2001.

[CBM20] J. Cao, R. Brossier, and L. Métivier. 3d acoustic-(visco) elastic coupled formulation and its spectral-element implementation on a cartesian-based hexahedral mesh. In *SEG Technical Program Expanded Abstracts 2020*, pages 2643–2647. Society of Exploration Geophysicists, 2020.

[CC18] P. Cupillard and Y. Capdeville. Non-periodic homogenization of 3-d elastic media for the seismic wave equation. *Geophysical Journal International*, 213(2):983–1001, 01 2018.

[CCV13] Zachary Clawson, Adam Chacon, and Alexander Boris Vladimirsky. Causal Domain Restriction for Eikonal Equations. *arXiv.org*, September 2013.

[CEL84] Michael G Crandall, Lawrence C Evans, and P-L Lions. Some properties of viscosity solutions of hamilton-jacobi equations. *Transactions of the American Mathematical Society*, 282(2):487–502, 1984.

[Cer05] V. Cerveny. *Seismic ray theory*. Cambridge university press, 2005.

[CHK13] Marcel Campen, Martin Heistermann, and Leif Kobbelt. Practical Anisotropic Geodesy. *Computer Graphics Forum*, 32(5):63–71, August 2013.

[CM18] Y. Capdeville and L. Métivier. Elastic full waveform inversion based on the homogenization method: theoretical framework and 2-d numerical illustrations. *Geophysical Journal International*, 213(2):1093–1112, 2018.

[CMA+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. A New Finsler Minimal Path Model With Curvature Penalization for Image Segmentation and Closed Contour Detection. *Computer Vision and Pattern Recognition (CVPR)*, pages 355–363, June 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.

[Cri09] E. Cristiani. A fast marching method for Hamilton-Jacobi equations modeling monotone front propagations. *Journal of Scientific Computing*, 39(2):189–205, 2009.

[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.

[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.

[D+59] Edsger W Dijkstra et al. A note on two problems in connexion with graphs. *Numerische mathematik*, 1(1):269–271, 1959.

[DCC+21] François Desquilbet, Jian Cao, Paul Cupillard, Ludovic Métivier, and Jean-Marie Mirebeau. Single pass computation of first seismic wave travel time in three dimensional heterogeneous media with general anisotropy. *Journal of Scientific Computing*, 89(1):1–37, 2021.

[DDBM19] Johann Dreo, François Desquilbet, Frederic Barbaresco, and Jean-Marie Mirebeau. Netted multi-function radars positioning and modes selection by non-holonomic fast marching computation of highest threatening trajectories & by cma-es optimization. In *2019 International Radar Conference (RADAR)*, pages 1–6. IEEE, 2019.

[DeV98] Ronald A DeVore. Nonlinear approximation. *Acta Numerica*, 7:51–150, 1998.

[DMM22] François Desquilbet, Jean-Marie Mirebeau, and Ludovic Métivier. Single pass eikonal solver in tilted transversely anisotropic media. 2022.

[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.

[DS97] J. Dellinger and W. Symes. Anisotropic finite-difference traveltimes using a hamilton-jacobi solver. In *SEG Technical Program Expanded Abstracts 1997*, pages 1786–1789. Society of Exploration Geophysicists, 1997.

[Eva10] Lawrence C Evans. *Partial Differential Equations*. American Mathematical Soc., 2010.

[FKW13] Zhisong Fu, Robert M Kirby, and Ross T Whitaker. A fast iterative method for solving the eikonal equation on tetrahedral domains. *SIAM Journal on Scientific Computing*, 35(5):C473–C494, 2013.

[FLT03] M. Fukushima, Z. Q. Luo, and P. Tseng. A sequential quadratically constrained quadratic programming method for differentiable convex minimization. *SIAM Journal on Optimization*, 13(4):1098–1119, 2003.

[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.

[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.

[Hes64] H. H. Hess. Seismic anisotropy of the uppermost mantle under oceans. *Nature*, 203(4945):629–631, 1964.

[HF07] M Sabry Hassouna and Aly A Farag. Multistencils fast marching methods: A highly accurate solution to the eikonal equation on cartesian domains. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(9):1563–1574, 2007.

[HJ16] Sumin Hong and Won-Ki Jeong. A multi-gpu fast iterative method for eikonal equations using on-the-fly adaptive domain decomposition. *Procedia Computer Science*, 80:190–200, 2016.

[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.

[KC99] S. Kim and R. Cook. 3d traveltime computation using second order eno scheme. *Geophysics*, 64:1867–1876, 1999.

[Kim99] S. Kim. On eikonal solvers for anisotropic traveltimes. In *SEG Technical Program Expanded Abstracts 1999*, pages 1875–1878. Society of Exploration Geophysicists, 1999.

[Knu98] Donald E Knuth. *The art of computer programming: Volume 3: Sorting and Searching*. Addison-Wesley Professional, 1998.

[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] Ron Kimmel and James A. Sethian. Computing geodesic paths on manifolds. *Proceedings of the National Academy of Sciences*, 95(15):8431–8435, July 1998.

[LBLM] P. Le Bouteiller, M. Benjemaa, and volume=84 number=2 pages=C107-C118 year=2019 publisher=Society of Exploration Geophysicists L. Métivier and J. Virieux, journal=Geophysics. A discontinuous galerkin fast-sweeping eikonal solver for fast and accurate traveltime computation in 3d tilted anisotropic media.

[LBMV18] Philippe Le Bouteiller, Mondher Benjemaa, Ludovic Métivier, and Jean 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, 2018.

[LCZ14] H. Lan, J. Chen, and Z. Zhang. A fast sweeping scheme for calculating p wave first-arrival travel times in transversely isotropic media with an irregular surface. *Pure and Applied Geophysics*, 171(9):2199–2208, 2014.

[Lec93] I. Lecomte. Finite difference calculation of first traveltimes in anisotropic media. *Geophysical Journal International*, 113(2):318–342, 1993.

[LFH11] Peter G. Lelièvre, Colin G. Farquharson, and Charles A. Hurich. Computing first-arrival seismic traveltimes on unstructured 3-D tetrahedral grids using the fast marching method. *Geophysical Journal International*, 184:885–896, 2011.

[LOP+03] G. Lambaré, S. Operto, P. Podvin, P. Thierry, and M. Noble. 3-D ray+Born migration/inversion - part 1: theory. *Geophysics*, 68:1348–1356, 2003.

[LQ12] Songting Luo and Jianliang Qian. Fast sweeping methods for factored anisotropic eikonal equations: multiplicative and additive factors. *Journal of Scientific Computing*, 52(2):360–382, 2012.

[LRW13] W. Liao, K. Rohr, and S. Wörz. Globally Optimal Curvature-Regularized Fast Marching For Vessel Segmentation. *Medical Image Computing and Computer-Assisted Interventation-MICCAI 2013. Springer Berlin Heidelberg.*, pages 550–557, 2013.

[LVF13] Siwei Li, Alexander Vladimirsky, and Sergey Fomel. First-break traveltime tomography with the double-square-root eikonal equationdsr tomography. *Geophysics*, 78(6):U89–U101, 2013.

[MD17] Jean-Marie Mirebeau and Johann Dreo. 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.

[MD20] J. M. Mirebeau and F. Desquilbet. Worst case and average case cardinality of strictly acute stencils for two dimensional anisotropic fast marching. In *Constructive Theory of Functions - 2019*, pages 157–180. Publishing House of Bulgarian Academy of Sciences, 2020.

[MGB+21] Jean-Marie Mirebeau, Lionel Gayraud, Rémi Barrère, Da Chen, and Francois Desquilbet. Massively parallel computation of globally optimal shortest paths with curvature penalization. 2021.

[Mir14a] Jean-Marie Mirebeau. Anisotropic fast-marching on cartesian grids using lattice basis reduction. *SIAM Journal on Numerical Analysis*, 52(4):1573–1599, 2014.

[Mir14b] Jean-Marie Mirebeau. Efficient fast marching with Finsler metrics. *Numerische Mathematik*, 126(3):515–557, 2014.

[Mir16] Jean-Marie Mirebeau. Adaptive, anisotropic and hierarchical cones of discrete convex functions. *Numerische Mathematik*, 132(4):807–853, 2016.

[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.

[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.

[Mus03] M. J. MP. Musgrave. *Crystal acoustics*. Acoustical Society of America New York, 2003.

[MvN92] T. J. Moser, T. van Eck, and G. Nolet. Hypocenter determination in strongly heterogeneous earth models using the shortest path method. *Journal Geophysical Research*, 97:6563–6572, 1992.

[Niq07] Milad Niqui. Exact arithmetic on the Stern–Brocot tree. *Journal of Discrete Algorithms*, 5(2):356–379, June 2007.

[Nol08] G. Nolet. *A Breviary of Seismic Tomography*. Cambridge University Press, Cambridge, UK, 2008.

[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.

[OS91] S. Osher and C. W. Shu. High-order essentially nonoscillatory schemes for hamilton-jacobi equations. *SIAM Journal on numerical analysis*, 28(4):907–922, 1991.

[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.

[PTVF07] W H Press, S A Teukolsky, W T Vetterling, and B P Flannery. Numerical recipes: the art of scientific computing. 3rd. 2007.

[PWZ17] A. Padh, M. Willis, and X. Zhao. Accurate quasi-p traveltimes in 3d transversely isotropic media using a high-order fast-sweeping-based eikonal solver. In *SEG Technical Program Expanded Abstracts 2017*, pages 369–373. Society of Exploration Geophysicists, 2017.

[QZZ07] J. Qian, Y. T. Zhang, and H. K. Zhao. A fast sweeping method for static convex hamilton-jacobi equations. *Journal of Scientific Computing*, 31(1-2):237–271, 2007.

[RS04] N. Rawlinson and M. Sambridge. Multiple reflection and transmission phases in complex layered media using a multistage fast marching method. *Geophysics*, 69(5):1338–1350, 2004.

[RS09] Christian Rasch and Thomas Satzger. Remarks on the $\mathcal{O}(N)$ implementation of the fast marching method. *IMA Journal of Numerical Analysis*, 29(3):806–813, 2009.

[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.

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

[Set96] J. A. Sethian. A fast marching level set method for monotonically advancing fronts. *Proceedings of the National Academy of Sciences of the United States of America*, 93:1591–1595, 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.

[SKD+07] Maxime Sermesant, Ender Konukoglu, Hervé Delingette, Yves Coudiere, Phani Chinchapatnam, Kawal S Rhode, Reza Razavi, and Nicholas Ayache. An anisotropic multi-front fast marching method for real-time simulation of cardiac electrophysiology. In *International Conference on Functional Imaging and Modeling of the Heart*, pages 160–169. Springer, 2007.

[Sla03] M. A. Slawinski. *Seismic Waves and Rays in Elastic Media*. Elsevier Science, 2003.

[SV01] J. A. Sethian and A. B. Vladimirsky. Ordered upwind methods for static Hamilton-Jacobi equations. *Proceedings of the National Academy of Sciences of the United States of America*, 98(20):11069–11074 (electronic), 2001.

[SV03] James A. Sethian and Alexander Boris Vladimirsky. Ordered upwind methods for static Hamilton-Jacobi equations: theory and algorithms. *SIAM Journal on Numerical Analysis*, 41(1):325–363, 2003.

[SW09] Seth Stein and Michael Wysession. *An introduction to seismology, earthquakes, and earth structure*. John Wiley & Sons, 2009.

[TBM19a] Julien Thurin, Romain Brossier, and Ludovic Métivier. Ensemble-based uncertainty estimation in full waveform inversion. *Geophysical Journal International*, 219(3):1613–1635, 2019.

[TBM+19b] P. T. Trinh, R. Brossier, L. Métivier, L. Tavard, and J. Virieux. Efficient 3D time-domain elastic and viscoelastic Full Waveform Inversion using a spectral-element method on flexible Cartesian-based mesh. *Geophysics*, 84(1):R75–R97, 2019.

[TCOZ03] Y. H. R. Tsai, L. T. Cheng, S. Osher, and H. K. Zhao. Fast sweeping algorithms for a class of hamilton-jacobi equations. *SIAM journal on numerical analysis*, 41(2):673–694, 2003.

[TH16] E. Treister and E. Haber. A fast marching algorithm for the factored eikonal equation. *Journal of Computational Physics*, 324:210–225, 2016.

[Tho86] L. Thomsen. Weak elastic anisotropy. *Geophysics*, 51(10):1954–1966, 1986.

[TNCC09] C. Taillandier, M. Noble, H. Chauris, and H. Calandra. First-arrival travel time tomography based on the adjoint state method. *Geophysics*, 74(6):WCB1–WCB10, 2009.

[Tsi95] J. N. Tsitsiklis. Efficient algorithms for globally optimal trajectories. *IEEE transactions on Automatic Control*, 40(9):1528–1538, September 1995.

[VAB+18] Konstantinos Varelas, Anne Auger, Dimo Brockhoff, Nikolaus Hansen, Ouassim Ait ElHara, Yann Semet, Rami Kassab, and Frédéric Barbaresco. A comparative study of large-scale variants of cma-es. In *International conference on parallel problem solving from nature*, pages 3–15. Springer, 2018.

[Vid88] J. Vidale. Finite-difference calculation of travel times. *Bulletin of the Seismological Society of America*, 78(6):2062–2076, 1988.

[Vla08] Alexander Boris Vladimirsky. Label-setting methods for Multimode Stochastic Shortest Path problems on graphs. *Mathematics of Operations Research*, 33(4):821–838, 2008.

[VN18] D. W. Vasco and K. Nihei. Broad-band trajectory mechanics. *Geophysical Journal International*, 216(2):745–759, 2018.

[Wah20] U. B. Waheed. A fast-marching eikonal solver for tilted transversely isotropic media. *Geophysics*, 85(6):S385–S393, 2020.

[WDB+08] Ofir Weber, Yohai S Devir, Alexander M Bronstein, Michael M Bronstein, and Ron Kimmel. Parallel algorithms for approximation of distance maps on parametric surfaces. *ACM Transactions on Graphics (TOG)*, 27(4):104–16, October 2008.

[WFNBZ20] Malcolm CA White, Hongjian Fang, Nori Nakata, and Yehuda Ben-Zion. PyKonal: a Python package for solving the eikonal equation in spherical and Cartesian coordinates using the fast marching method. *Seismological Research Letters*, 91(4):2378–2389, 2020.

[WYF15] U. B. Waheed, C. E. Yarman, and G. Flagg. An iterative, fast-sweeping-based eikonal solver for 3d tilted anisotropic media. *Geophysics*, 80(3):C49–C58, 2015.

[Zha05] H. Zhao. A fast sweeping method for eikonal equations. *Mathematics of computation*, 74:603–627, 2005.