Entropic Unbalanced Optimal Transport: Application to Full-Waveform Inversion and Numerical Illustration

Miao Yu

Thèse préparée à l'institut de physique du globe de Paris
Université de Paris

École doctorale Sciences de la Terre et de l'environnement
et physique de l'Univers STEP'UP n°560

Sismologie

Entropic Unbalanced Optimal Transport :

Application to Full-Waveform Inversion and Numerical Illustration

par Miao Yu

Thèse de doctorat de Sciences de la Terre et de l'environnement

- dirigée par

  Jean-Pierre Vilotte et Jean-David Benamou

- présentée et soutenue publiquement le

  17 décembre 2021

- devant un jury composé de

  Ludovic, Métivier, Chargé de Recherche, LJK, CNRS, Université Grenoble Alpes, Rapporteur
  Hérvé Chauris, Professeur Université, Mines Paris-Tech, Université Paris Sorbonne Lettres, Rapporteur
  Géraldine, Pichot, Chargé de Recherche, INRIA Paris, Examinateur
  Anne, Mangeney, Professeure Université, IPGP, Université de Paris, Examinateur
  Jean-David, Benamou, Directeur de Recherche, INRIA Paris, co-Directeur de Thèse
  Jean-Pierre, Vilotte, Physicien des Observatoires, IPGP, Université de Paris, Directeur de Thèse

Entropic Unbalanced Optimal Transport :

Application to Full-Waveform Inversion and Numerical Illustration

par Miao Yu

Résumé

Les méthodes de tomographie sismique visent à inférer les propriétés physique et reconstruire le "modèle", i.e. les structures de l'intérieur de la Terre, à partir des ondes mécaniques - radiées par des sources naturelles ou anthropogéniques - enregistrées par des récepteurs en surface sous la forme de sismogrammes. Les méthodes d'inversion de formes d'onde ont été activement développées dans les contextes académiques et industriels et sont devenus des outils puissants pour améliorer l'estimation des propriétés physiques et des structures d'objets géologiques depuis les échelles globales jusqu'aux échelles locales de la géophysique d'exploration.

L'inversion de formes d'onde est formulée comme un problème d'optimisation non linéaire, associé à un système d'équations aux dérivées partielles. Il est classiquement résolu par des méthodes d'optimisation locale via la minimisation itérative d'une fonction coût qui mesure la différence entre les sismogrammes observés et synthétiques, et utilisent des méthodes d'état adjoint. Les méthodes d'état adjoint permettent le calcul des dérivées de la fonction coût par rapport aux paramètres du modèle en combinant le champ d'onde direct et un champ d'onde adjoint gouverné par un système d'équations adjointes et des conditions adjointes complémentaires.

Les méthodes d'inversion de forme d'onde, qui inversent simultanément les courtes et grandes longueurs d'onde, souffrent malheureusement en pratique de difficultés qui restreignent leur utilisation pratique. Leurs capacités se détériorent du fait du déficit en basse fréquence des observations et d'un bon modèle initial. Des limitations qui sont associées à la nature mal posée du problème inverse qui peut facilement être piégé dans un minimum local.

Une direction proposée, afin de réduire la dépendance vis à vis du modèle initial, est de remplacer la fonction classique, basée sur une distance de type moindres carrés, par de nouvelles fonctions coûts, pouvant impliquer une transformation non linéaire du signal, afin de promouvoir la convexité et élargir le bassin d'attraction du minimum global.

La théorie du Transport Optimal (OT) a récemment été utilisée dans le cadre des problèmes inverses et de l'apprentissage automatique. Le transport optimal généralise les propriétés de la distance euclidienne au carré à l'espace des distributions de probabilité. La valeur optimale (au carré) du transport lui-même définit une distance appelée distance 2-Wasserstein. Cette quantité est à nouveau convexe mais maintenant sur l'ensemble des distributions de probabilité.

Le Transport Optimal est déjà utilisé en FWI, cette thèse en fait partie. L'approche OT est encore largement ouverte sur trois fronts : les formes d'onde sismiques ne sont ni positives ni de masse totale normalisée. La convexité par rapport au modèle n'est pas garantie et finalement le calcul réel de la distance OT est coûteuse.

Dans ce travail, nous utilisons et combinons - d'un point de vue académique - deux extensions récentes d'OT dans le contexte FWI. D'abord la distance OT "non-équilibrée", qui définit rigoureusement une distance sur l'ensemble des mesures de Radon positives contournant ainsi le problème de normalisation des données (mais pas le problème de positivité). Puis le cadre du Transport Optimal entropique et en particulier la variante simple et facile à calculer appelée divergence de Sinkhorn fournissant une bonne approximation de la distance 2-Wasserstein. La divergence de Sinkhorn peut être naturellement étendue au transport "non-équilibré".

Nous utilisons ces outils pour construire et mettre en œuvre une fonction coût OT "non-équilibrée". Nous discutons de son utilisation dans le contexte FWI au travers d'un certain nombre d'exemples académiques et de problèmes de référence classiques.

Mots-clés

Transport-optimal entropique non-équilibré, divergence de Sinkhorn, inversion de formes d'onde,

Abstract

Seismic tomography aims at inferring physical properties and reconstructing quantitatively the "model", i.e. structures of the Earth interior, from the mechanical waves – radiated by natural and man-made seismic sources – that are recorded at the surface by receivers in the form of seismograms. Over the past decades, Full-Waveform inversion (FWI) has been actively developed in both academia and industry, and has proven to be a powerful tool that dramatically improved the capability to estimate physical properties and structures of various geological targets from global to local scales.

Full-waveform inversion (FWI) is formulated as a nonlinear, PDE-based optimisation problem that is classically solved by iterative minimisation of an objective function – measuring the misfit between synthetic and observed seismic waveforms – using adjoint-based solution methods. Adjoint-based solution methods allow the computation of the derivative of the objective function with respect to the model parameters by combining the synthetic forward wavefield and an adjoint wavefield governed by a set of adjoint equations and adjoint subsidiary conditions.

In practice, however, solving FWI problems using local, nonlinear optimisation methods is facing challenges that preclude routine use. The quality of the inversion, simultaneously dealing with long and short wavelengths information, is degraded by the lack of low frequencies and also depends on a good starting model. These limitations are linked to the ill-posed nature of the inverse problem which can be easily trapped into a local minimum.

One proposed research direction to reduce the dependency on the initial model, is to replace the classical least-squares based misfit by other objective functions – possibly involving nonlinear transformation of the seismic signal –hence promoting the convexity and trying to enlarge the basin of attraction of the global minimum.

Optimal Transport (OT) theory has recently been use in inverse problems and machine learning. Optimal Transport lifts the properties of the squared Euclidean distance to the space of probability distributions. The optimal value (squared) of the transport itself defines a distance called the 2-Wasserstein distance. This quantity is again convex but now on the set of probability distributions.

This trend is already active for FWI, this thesis is part of it. The OT approach is still largely open on three fronts : Seismic Waveforms are not probability distributions, lacking positivity and normalised total mass. Convexity with respect to the model is not guaranteed and finally the actual computation of the OT distance is not cheap.

In this work we use and combine – from an academic point of view – two recent extensions of OT in the context of FWI. First the "unbalanced" OT distance, which rigorously defines a distance on the set of positive Radon measure thus by-passing the data normalisation issue (but not the positivity problem). Then, the entropic regularisation OT framework and in particular the simple and easy to compute variant called Sinkhorn divergence providing a good approximation of the 2-Wasserstein distance. The Sinkhorn divergence can be naturally extended to unbalanced OT.

We use these tools to construct and implement our unbalanced OT misfit and discuss its use in the context of full-waveform inversion through a number of academic examples and classical benchmark problems.

Keywords

Unbalanced entropic optimal transport, Sinkhorn divergence, full-waveform inversion

IPGP
INSTITUT DE PHYSIQUE DU GLOBE DE PARIS

Université de Paris

To my loving family.

Contents

List of Figures

Résumé en Français

La tomographie sismique vise à déduire des propriétés physiques et à reconstruire quantitativement des structures de l'intérieur de la Terre à partir des ondes mécaniques qui sont rayonnées par des sources sismiques naturelles ou artificielles, et enregistrées à la surface par des récepteurs sous la forme de sismogrammes, également appelés traces sismiques. Il s'agit d'un problème fondamental en géophysique avec des applications sociétales et industrielles telles que la découverte et l'exploitation de ressources énergétiques, l'investigation de sites pour le génie civil, les gisements minéraux et l'approvisionnement en eau souterraine, l'aléa sismique et l'atténuation des risques.

En sismologie et en géophysique d'exploration le développement de nouveau systèmes d'acquisition et de l'utilisation de nouvelles sources (naturelles et/ou artificielles) ont favorisé des méthodes d'imagerie sismique haute résolution exploitant des quantités et une diversité de données de plus en plus importantes. Cela a conduit à des modèles de profondeur de résolution de plus en plus élevée à des échelles locales, allant de dizaines de mètres à des dizaines de kilomètres, selon l'énergie de la source et la géométrie du système d'acquisition.

Les interactions entre les ondes et le milieu géologique dépendent : de la bande de fréquence, de la durée des sources sismiques et de leur énergie rayonnée ; des échelles d'hétérogénéité caractéristiques du milieu ; et du trajet de propagation de l'onde entre la source et les récepteurs. L'extraction d'information à partir des sismogrammes enregistrés n'est pas une simple tâche du fait que ces interactions sont intégrées le long de la distance de propagation.

La méthode d'inversion de forme d'onde (FWI), ainsi que la méthode élégante et constructive de l'état adjoint, inverse simultanément les courtes et grandes longueurs d'onde du signal et ainsi de simplifier le flux de travail en imagerie sismique classique impliquant une étape de tomographie suivie d'une étape de migration. la méthode FWI est formulé comme un problème d'optimisation non linéaire dans lequel l'état du systèmes est caractérisé par le champ d'ondes dont la réalisation physique est régie, sous certaines approximations, par un systèmes de PDEs pour une paramétrisation donnée du modèle.

Les méthodes de l'état adjoint permettent le calcul du gradient d'une fonction coût - mesurant la différence entre les observations et les observables prédites - par rapport aux paramètres du modèle en combinant le champ d'onde incident prédit par les équations d'onde, ainsi que les valeurs initiales et les conditions aux limites associées, et le champ d'onde adjoint régi par un ensemble d'équations adjointes, ainsi que les valeurs finales et conditions aux limites adjointes associées. Au cours des dernières décennies, FWI a été activement étudié dans le monde académique et industriel, tirant parti de : (1) le développement de nouvelles méthodes

de modélisation physique et numérique pour la propagation du champ d'ondes 3D - de l'acoustique au viscoélastique - dans des milieux hétérogènes et complexes ; (2) l'augmentation rapide des capacités du calcul haute performance ; (3) la quantité et la couverture croissantes des données sismiques large bande.

Dans la pratique, cependant, la résolution des problèmes FWI à l'aide de méthodes d'optimisation locales non linéaires est confrontée à des défis qui limitent son utilisation en production: (1) la physique des ondes sismiques est complexe et la fonction coût est peu sensible aux composants du modèle dont les nombres d'onde sont plus petits que les longueurs d'ondes sismiques résolus par l'opérateur de propagation; (2) l'absence de basses fréquences et le bruit dans les observations, ainsi qu'une mauvaise connaissance du modèle initial conduisent à un problème d'inversion mal posé et détériorent sa précision.

Les données sismiques sont par nature des signaux oscillatoires, dont la composant à fréquence nulle n'est pas enregistrée. Le problème d'optimisation étant hautement non linéaire avec une fonction coût multimodale, un mauvais modèle cinématique initial peut piéger la solution dans des minima locaux. Dans la forme standard de FWI, la fonction coût est basée sur la distance des moindres carrés, qui mesure des différences locales en intensité, et est oscillante et non convexe dans les directions associées aux composantes du modèle dont les nombres d'onde sont plus longues que la moitié de la longueur d'onde dominante. Pour des erreurs cinématiques importantes, les algorithmes d'optimisation locale feront correspondre des phases incorrectes entre les signaux prédits et observés, rendant la méthode sujette à ce que l'on appelle l'effet de "cycle skipping". L'inversion converge vers un minimum local, éventuellement éloigné du minimum global, du fait des erreurs cinématiques introduites par le mauvais modèle de départ.

La thèse fournit un rapide aperçu des fondements de ma méthode FWI, et de l'état adjoint, avec une illustration dans le cade de l'approximation acoustique scalaire, ainsi qu'un revue des différentes méthodes proposées pour prévenir les problèmes associés à cette méthodes, dans l'espace des données et/ou des images.

Cette thèse est une contribution à une nouvelle approche qui se développe activement aujourd'hui pour la formulation de nouvelles fonctions coût des les problèmes inverses (et l'apprentissage statistique) basés sur les distances de transport optimal (OT). Les distances OT transforment la mesure intensive locale des moindres carrés en une mesure globale, sensible a tous les défauts de correspondance en espace, temps et amplitude. Les distances basées sur l'OT sont de nature lagrangienne et en particulier convexes par rapport aux translations et aux dilatations. Cette propriété en fait de bons candidats pour résoudre le problème du "cycle skipping".

A ce jour, plusieurs familles de fonctions coût basées sur l'OT ont été proposées :

- Une première famille, proposée à l'origine par Engquist et ses collaborateurs est liée à la métrique 2-Wasserstein. Elle est limitée à des mesures de probabilité. Les signaux sismiques nécessitent alors une transformation et une normalisation ad-hoc pour être comparés à l'aide de cette métrique. Cela peut altérer les informations de phase et d'amplitude, et favoriser la non-convexité par rapport aux décalages temporels. Le calcul de la métrique 2-Wasserstein en utilisant la formulation de Monge-Ampère soit avec des différences finies optimisées, des solveurs linéaires, ou une stratégie semi-discrète ajoute un coût de calcul significatif à le FWI. Dans la pratique l'application de la distance 2-Wasserstein a souvent

été restreinte au transport 1D-optimal, c'est-à-dire à la comparaison trace par trace moins couteuse.

- Une deuxième famille, proposée par Métivier et collaborateurs fait appel à la métrique 1-Wasserstein, instance particulière de la métrique de Kantorowitch-Rubinstein, associée à la formulation duale du problème de transport optimal. Cette métrique OT est désormais définie pour des mesures signées et peut être utilisée pour gérer un transport optimal déséquilibré, évitant la transformation et la normalisation ad hoc du signal. Cela conduit à un problème d'optimisation convexe sous contraintes linéaires qui peut être résolu en utilisant des méthodes d'optimisation convexe non lisses qui permettent de réduire significativement le coût de calcul par rapport à la métrique 2-Wasserstein. La convexité de la fonction objectif n'est cependant pas garantie vis-à-vis des transformations en temps et en amplitude.

- Une troisième famille récemment introduite par Métivier et collaborateurs est liée à une distance de transport optimale dans l'espace étendu du graphe des données. Faisant ainsi suite aux développements originaux de la distance de transport Lp pour l'analyse du signal par Thorpe et collaborateurs. Des graphes discrets des signaux prédits et observés sont comparés plutôt que les signaux eux-mêmes. Le signal temporel est représenté, après discrètisation temporelle, par une mesure empirique dans le domaine 2D temps-amplitude, ce qui assure la positivité des données comparées. Cette nouvelle fonction coût montre des propriétés de convexité prometteuses en ce qui concerne les transformations en temps et en l'amplitude. En 2D, la distance peut être calculée avec un coût supplémentaire modéré en tant qu'un problème d'affectation linéaire. Cependant, l'extension à des dimensions de domaine plus élevées reste à développer et à évaluer, et cela pourrait augmenter considérablement le coût de calcul.

Dans cette thèse, nous avons développé une troisième famille de fonctions coût, continues et différentiables, basées sut l'OT issues en combinant deux extensions récentes:

- La variante OT « déséquilibrée » introduite récemment et qui définit rigoureusement une distance sur l'espace des mesures de Radon positives contournant ainsi le problème de la normalisation des données. Cette classe de distance OT a été déjà fait l'objet d'une application FWI par Li et collaborateurs en 2021.

- La méthode de régularisation entropique de OT qui permet une implémentation efficace à l'aide de l'algorithme de Sinkhorn. L'OT entropique ne définit pas réellement une distance mais fournit une approximation du second ordre de la distance 2-Wasserstein sous la forme dite de divergence Sinkhorn, qui réduit significativement le coût de calcul. La divergence de Sinkhorn peut être étendue à un OT déséquilibré. Cette approche nécessite une separation du signal en partie positive et partie negative ainsi qu'un filtrage.

Nous donnons dans ce document un aperçu du cadre théorique de ces deux extensions récentes de l'OT conduisant à des fonctions objectifs lisses et différentiables dans le contexte de FWI, dans le contexte général de la théorie du transport optimal, ainsi qu'une revue des différentes approches existantes. Ces extensions

récentes sont en particulier analysées dans le contexte d'une application) la méthode de FWI au travers d'études paramétriques 1D et 2D. Leur implémentation, exploitant le.s propriétés de tensorisation pour des espaces discrets cartésiens sont présentées ainsi qu'une évaluation de la complexité de l'algorithme de Sinkohrn.

Nous illustrons numériquement l'utilisation de ces fonctions coût dans le cadre de FWI au travers de cas académiques et d'un example de référence (Marmousi-I). Les illustrations numériques font cependant un certain nombre d'hypothèses physiques par souci de simplicité qui doivent être gardées à l'esprit :

- Nous utilisons l'approximation acoustique classique des ondes élastiques et supposons une densité constante dans les illustrations 2D ; dans ce cas, la vitesse du son est le seul paramètre du modèle. Dans l'approximation acoustique, la conversion de mode n'est pas prise en compte et le milieu est isotrope. En pratique, lorsqu'il s'agit de milieux hétérogènes, une telle approximation peut être remise en cause Cance et Capdeville (2015)

- Nous supposons que la source sismique est ponctuelle et connue, alors qu'en pratique la source doit être déterminée ou calibrée, généralement à partir d'une arrivée directe.

- Nous supposons que les observables sont des échantillons ponctuels de la solution de l'équation des ondes acoustiques, plutôt que des versions filtrées (à la fois dans l'espace et dans le temps) des perturbations de pression mesurées par les récepteurs.

- Nous ne traitons pas la problématique des bruits d'acquisition dans les observations (ex : dysfonctionnement des détecteurs, bruit sismique ambiant, diffusion incohérente des structures qui ne sont pas à imager). En tant que telles, les illustrations numériques 2D présentées dans ce travail doivent être considérées comme des illustrations académiques qui doivent être étudiées plus avant et évaluées dans des contextes plus réalistes.

Les résultats de cette étude ne permettent pas de conclure quand à l'efficacité de cette approche par rapport aux autres methode existantes mais met partiellement en lumière l'impact non trivial des différents paramêtres de régularisation et de lissages intervenants dans la mise en oeuvre de ces distances OT entropiques. Elle permet toutefois de confirmer leur efficacité computationelle.

Chapter 1

Introduction

Seismic tomography aims at inferring physical properties and reconstructing quantitatively structures of the Earth interior from the mechanical waves that are radiated by natural and man-made seismic sources and recorded at the surface at receivers in the form of seismograms also called seismic traces. This is a fundamental problem in geophysics with applications to societal concerns such as the discovery and exploitation of energy resources, site investigation for civil engineering, mineral deposits and underground water supplies, seismic hazard and risk mitigation.

With respect to other physics-based osculating methods – *e.g.* gravimetry, geomagnetism, and electromagnetism – seismic waves have large source-dependent penetration depths and provide high-resolution reconstruction due to their source-dependent short wavelengths. They are today recorded with increasing accuracy by extended arrays with a large number of receivers – seismometers, accelerometers, hydrophones, geophones, fibre optic cables – at the surface (land and sea) or near-surface (boreholes) of the Earth.

In seismology, passive source tomography that makes use of natural seismic sources – *e.g.* earthquakes radiating very energetic seismic waves – have been actively developed in the last decades, and provide significant insights of the deep Earth interior from global to regional scales. The main limitation being the distribution and the frequency content of the natural sources, together with the available data coverage at these scales. In exploration geophysics, the development of controlled man-made sources –*e.g.* explosive and vibro-seis sources on land, air-gun sources at sea – and of long and very long offset acquisition systems has foster the development of high-resolution seismic imaging methods exploiting increasingly large amount of data. This led to higher and higher resolution subsurface models at local scales, from dozens of meters to dozens of kilometers, depending on the source energy and the acquisition system geometry.

Interactions between the waves and the geological medium depend on: the limited frequency-band, the duration and radiated energy of the seismic sources; the characteristic heterogeneity scales of the medium; and of the source-receiver wave propagation path and distance. Extracting information from the recorded seismograms is not a trivial task as these interactions get integrated over the propagation distance. Usable data in seismic tomography have long been restricted to secondary observables such as travel times, phase speeds or waveforms – a small portion of the full recorded seismograms – by the limited capability of the physics and numerical wave propagation modelling.

Most discoveries related to the structure of the Earth interior – such as the Earth inner core, asthenosphere and the major seismic discontinuities – have long been based on ray theory describing short-wavelength energy propagation and travel-time attributes. In the ray theoretical framework, the arrival time of the seismic phases are sensitive to the wave speeds along the ray path connecting source and receiver. Travel time tomography (*Aki et al.*, 1977; *Luo and Schuster*, 1991; *Pratt and Goulty*, 1991; *Schuster and Quintus-Bosz*, 1993; *Trampert and Woodhouse*, 1995; *Ekström et al.*, 1997; *Nemeth et al.*, 1997; *Zelt*, 1999; *Boschi and Dziewonski*, 2000; *Rawlinson and Sambridge*, 2003; *Romanowicz*, 2003) is a fast and cost effective tool. It is however limited to well-isolated body-wave phases or surface waves, for which the fundamental and higher modes are well separated. High-frequency approximation does not take into account the source finite-frequency band, and is only applicable to smooth media, in which the characteristic scales of the inhomogeneities are much larger than the dominant wavelength. As such the velocity structure derived from travel-time tomography is only sub-optimal, and ray-based method can yield distorted results and fail at caustics (*Červený and Pšenčík*, 2011).

Efforts to overcome the limitations of ray theory led to the development of finite-frequency travel-time tomography, a semi-analytical extension of the ray-based inverse problem accounting for spatially extended, finite-frequency 3D wave-sensitivity kernels, together with the integration of different data sets (*Červený and Soares*, 1992; *Masters et al.*, 1996; *Friederich*, 1999, 2003; *Ritsema et al.*, 1999; *Dahlen et al.*, 2000; *Romanowicz*, 2000; *Gu et al.*, 2001; *Montelli et al.*, 2004; *Yoshizawa and Kennett*, 2004, 2005; *Boschi*, 2006; *Sigloch et al.*, 2008). In particular asymptotic finite-frequency kernels allowed to exploit weighted, low-amplitude energy wave packets in the seismograms (*Li and Romanowicz*, 1996; *Mégnin and Romanowicz*, 2000; *Gung and Romanowicz*, 2004) leading to improved global tomographic models.

In exploration seismology, more controlled setup and multifold acquisition systems have enabled accurate, passive seismic imaging methods making use of both travel times and reflectivity energy information. However limited offsets of the seismic reflection surveys and limited frequency band of the source make seismic imaging poorly sensitive to intermediate wavelengths (*Jannane et al.*, 1989).

As such, two broad regimes of wave interactions were mainly considered: a transmission regime, in which wave interactions are dominated by the slow variations of the physical properties, and the seismic phases and their propagation direction are slightly perturbed with a dominant forward scattering; a reflectivity regime, in which wave interactions are dominated by the fast variations of the physical properties, and the seismic phases and their propagation directions are strongly perturbed with a significant back scattering. However, it is worth to note that other types of interactions such as multiple scattering may also contribute to the recorded traces.

As one phase is mixing different interaction regimes during their propagation, the recorded seismic traces contain both kinematic information and reflectivity information. Following the early breakthroughs by *Claerbout* (1971, 1976), two-step seismic imaging workflows orchestrating a tomography mode and a migration mode were actively developed, assuming that reflectivity result from small perturbation in velocity under the Born approximation, and that data – up to a first-order perturbation – linearly depend on the short

wavelengths of the velocity model.

In tomography mode, long wavelength components of the velocity model are reconstructed using the kinematic information of the wave propagation. In migration mode, short wavelengths images of the sub-surface structures, which appear as a discontinuity for a given seismic signal frequency, are reconstructed after kinematic corrections by amplitude summation and back-projection using different types of migration methods (*Claerbout and Doherty*, 1972; *Gazdag*, 1978; *Stolt*, 1978; *Baysal et al.*, 1983; *Nemeth et al.*, 1999; *Yilmaz*, 2001; *Biondi and Symes*, 2004) formulated in the time or the frequency domain.

The notion of full wave form inversion (FWI), which is the context of this thesis, was first introduced to the geophysical community in the early '80s together with the elegant and physically insightful adjoint-state method (*Bamberger et al.*, 1977, 1982; *Lailly*, 1983; *Tarantola*, 1984a,b, 1987, 1988). It was later extended to the frequency domain (*Pratt and Worthington*, 1990; *Pratt*, 1990, 1999; *Sirgue and Pratt*, 2004).

FWI goes one-step further by inverting simultaneously the short and long wavelengths with the potential to bridge the gap between the transmission and the reflectivity regimes, overcoming the limitations of the conventional sequential seismic-imaging workflow involving a tomography step followed by a migration step.

FWI is formulated as a nonlinear PDE-based optimisation problem that considers the full wavefield information to estimate the Earth interior properties. The adjoint-state methods allows the computation of the derivative of an objective function – measuring the difference between the observations and the predicted observables – with respect to the model parameters by combining the predicted incident wavefield governed by the wave equations, together with appropriate initial and boundary equations, and the adjoint wavefield governed by a set of adjoint equations, together with adjoint subsidiary terminal and boundary conditions.

In the last decades, FWI has been actively studied in both academia and industry, taking advantage of: (1) the development of new physical and numerical modelling methods for 3D full wavefield propagation – from acoustic to viscoelastic – in heterogeneous, complex media; (2) the rapidly increasing capability and capacity of high-performance computing; (3) the increasing amount and coverage of broadband seismic data.

FWI has proven to be a powerful tool that dramatically improved the capability to estimate physical properties and structures of various geological targets from global to regional scales (*Komatitsch et al.*, 2002; *Tromp et al.*, 2005; *Fichtner et al.*, 2006a,b; *Tape et al.*, 2007; *Liu and Tromp*, 2008; *Fichtner et al.*, 2008; *Fichtner*, 2011; *Peter et al.*, 2011; *Fichtner et al.*, 2013; *Zhu et al.*, 2015; *Komatitsch et al.*, 2016; *Tromp*, 2020) and local scales in exploration geophysics (*Pratt et al.*, 1996; *Pratt and Schipp*, 1999; *Virieux and Operto*, 2009; *Brossier et al.*, 2009; *Plessix and Perkins*, 2010; *Sirgue et al.*, 2010; *Plessix et al.*, 2012; *Warner et al.*, 2013; *Stopin et al.*, 2014; *Vigh et al.*, 2014; *Operto et al.*, 2015; *Shen et al.*, 2018; *Borisov et al.*, 2019).

In practice, however, solving FWI problems using local, nonlinear optimisation methods is facing challenges that preclude routine use.

The physics of seismic waves is complex and the objective functional is poorly sensitive to wavenumber components of the model that are shorter than seismic wavelengths and only through information that is homogenised by the wave propagation operator (*Cupillard and Capdeville*, 2018; *Capdeville and Métivier*, 2018; *Hedjazian et al.*, 2021). Accurate and computationally efficient models of wave propagation – from

pure acoustic waves to anisotropic viscoelasticity – together with well-defined physics-based and data-based parametrisation of the model need to be carefully considered in the FWI, specially in the case of multi-parameter estimation and multi-mode modelling that remain very challenging (*Operto et al.*, 2013). Despite the rapidly increasing computational capabilities, numerical models can only process limited frequencies in 3D. While increasing the level of parametrisation allows more realistic models, the inversion problem becomes more ill-posed and even non-unique when the model space dimension increases.

The accuracy deteriorates from the lack of low frequencies, observation noise and poor starting models (*Gauthier et al.*, 1986; *Mora*, 1987; *Luo and Schuster*, 1991; *Fichtner et al.*, 2008; *Virieux et al.*, 2014), unless some prior information is employed (*Bunks et al.*, 1995). In exploration geophysics, while ultra long-offsets allow to recover partially low frequencies, the seismic traces contain a full-variety of integrated, frequency-dependent information associated with different resolution and interaction regimes during the propagation. This makes the inversion problem highly nonlinear with multimodal objective function. When using local gradient-based optimisation methods, a poor initial model can cause the nonlinear optimisation problem to be trapped easily in a local minimum *e.g.* *Tarantola* (2005). Global optimisation methods – such as Monte Carlo (*Sambridge and Gallagher*, 2011; *Ray et al.*, 2016; *Biswas and Sen*, 2017; *Ray et al.*, 2018), and Hamiltonian Monte Carlo (*Fichtner et al.*, 2019; *Fichtner and Zunino*, 2019; *Gebraad et al.*, 2020) sampling methods, simulated annealing (*Datta and Sen*, 2016), particle swarm optimisation (*Chen and Wang*, 2017) – could avoid theoretically this problem and quantify the uncertainties. However, owing to the large dimension of the model space, these methods remain currently intractable and far from reaching the computational efficiency needed to handle realistic large-scale seismic inversion problems in exploration geophysics.

Seismic data are oscillatory signals by nature, as the zero-frequency part is not recorded (*Chauris*, 2021). Since the optimisation problem is highly nonlinear with multimodal objective functions a poor initial kinematic model may cause the solution to be trapped in local minima. In the standard form of FWI, the objective function, measuring the misfit between the predicted observables and the observations, is based on the least-squares distance, which is oscillatory and non convex in directions associated with wavenumber components of the model that are longer than a half the dominant wavelength. For large kinematic errors, local optimisation algorithms will match incorrect phases between the predicted and the observed signals, making the method prone to the so-called cycle-skipping effect. The inversion converges in a local minimum, possibly far from the global minimum, due to kinematic errors introduced by the poor starting model.

Different strategies have been developed to mitigate the non convexity of the FWI problem and to reduce the dependency on the initial model:

- Improve the kinematic accuracy of the initial model using for example reflection tomography(*Woodward et al.*, 1973) or stereo-tomography methods (*Lambaré*, 2008), so that the initial model is not too far from the correct model and the kinematic errors are reduced. Following a similar approach, it has also been proposed to enhance the tomographic components at the early iterations and gradually reduce its weights towards convergence (*Tang et al.*, 2013), or to further separate the gradient components based on the scattering angle at the imaging point (*Alkhalifah*, 2015). Although these methods do enhance

the low wavenumber components of the FWI gradient, they reduce the flexibility of the FWI workflow and can be rather time consuming.

- Mitigate multi-modalities and enlarge the basin of attraction of the objective function using a data hierarchical approach: seeding it with low frequency data only, and slowly enlarge the data bandwidth as the iterations of the gradient-based methods progress (*Pratt*, 1999; *Shipp and Singh*, 2002; *Brossier et al.*, 2009; *Wang and Rao*, 2009). Extension to the time domain involves different temporal and/or offset selection of the observations. The idea is to reduce the number of propagated wavelengths to be recovered simultaneously. Among the many factors that affect in practice the success of this approach, the lowest starting frequency is an essential one. Broadband data with high signal-to-noise ratio may not be always available or of sufficient quality preventing reliable model reconstruction. Moreover, smart selection of the data is required for the different inversion stages, which remains often rather empirical. Following similar approach, low frequency data extrapolation based on the phases and amplitudes in the observed band has also been proposed, fitting smooth non-oscillatory functions to represent and extrapolate wave physics to the unrecorded frequency band (*Li and Demanet*, 2015).

Other solutions based on a reformulation of the FWI have been extensively studied in the past decades to reduce the non-linearity of the continuous PDE-based optimisation problem and to increase the convexity of the inversion in higher-dimensions. This has led to different extension strategies that aim to restore the dimensional balance between the model and the data space by considering artificial, extended models through the introduction of redundant coordinates in the model space. The common feature in all model extension strategies is their tendency to promote the same sort of long-wavelength velocity updates as does travel time tomography, *i.e.* to extract kinematic information from the data.

Enforcing a consistency condition that link the unphysical, extended model to the original physical one is widely viewed as a reasonable way of retrieving the background model in which waves propagate. There are at least two different ways of enforcing the consistency relation:

- A first strategy in the image (reflectivity) domain was originally developed within the framework of wave equation migration analysis (*Symes*, 2008) and relies on a scale separation assumption. The idea is to introduce additional degrees of freedom at the reflectivity level extending the model space along subsurface offset, plane-wave ray-parameter, or time-lag (*Shen et al.*, 2003; *Shen*, 2004; *Sava and Fomel*, 2006; *Sun and Symes*, 2012; *Biondi and Almondin*, 2014). The extended modelling operator is now surjective under very general conditions in 3D, and as such an extended model can always be found to match any data, regardless of wether the background kinematic model is correct. The construction delocalises scattering events and the extended image (reflectivity) should take on near-zero values because the wave fields are not supposed to interact constructively. The consistency condition is then derived by penalising the additional parameters in order to focus the energy of the extended reflectivity image (*Chauris and Noble*, 2001; *Biondi and Symes*, 2004; *Sava and Fomel*, 2006; *Symes*, 2008; *Shen and Symes*, 2008; *Biondi and Almondin*, 2014; *Cocher et al.*, 2017) . The main drawback is

the additional computational cost to the already computationally demanding FWI, associated to the construction of the extended reflectivity images. Moreover since only primary reflections are used to generate the images, considerable preprocessing – such as removing the multiples, muting the direct arrivals, and diving waves – is required.

- A second strategy is the source-receiver extension strategy introducing the seismic source indexes as additional coordinates in the model parameter space (*van Leeuwen and Herrmann*, 2013; *Warner and Guasch*, 2016; *Huang et al.*, 2017, 2018), initially including trace-based Wiener filters to compare the observed and modelled data and later on the whole wavefield so that the reconstructed wavefield fits the data by design. The imaging condition involves incident and adjoint fields computed in the extended model independently for each source. The consistency condition again relies on the concept of differential semblance optimisation (*Symes and Carrazone*, 1991), which is a good though not fully compelling choice of regularisation for model-extended waveform inversion. The applicability of these methods is by no mean feat when multiple arrivals are present in the transmission data (*Plessix et al.*, 2000), as well as their extension to the time domain (*Aghamiry et al.*, 2020). Current solutions either rely on crude approximations (*Wang et al.*, 2016) or on advanced iterative solutions, which are computationally demanding (*Aghamiry et al.*, 2020)

Apart from finding an adequate initial kinematic model, another line of investigation is to reformulate the FWI using alternative ways to measure the difference between predicted observables and observations in an attempt to mitigate the non convexity of the objective function with respect to the model parameters and to enlarge the basin of attraction of the global minimum. Different families of misfit function with enhanced convexity have been proposed in the past decades:

- A first family involves transforming the signal before measuring the difference between the predicted observable and observations in the sense of the least-squares distance. Examples of such transformation include: reconstruction of the envelope and the unwrapped phase of the signal (*Wu et al.*, 2014; *Luo and Wu*, 2015), extraction of the instantaneous phase and enveloppe of the signal by the Hilbert transform (*Fichtner et al.*, 2008; *Bozdağ et al.*, 2011; *Alkhalifah and Choi*, 2012; *Luo et al.*, 2018), normalised integration of the signal leading to trace cumulative distribution (*Liu et al.*, 2012; *Donno et al.*, 2013).

- Another family involves more global measures in replacement of the least-squares distance. Correlation-based objective functions, promoting the minimisation of travel-time shifts, were initially proposed (*Luo and Schuster*, 1991; *Dahlen et al.*, 2000; *Zhao et al.*, 2000; *van Leeuwen and Mulder*, 2008, 2010; *Choi and Alkhalifah*, 2012), leading to the so-called wave equation tomography strategy. However, crosscorrelation-based measures tend to fail in the case of multiple energetic seismic arrivals. Matching filter-based objective functions (*Luo and Sava*, 2011; *Warner and Guasch*, 2016; *Huang et al.*, 2017, 2018; *Zhu and Fomel*, 2016) were then developed to circumvent these limitations, making use of a normalised deconvolution, *i.e.* a Wiener filter, of the predicted and observed seismic traces , leading to the so-called adaptive wave inversion method. The resulting objective function penalises the energy of

the filter away from a band-pass Dirac filter, which would have resulted from a correct model. However, these objective functions can still suffer from cycle-skipping and from a loss of resolution in the case of complex multi-arrival signals.

Matching-filter-based objective functions can be seen as an attempt to regularise the ill-conditioned inverse problem by controlling the slowly decaying energy associated to multiple arrivals and by removing effects of the small eigenvalues – associated to wavenumber components of the model that are shorter than seismic wavelengths – of the linearised forward wave operator.

Finally, a new trend has recently emerged in inverse problems and machine learning promoting optimal transport (OT) distances to measure the difference between the predicted observables and the observations, defining new objective or fidelity functions. Those newly proposed methods transform the local, sample-by-sample least-squares measure to a global one, trace by trace, or even source gather by source gather. OT-based distances are appealing because of their Lagrangian nature and convexity with respect to translations and dilations of a prescribed probability distribution. To date several families of OT-based functions have been proposed:

- A first family, proposed originally by Engquist and collaborators (*Engquist and Froese*, 2014; *Engquist et al.*, 2016; *Qiu et al.*, 2017; *Yang and Engquist*, 2018; *Yang et al.*, 2018; *Sun and Alkhalifah*, 2019; *Engquist and Yang*, 2019; *Engquist et al.*, 2020; *Engquist and Yang*, 2021) is linked to the 2-Wasserstein metric associated to balanced optimal transport, which shows nice convex properties with respect to time shifts and rotations. The classical 2-Wasserstein distance however only applies to probability measures. Raw oscillatory seismic signals require ad-hoc transformation and normalisation (*Qiu et al.*, 2017; *Yang and Engquist*, 2018) to be compared using this metric. This can alter the phase and amplitude information, and prone non-convexity with respect to time shifts. Computing the 2-Wasserstein metric using the Monge-Ampère formulation either with optimised finite-difference, linear solvers (*Benamou et al.*, 2014; *Froese*, 2012) or semi-discrete strategy (*Mérigot*, 2011) add significant computational cost to the FWI and in practice application of the 2-Wasserstein distance has often been restricted 1D-optimal transport, *i.e.* trace-by-trace comparison, for which an analytical solution exists.

- A second family, proposed by Métivier and collaborators (*Métivier et al.*, 2016a,b; *Métivier et al.*, 2016), resorts to the 1-Wasserstein metric, a particular instance of the Kantorowitch-Rubinstein metric, associated to the dual formulation of the optimal transport problem. This OT metric is now defined for signed measures and can be lifted to handle unbalanced optimal transport, avoiding ad-hoc signal transformation and normalisation. It leads to a convex optimisation problem under linear constraints that can be solved using proximal splitting methods (*Combettes and Pesquet*, 2011), reducing significantly the additional computational compared to the 2-Wasserstein metric. The convexity of the objective function is however not guaranteed with regard to time and amplitude transformations.

- A third family recently introduced by Métivier et al (*Métivier et al.*, 2018; *Metivier et al.*, 2019) is linked to a graph-space optimal transport distance, following original developments in transport L^p distance

for signal analysis by *Thorpe et al.* (2017). Discrete graphs of the predicted and observed signals are compared rather than the signals themselves. The time-signal is represented, after time discretisation, by a point cloud in the 2D time-amplitude domain, which ensure the positivity of the compared data. This new objective function shows promising convexity properties with respect to time and amplitude transformation, and in 2D the distance can be computed with moderate additional cost as a linear assignment problem. However extension to higher domain dimensions is still under development and remains to be evaluated, and it might increase significantly the computational cost.

To date, OT-based misfit functions have mostly been illustrated using 2D academic and benchmark cases, and only few of them have been practically demonstrated for field data (*Poncet et al.*, 2018; *Messud and Sedova*, 2019; *Sedova et al.*, 2019; *Górszczyk et al.*, 2020, 2021; *Pladys et al.*, 2021).

In this thesis, we developed a third family of OT-based, continuous and differentiable objective functional, derived from combining two recent extensions of OT:

- an "unbalanced" OT variant introduced in (*Chizat et al.*, 2018a; *Chizar et al.*, 2018; *Liero et al.*, 2018; *Kondratyev et al.*, 2016a), which rigorously defines a distance on the set of positive Radon measures thus bypassing the data normalisation issue. This class of OT distance was proposed but not investigated in details in *Li et al.* (2021).

- an entropic regularisation of the OT framework and its efficient implementation using the Sinkhorn algorithm *Cuturi* (2013) *Peyré and Cuturi* (2019). Entropic OT does not actually define a distance but provides a second-order approximation of the 2-Wasserstein distance, and the so-called Sinkhorn divergence variant (*Genevay et al.*, 2018; *Feydy et al.*, 2019) leads to a very efficient computational implementation. The Sinkhorn divergence can be naturally extended to unbalanced OT (*Séjourné et al.*, 2019).

We provide an overview of the theoretical framework of these two recent OT extensions leading to smooth and differentiable objective functions in the context of FWI. We numerically illustrate, in the 2D acoustic case, the use of these objective functions in the context of FWI with academic examples and the Marmousi benchmark data sets.

The numerical illustrations make however a number of physical assumptions for sake of simplicity that need to be kept in mind:

- We use the classical acoustic approximation of elastic waves and assume constant density in the 2D illustrations ; in this case, the sound speed velocity is the only model parameter. In the acoustic approximation mode conversion is not considered and the medium is isotropic. In practice, when dealing with heterogeneous media such an approximation can be challenged *Cance and Capdeville* (2015)

- We assume that the seismic source is ponctual and known, whereas in practice the source needs to be determined or calibrated, usually from a direct arrival.

- We assume that the observables are point samples of the solution of the acoustic wave equation, rather than filtered versions (both in space and time) of pressure disturbances as measured by the receivers.
- We do not deal with the issue of acquisition noise in the observations (*e.g.* malfunctioning detectors, ambient seismic noise, incoherent scattering from structures that are not to be imaged).

As such the 2D numerical illustrations presented in this work has to be considered as academic illustrations that need to be further investigated and evaluated in more realistic contexts.

Outline of the manuscript

Chapter 2 (Seismic Background) presents the Full-waveform Inversion (FWI) problem of reconstructing a background model and the classical adjoint state method used to compute the gradient of the misfit function, and provides a discussion the so-called "cycle skipping" problem in the case of the classical least-squares misfit function.

Chapter 3 (OT Background) a general presentation of the Optimal Transport (OT) theory and of the main contributions using OT misfit functions in the FWI context is provided.

Chapter 4 (Entropic regularization of OT and its generalizations) presents recent branches of OT. First its entropic penalization and companion Sinhkhorn algorithm . Then its "unbalanced" extension allowing to define a bona fide distance between positive distributions with total mass. Finally how the two extension can be combined and how to correct the bias introduced by the entropic penalization (Sinkhorn Divergence).

Chapter 5 (The unbalanced optimal transport misfit function) summarizes the construction of the unbalanced misfit function from the model onward. We also illustrate its properties on two simple 1D and 2D parametric models.

Chapter 6 (2D FWI numerical illustrations) illustrates numerically the use in the FWI context of the misfit function presented in chapter 4 on synthetic data, through simple canonical models, and discusses the results.

Chapter 7 (Conclusion) gives partial conclusions supported by this study. We also describe a possible follow up work program.

Chapter 2

Seismic full-waveform inversion: an overview

This section provides a brief overview of the full-waveform inversion and of the adjoint-state formulation, following closely the reviews of *Tarantola* (1988); *Plessix* (2006); *Virieux and Operto* (2009); *Fichtner* (2011); *Virieux et al.* (2014); *Chauris* (2021)

2.1 A general setting

In the context of full-waveform inversion, the state of the dynamical system is determined by the wavefield $\mathbf{u}(\mathbf{x}, t) : \mathbf{\Omega} \times [0, T] \to \mathbb{R}^{n_u}$ for some n_u. We assume that $\mathbf{u}$ belongs to a subset of $\mathcal{W} = L^2([0, T], [L^2(\mathbf{\Omega})]^{n_u})$, The physical or computational domain $\mathbf{\Omega} \subset \mathbb{R}^{n_o}$ can be open or bounded with boundary $\mathbf{\Gamma}$.

The evolution of the dynamical system is governed, under some physical approximation, by a state equations, *i.e.* a system of partial differential equations, represented by the differential operator $\mathbf{L}$, the behavior of which depends on model parameters $\mathbf{m}$:

$$\mathbf{L}\left(\mathbf{u}(\mathbf{x}, t), \mathbf{m}(\mathbf{x})\right) - \mathbf{f}(\mathbf{x}, t) = \mathbf{0}, \quad \forall \mathbf{x} \in \mathbf{\Omega}, \forall t \in [0, T], \tag{2.1}$$

where the seismic source $\mathbf{f} : \mathbf{\Omega} \times [0, T] \to \mathbb{R}^{n_p}$ is in $\mathcal{W}$, and $\mathbf{m} : \mathbf{\Omega} \to \mathcal{M} \subset \mathbb{R}^{n_m}$ the model parameters that define a vector-valued field, with $\mathcal{M}$ the admissible model parameters space. The solution of (2.1) defines a physical realisation of the system $\mathbf{u}(\mathbf{m}) = \mathbf{u}(\mathbf{x}, t; \mathbf{m})$ for fixed model parameters $\mathbf{m}$.

The differential operator $\mathbf{L}$ governing the wavefield propagation, *i.e.* from acoustic to viscoelastic, is generally a linear operator:

$$\mathbf{L}(\mathbf{u}, \mathbf{m}) \equiv \mathcal{L}[\mathbf{m}]\, \mathbf{u}, \tag{2.2}$$

where $\mathcal{L}\,[\mathbf{m}]$ is the wave-propagation operator associated to the physical approximation.

When $\mathcal{L}\,[\mathbf{m}]$ is of order Q in times, the state equations (2.2) must be complemented with $Q - 1$ initial conditions at $t = 0$ (we assume a null initial state to simplify)

$$\frac{\partial^q \mathbf{u}}{\partial t^q}(\mathbf{x}, t = 0) = \mathbf{0}; \quad \forall \mathbf{x} \in \mathbf{\Omega}, \forall q = 0, .., Q - 1, \tag{2.3}$$

and possibly with appropriate boundary conditions along $\mathbf{\Gamma}$ when the physical domain is bounded.

In the data space $\mathcal{D}$, observations $\mathbf{d}_{\text{obs}}$ are seismograms recorded at n_r receivers – *e.g.* seismometers, hydrophones, geophones – for a source $\mathbf{f}$:

$$\mathbf{d}_{\text{obs}} : \{\mathbf{d}_{\text{obs}}(\mathbf{x}_r, t),\, r = 1, \ldots, n_r\}\,; \quad t \in [0, T]\,, \tag{2.4}$$

where $\mathbf{x}_r$ denotes the spatial position vector of the receiver r.

The predicted observables are defined accordingly as

$$\mathbf{d}_{\text{cal}}(m) : \{\mathbf{d}_{\text{cal}}(\mathbf{x}_r, t; \mathbf{m}),\, r = 1, \ldots, n_r\} = \mathcal{R}\mathbf{u}(\mathbf{x}, t; \mathbf{m}); \quad t \in [0, T]\,, \tag{2.5}$$

where $\mathcal{R} : \mathcal{W} \to \mathcal{D}$ is an extraction operator mapping the physical realisation $\mathbf{u}(\mathbf{m})$ onto the receiver positions $\mathbf{x}_r$

$$\mathcal{R} : \mathbf{u}(\mathbf{x}, t; \mathbf{m}) \to \{\mathbf{u}(\mathbf{x}_1, t; \mathbf{m}), \ldots, \mathbf{u}(\mathbf{x}_{n_r}, t; \mathbf{m})\}\,. \tag{2.6}$$

Note that the definition of $\mathcal{R}$ supposes that $\mathbf{u}(\mathbf{x}, t; \mathbf{m})$ is continuous at points $\mathbf{x} = \mathbf{x_r}$; this holds generally when the seismic source is regular enough and when the medium $\mathbf{m}(\mathbf{x})$ is constant around the receivers $\mathbf{x} = \mathbf{x}_r$

FWI is classically formulated as a nonlinear optimisation problem associated to an objective functional $\mathcal{J} : \mathcal{M} \to \mathbf{R}$ attached to the observations and that acts on the physical observables:

$$\mathcal{J}(\mathbf{m}) = h\left(\mathbf{d}_{\text{cal}}(\mathbf{m}), \mathbf{m}; \mathbf{d}_{\text{obs}}\right) = h\left(\mathcal{R}\mathbf{u}(\mathbf{m}), \mathbf{m}; \mathbf{d}_{\text{obs}}\right), \tag{2.7}$$

where the functional $h(\cdot)$, which may depend explicitly on $\mathbf{m}$, measures the difference between the observations and the predicted observables associated to the physical realisation $\mathbf{u}(\mathbf{m}) = \mathbf{u}(\mathbf{x}, t; \mathbf{m})$.

2.2 The optimisation problem and its solution.

The optimisation problem can be stated as finding the optimal model parameters $\mathbf{m}^\star$ such that $\mathcal{J}(\mathbf{m}^\star)$ is the global minimum of $\mathcal{J}$ for a given set of observations:

$$\mathbf{m}^* = \underset{\mathbf{m} \in \mathcal{M}}{\operatorname{argmin}}\, \mathcal{J}(\mathbf{m}) \quad \text{where} \quad \mathcal{J}(\mathbf{m}) = h\left(\mathbf{d}_{\text{cal}}(\mathbf{m}), \mathbf{m}; \mathbf{d}_{\text{obs}}\right). \tag{2.8}$$

This is not a trivial task as $\mathcal{J}$ depends on $\mathbf{m}$ through $\mathbf{d}_{\text{cal}}$ and the relation between $\mathbf{d}_{\text{cal}}$ and $\mathbf{m}$ is given by the state equation (2.1)-(2.2) and (2.6) and is clearly non linear.

Global optimisation methods, which imply exploring the admissible model space $\mathcal{M}$, still remain unfeasible today for most seismology applications owing to the large number of model parameters.

In practice, the minimisation proceeds iteratively toward the global minimum $\mathbf{m}^\star$, starting from a plausible initial model $\mathbf{m}^0$, using local optimisation approach. Gradient-based methods *Nocedal and Wright* (2006),

such as steepest descent, conjugate gradient or Newton-like methods can be formulated as:

$$\mathbf{m}_{k+1} = \mathbf{m}_k + \alpha_k \, \mathbf{g}_k \quad \text{with} \quad \mathbf{g}_k \cdot \nabla_{\mathbf{m}} \mathcal{J}(\mathbf{m}_k) < 0, \tag{2.9}$$

where $\mathcal{J}(\mathbf{m}_{k+1}) < \mathcal{J}(\mathbf{m}_k)$, and $\mathbf{g}_k$ is the local descent direction, with $\nabla_{\mathbf{m}}(\cdot)$ the gradient operator with respect to the model parameters, known as the Fréchet derivative.

The step length $\alpha_k \in \mathbb{R}_+$, must be chosen by a line search process so that $\mathcal{J}(\mathbf{m}_k - \alpha_k \mathbf{g}_k)$ is minimal, and must satisfies the strong Wolfe conditions (*Wolfe*, 1969, 1971), which are sufficient conditions to assure the convergence:

$$\mathcal{J}(\mathbf{m}_k + \alpha_k \mathbf{g}_k) < \mathcal{J}(\mathbf{m}_k) + c_1 \, \alpha_k \, \nabla_{\mathbf{m}}(\mathbf{m}_k) \cdot \mathbf{g}_k \tag{2.10}$$

$$|\nabla_{\mathbf{m}} \mathcal{J}(\mathbf{m}_k + \alpha_k \mathbf{g}_k)| < c_2 \, |\nabla_{\mathbf{m}} \mathcal{J}(\mathbf{m}_k) \cdot \mathbf{g}_k| \, , \tag{2.11}$$

with $0 < c_1 < c_2 < 1$. The choice of the step length α_k is therefore a balance between stability and speed.

The steepest descent algorithm corresponds to $\mathbf{g}_k = -\nabla_{\mathbf{m}} \mathcal{J}(\mathbf{m}_k)$ and the conjugate gradient algorithm to $\mathbf{g}_k = -\nabla_{\mathbf{m}} \mathcal{J}(\mathbf{m}_k) / \beta_k \, \mathbf{g}_{k-1}$, with $\beta_k \geq 0$; whereas for Newton methods $\mathbf{g}_k = -\mathcal{H}_{\mathcal{J}}^{-1}(\mathbf{m}_k) \nabla_{\mathbf{m}} \mathcal{J}(\mathbf{m}_k)$ where the functional Hessian $\mathcal{H}_{\mathcal{J}}$ is defined as

$$\mathcal{H}_{\mathcal{J}}(\mathbf{m}_k) = \frac{\nabla_{\mathbf{m}} \partial \mathcal{J}}{\partial \mathbf{m}}(\mathbf{m}_k), \tag{2.12}$$

and positive definite Hessian assures $\mathcal{J}(\mathbf{m}_{k+1}) < \mathcal{J}(\mathbf{m}_k)$.

Gradient descent typically converges slowly, which is a significant impediment for large-scale problems, whereas the Newton method, which takes into account the curvature of the objective function, converges faster than gradient descent methods in the neighborhood of a local minimum when the Hessian of $\mathcal{J}$ is close to being positive semi-definite.

Newton method is in general much more complicated to set up than gradient descent methods and much more computationally demanding as the Hessian of $\mathcal{J}$ is a large matrix, costly to store and to invert.

Practical alternatives are the quasi-Newton methods which attempt to partially invert the Hessian of $\mathcal{J}$.

$$\mathbf{m}_{k+1} = \mathbf{m}_k - \alpha_k \, \mathbf{Q}_k \, \nabla_{\mathbf{m}} \mathcal{J}(\mathbf{m}_k), \tag{2.13}$$

where $\mathbf{Q}_k \approx \mathcal{H}_{\mathcal{J}}^{-1}$, is symmetric and positive definite.

Among the quasi-Newton methods, the l-BFGS (*Nocedal*, 1980; *Liu and Nocedal*, 1989), a limited memory implementation of the Broyden-Fletcher-Goldfarb-Shanno quasi-Newton method, leads to an efficient iterative algorithm for computing recursively $\mathbf{Q}_k$ involving only the functional derivative of $\mathcal{J}$ at iteration k and at the l previous iterations.

The objective functional (2.7) can be relatively insensitive to wavenumber components of the model that are shorter than the seismic wavelengths. The eigenvalues of the Hessian of $\mathcal{J}$ corresponding to these components are nearly zero and the optimisation problem (2.8) becomes locally ill-posed.

The eigenvectors associated to these eigenvalues are directions in which $\mathcal{J}(\mathbf{m})$ have very small curvature in the vicinity of $\mathbf{m}^\star$, the curvature being twice the second directional derivative in the eigen-directions. Small perturbations of the data or of the model $\mathbf{m}$ induce modifications of $\mathcal{J}$ that may result in large movements of its global minimum in problematic directions in the vicinity of the null-space of the Hessian of $\mathcal{J}$.

All gradient-based minimisation algorithms critically rely on the ability to efficiently compute the directional derivative of the objective function with respect to the model parameters.

2.3 Adjoint-state methods and gradient computation

Adjoint-state methods (*Lions*, 1971; *Chavent*, 1974; *Chavent et al.*, 1975; *Bamberger et al.*, 1982; *Tarantola*, 1988; *Lailly*, 1983; *Plessix*, 2006; *Fichtner*, 2011) allow to compute the directional derivatives of $\mathcal{J}$ with optimal efficiency.

The adjoint wave equations can be derived from the wave equations, and the properties of the adjoint wave field, solution of the adjoint wave equations, are determined by the adjoint source, which is completely specified by the misfit function (2.7).

Generalisation of the adjoint method allows to compute the Hessian of $\mathcal{J}$, which plays a fundamental role in Newton-like methods of non-linear minimisation.

The variation of the wave field $\mathbf{u}$, a functional of $\mathbf{m}$, with respect to $\mathbf{m}$ in the direction of $\delta\mathbf{m}$ is given by the functional derivative, known as the Gâteau derivative:

$$\delta\mathbf{u} = \nabla_{\mathbf{m}}\mathbf{u}\,\delta\mathbf{m}, \tag{2.14}$$

where $\nabla_m(\cdot)$ is the gradient operator with respect to the model parameters, *e.g.* the Fréchet derivative.

The corresponding functional derivative of $\mathcal{J}$ is given by:

$$\delta\mathcal{J} = \nabla_{\mathbf{m}}\mathcal{J}\,\delta\mathbf{m} = \nabla_{\mathbf{m}}h(\mathcal{R}\mathbf{u}(\mathbf{m}),\mathbf{m};\mathbf{d}_{\text{obs}})\,\delta\mathbf{m} + \left\langle \nabla_{\mathbf{u}}h(\mathcal{R}\mathbf{u}(\mathbf{m}),\mathbf{m};\mathbf{d}_{\text{obs}}),\delta\mathbf{u}\right\rangle_{\mathcal{W}}, \tag{2.15}$$

where $\langle\cdot,\cdot\rangle_{\mathcal{W}}$ denotes the inner product in $\mathcal{W}$,

$$\langle\mathbf{u}_1,\mathbf{u}_2\rangle_{\mathcal{W}} = \int_0^T \left(\int_{\mathbf{\Omega}} \mathbf{u}_1(\mathbf{x},t)\mathbf{u}_2(\mathbf{x},t)\,d\mathbf{x}\right) dt, \tag{2.16}$$

and the chain rule is used here together with the definition of $\delta\mathbf{u}$.

The main problem with (2.15) is that explicitly computing the full kernel of the operator $\nabla_{\mathbf{m}}\mathbf{u}$ can be highly inefficient. The adjoint-state method is a very good way to eliminate $\nabla_{\mathbf{m}}\mathbf{u}\delta\mathbf{m}$ so that $\delta\mathcal{J}$ can be computed in more favorable complexity.

In order to achieve this, the state equation $\mathcal{L}[\mathbf{m}]\,\mathbf{u} - \mathbf{f} = \mathbf{0}$ can be linearised, assuming a sufficiently smooth seismic source, leading to:

$$\Big[\nabla_{\mathbf{m}}\mathcal{L}[\mathbf{m}]\ \delta\mathbf{m}\Big]\,\mathbf{u} + \mathcal{L}[\mathbf{m}]\ \delta\mathbf{u} = \mathbf{0} \tag{2.17}$$

which shows that $\delta\mathbf{u}$ can be eliminated by composition with $\mathcal{L}[\mathbf{m}]$ on the left. The main idea of the adjoint-state method is that a copy of $\mathcal{L}[\mathbf{m}]$ $\delta\mathbf{u}$ can materialise in (2.15) provided that $\nabla_{\mathbf{u}}h$ is seen as the adjoint of $\mathcal{L}[\mathbf{m}]$ applied to some field $\boldsymbol{\lambda}$,

$$[\mathcal{L}[\mathbf{m}]]^{\dagger}\ \boldsymbol{\lambda} = -\nabla_{\mathbf{u}}h(\mathcal{R}\mathbf{u}(\mathbf{m}),\mathbf{m};\mathbf{d}_{\text{obs}}) = -\mathcal{R}^{T}\nabla_{\mathbf{d}_{\text{cal}}}h(\mathbf{d}_{\text{cal}}(\mathbf{m}),\mathbf{m};\mathbf{d}_{\text{obs}}), \tag{2.18}$$

with $\boldsymbol{\lambda}$ naturally called the adjoint field and (2.18) the adjoint-state equation. Then,

$$\begin{aligned}\delta\mathcal{J} &= \nabla_{\mathbf{m}}h(\mathcal{R}\mathbf{u}(\mathbf{m}),\mathbf{m};\mathbf{d}_{\text{obs}})\delta\mathbf{m} - \left\langle (\mathcal{L}[\mathbf{m}])^{\dagger}\boldsymbol{\lambda}, \delta\mathbf{u}\right\rangle_{\mathcal{W}} \\ &= \nabla_{\mathbf{m}}h(\mathcal{R}\mathbf{u}(\mathbf{m}),\mathbf{m};\mathbf{d}_{\text{obs}})\ \delta\mathbf{m} - \left\langle\boldsymbol{\lambda}, \mathcal{L}[\mathbf{m}]\delta\mathbf{u}\right\rangle_{\mathcal{W}}\end{aligned}$$

Taking into account (2.17), the first variation of $\mathcal{J}$, now reads

$$\delta\mathcal{J} = \nabla_{\mathbf{m}}h(\mathcal{R}\mathbf{u}(\mathbf{m}),\mathbf{m};\mathbf{d}_{\text{obs}})\ \delta\mathbf{m} + \left\langle\boldsymbol{\lambda}, [\nabla_{\mathbf{m}}\mathcal{L}[\mathbf{m}]\,\delta\mathbf{m}]\mathbf{u}\right\rangle_{\mathcal{W}} \tag{2.19}$$

which is often much easier to compute than (2.15)

2.4 Physics approximation: the acoustic wave

Physical approximation of the dynamical system is an important component in full-waveform inversion. In this thesis, the behaviour of the dynamical system is described by the 2D acoustic wave equation.

Linearised isotropic, inviscid fluid flow, in the physical domain $\boldsymbol{\Omega} \subset \mathbb{R}^2$, leads to

$$\frac{\partial\mathbf{u}}{\partial t} = -\frac{1}{\rho^{\star}(\mathbf{x})}\nabla p, \qquad \frac{\partial p}{\partial t} = -\kappa^{\star}(\mathbf{x})\nabla\cdot\mathbf{u} \tag{2.20}$$

where the model parameters $\mathbf{m} \in \mathcal{D}$ are $\rho^{\star} : \boldsymbol{\Omega} \to \mathbb{R}_+$, the background mass density, and $\kappa^{\star} : \boldsymbol{\Omega} \to \mathbb{R}$, the background bulk modulus with $\kappa^{\star} = \rho^{\star}[\mathbf{c}^{\star}]^2$ with $\mathbf{c}^{\star}$ the sound speed.

The physical state of the dynamical system is $\mathbf{w} = (\mathbf{u}, p) \in \mathcal{W}$, with $\mathbf{u} : \boldsymbol{\Omega}\times\mathbb{R}_+ \to \mathbb{R}^2$ the displacement perturbation and $p : \boldsymbol{\Omega}\times\mathbb{R}_+ \to \mathbb{R}$ the pressure perturbation.

The behaviour of the system is described by the linear wave operator

$$\mathcal{L}(\mathbf{m})\mathbf{w}(\mathbf{x},t) = \left[\frac{\partial}{\partial t} + \mathcal{S}(\mathbf{m})\right]\mathbf{w}(\mathbf{x},t) \quad \text{with} \quad \mathcal{S}(\mathbf{m},\mathbf{x}) = \begin{pmatrix} 0 & \frac{1}{\rho^{\star}(\mathbf{x})}\nabla \\ \kappa^{\star}(\mathbf{x})\nabla\cdot & 0 \end{pmatrix} \tag{2.21}$$

which defines a hyperbolic system of partial differential equations, and $\mathcal{S}$ is anti-self-adjoint, *i.e.* $\mathcal{S}^{\dagger} = -\mathcal{S}$, with respect to the inner product

$$\langle\mathbf{w},\tilde{\mathbf{w}}\rangle = \frac{1}{2}\int_{\boldsymbol{\Omega}}\left(\rho^{\star}\,\mathbf{u}\cdot\tilde{\mathbf{u}} + \frac{1}{\kappa^{\star}}p\,\tilde{p}\right)\,d\mathbf{x} \tag{2.22}$$

in which the factor 1/2 is chosen to be consistent with the physics convention for the energy.

The acoustic wave operator can be rewritten in term of the pressure perturbation only

$$[\mathcal{L}_p(\mathbf{m})\, p](\mathbf{x},t) = \left[\frac{1}{\kappa^\star(\mathbf{x},t)}\frac{\partial^2}{\partial t^2} - \nabla\cdot\left(\frac{1}{\rho^\star(\mathbf{x})}\nabla\right)\right] p(\mathbf{x},t) \tag{2.23}$$

leading to the scalar state equation

$$[\mathcal{L}_p(\mathbf{m})\, p](\mathbf{x},t) - f_p(\mathbf{x},t) = \frac{1}{\kappa^\star(\mathbf{x})}\frac{\partial^2 p(\mathbf{x},t)}{\partial t^2} - \nabla\cdot\left(\frac{1}{\rho^\star(\mathbf{x})}\nabla p(\mathbf{x},t)\right) - f_p(\mathbf{x},t) = 0 \tag{2.24}$$

where $f_p : \mathbf{\Omega}\times\mathbb{R}_+ \to \mathbb{R}$ is the pressure perturbation associated to a seismic source, generally considered as a point source: $f_p(\mathbf{x},t) = s(t)\delta(\mathbf{x}-\mathbf{x}_s)$ where $\mathbf{x}^s \in \mathbf{\Omega}$ is the position of the source and $s(t)$ the source time function often modelled by a causal Ricker wavelet

$$s(t) = \left[1 - 2\pi^2 f_0^2\,(t-t_0)^2\right]\, \mathrm{e}^{-\pi^2 f_0^2 (t-t_0)^2} \tag{2.25}$$

where the time shift t_0 assures the causality, *i.e.* $s(t) = 0$, for $t < 0$, and f_0 is the characteristic frequency of the wavelet.

The state equation (2.24) needs to be completed by the initial conditions

$$\frac{\partial p}{\partial t}(\mathbf{x},0) = \dot{p}^0(\mathbf{x}) = 0 \quad \text{and} \quad p(\mathbf{x},0) = p^0(\mathbf{x}) = 0; \quad \forall \mathbf{x}\in\mathbf{\Omega}. \tag{2.26}$$

When the physical domain is limited by an upper free surface $\mathbf{\Gamma}_1$, the associated boundary condition is

$$p|_{\mathbf{\Gamma}_1} = p(\mathbf{x},t) = 0, \quad \forall \mathbf{x}\in\mathbf{\Gamma}_1,\, \forall t\in[0,T]\,. \tag{2.27}$$

When the physical domain is unbounded, the computational domain must be truncated by artificial interfaces $\mathbf{\Gamma}_2$ together with an appropriate absorbing boundary condition (ABCs). The simplest form of ABC is the zero-order absorbing boundary condition (*Clayton and Engquist*, 1977; *Engquist and Majda*, 1977; *Nataf*, 2013) derived from the approximation of the one-way wave equation,

$$\frac{\partial p}{\partial t}(\mathbf{x},t) + \mathbf{c}^\star(\mathbf{x})\nabla p\cdot\mathbf{n} = 0, \quad \forall \mathbf{x}\in\mathbf{\Gamma}_2, \quad \mathbf{c}^\star(\mathbf{x}) = \sqrt{\frac{\kappa^\star(\mathbf{x})}{\rho^\star(\mathbf{x})}}, \tag{2.28}$$

where $\mathbf{n}$ the unit outward normal associated to $\mathbf{\Gamma}_2$.

Designing good absorbing boundary conditions is a somewhat difficult problem that has a long history. The currently most popular solution to this problem is to slightly expand the computational domain through an absorbing perfectly matched layer (PML) *Berenger* (1994). The PML shows great superiority in the context of acoustic wave simulation over the classical ABC mentioned above (*Arbanel and Gottlieb*, 1997; *Diaz and Joly*, 2006; *Bermúdez et al.*, 2006; *Gao et al.*, 2017). In particular, convolutional implementations of complex frequency shifter perfectly matched layer (*Roden and Gedney*, 2000; *Festa and Vilotte*, 2005;

Komatitsch and Martin, 2007) have become widely used in acoustic wave simulation (*Pasalic and McGarry*, 2010; *Gao et al.*, 2017). Here, to simplify the presentation, we adopt the first order absorbing condition (2.28)

This leads to the forward modelling problem

$$\begin{cases} \dfrac{1}{\kappa^\star(\mathbf{x})}\dfrac{\partial^2 p(\mathbf{x},t)}{\partial t^2} - \nabla\cdot\left(\dfrac{1}{\rho^\star(\mathbf{x})}\nabla p(\mathbf{x},t)\right) - f_p(\mathbf{x},t) = 0; & \forall \mathbf{x}\in\mathbf{\Omega},\ \forall t\in[0,T] \\ \dfrac{\partial p}{\partial t}(\mathbf{x},0) = \dot{p}^0(\mathbf{x}) = 0 \quad \text{and} \quad p(\mathbf{x},0) = p^0(\mathbf{x}) = 0; & \forall \mathbf{x}\in\mathbf{\Omega} \\ p(\mathbf{x},t) = 0; & \forall \mathbf{x}\in\mathbf{\Gamma}_1,\ \forall t\in[0,T] \\ \dfrac{\partial p}{\partial t}(\mathbf{x},t) + \mathbf{c}^\star(\mathbf{x})\nabla p(\mathbf{x},t)\cdot\mathbf{n}(\mathbf{x}) = 0, & \forall \mathbf{x}\in\mathbf{\Gamma}_2. \end{cases} \tag{2.29}$$

In order to specify the adjoint acoustic wave operator $\mathcal{L}_p^\dagger$ let consider $\forall\lambda\in\mathcal{W}^\dagger$;

$$\begin{aligned} \left\langle \mathcal{L}_p(\mathbf{m})p,\lambda\right\rangle_{\mathcal{W}} &= \left\langle \frac{1}{\kappa^\star(\mathbf{x})}\frac{\partial^2 p(\mathbf{x},t)}{\partial t^2} - \nabla\cdot\left(\frac{1}{\rho^\star(\mathbf{x})}\nabla p(\mathbf{x},t)\right),\lambda\right\rangle_{\mathcal{W}} \\ &= \int_0^T\int_{\mathbf{\Omega}} \frac{1}{\kappa^\star}\frac{\partial^2 p}{\partial t^2} - \nabla\cdot\left(\frac{1}{\rho^\star}\nabla p\right)\lambda\, d\mathbf{x}\, dt. \end{aligned} \tag{2.30}$$

After two integration by part:

$$\begin{aligned} \int_0^T\int_{\mathbf{\Omega}}\left(\frac{1}{\kappa^\star(\mathbf{x})}\frac{\partial^2 p}{\partial t^2}\right)\lambda\, d\mathbf{x}\, dt &= \int_0^T\int_{\mathbf{\Omega}}\left(\frac{1}{\kappa^\star}\frac{\partial^2\lambda}{\partial t^2}\right)p\, d\mathbf{x}\, dt \\ &+ \int_0^T\left[\left(\frac{1}{\kappa^\star}\frac{\partial p}{\partial t}\right)\lambda\right]_{t=0}^{t=T} d\mathbf{x} - \int_{\mathbf{\Omega}}\left[\left(\frac{1}{\kappa^\star(\mathbf{x})}\frac{\partial\lambda}{\partial t}\right)p\right]_{t=0}^{t=T} d\mathbf{x}. \end{aligned} \tag{2.31}$$

The initial conditions (2.26) imply that the first two last terms in the right hand side can be eliminated when imposing the terminal conditions for λ

$$\lambda(\mathbf{x},T) = 0 \quad \text{and} \quad \frac{\partial\lambda}{\partial t}(\mathbf{x},T) = 0; \quad \forall\mathbf{x}\in\mathbf{\Omega}. \tag{2.32}$$

Similarly applying the divergence theorem and the identity

$$\nabla\cdot\left(\frac{1}{\rho^\star}\lambda\nabla p\right) - \nabla\cdot\left(\frac{1}{\rho^\star}p\nabla\lambda\right) = \lambda\nabla\cdot\left(\frac{1}{\rho^\star}\nabla p\right) - p\nabla\cdot\left(\frac{1}{\rho^\star}\nabla\lambda\right), \tag{2.33}$$

together with the boundary condition (2.27) yields to

$$\begin{aligned} \int_0^T\int_{\mathbf{\Omega}}\nabla\cdot\left(\frac{1}{\rho^\star}\nabla p\right)\lambda\, d\mathbf{x}\, dt &= \int_0^T\int_{\mathbf{\Omega}}\nabla\cdot\left(\frac{1}{\rho^\star}\nabla\lambda\right)p\, d\mathbf{x}\, dt \\ &+ \int_0^T\int_{\mathbf{\Gamma}_1}\left(\frac{1}{\rho^\star}\lambda\nabla p\right)\cdot\mathbf{n}\, ds\, dt + \int_0^T\int_{\mathbf{\Gamma}_2}\left(\frac{1}{\rho^\star}\lambda\nabla p\right)\cdot\mathbf{n}\, ds\, dt \\ &- \int_0^T\int_{\mathbf{\Gamma}_2}\left(\frac{1}{\rho^\star}p\nabla\lambda\right)\cdot\mathbf{n}\, ds\, dt, \end{aligned} \tag{2.34}$$

where ds is a line element and $\mathbf{n}$ is the unit outward-pointing normal on the boundaries.

Making use of the absorbing boundary condition (2.28)

$$\begin{aligned}\int_0^T \int_{\mathbf{\Gamma}_2} \left(\frac{1}{\rho^\star}\lambda\nabla p\right)\cdot\mathbf{n}\, ds\, dt &= -\int_0^T \int_{\mathbf{\Gamma}_2} \left(\frac{1}{\rho^\star}\lambda\right)\frac{1}{c}\frac{\partial p}{\partial t}\, ds dt \\ &= \int_0^T \int_{\mathbf{\Gamma}_2} \left(\frac{1}{\rho^\star}p\right)\frac{1}{c}\frac{\partial \lambda}{\partial t} - \left[\int_{\mathbf{\Gamma}2}\frac{1}{\rho^\star}\frac{1}{c}\lambda\, p\, ds\right]_{t=0}^{t=T},\end{aligned} \tag{2.35}$$

where the last term can be eliminated when imposing the initial conditions (2.26) on p and the terminal condition (2.32) on λ.

Gathering the previous results leads

$$\begin{aligned}\left\langle \frac{1}{\kappa^\star}\frac{\partial^2 p}{\partial t^2} - \nabla\cdot\left(\frac{1}{\rho^\star}\nabla p\right), \lambda\right\rangle_{\mathcal{W}} &= \left\langle \frac{1}{\kappa^\star}\frac{\partial^2 \lambda}{\partial t^2} - \nabla\cdot\left(\frac{1}{\rho^\star}\nabla\lambda\right), p\right\rangle_{\mathcal{W}} \\ &+ \int_0^T \int_{\mathbf{\Gamma}_2} \frac{1}{\rho^\star}\left(\frac{1}{c}\frac{\partial\lambda}{\partial t} - \nabla\lambda\cdot\mathbf{n}\right) p\, ds\, dt \\ &- \int_0^T \int_{\mathbf{\Gamma}_1} \left(\frac{1}{\rho^\star}\lambda\nabla p\right)\cdot\mathbf{n}\, ds\, dt.\end{aligned} \tag{2.36}$$

The two last terms in (2.36) can be eliminated by imposing to the adjoint field the boundary condition

$$\begin{cases} \lambda(\mathbf{x},t) = 0; & \forall \mathbf{x}\in\mathbf{\Gamma}_1,\ \forall t\in[0,T] \\ \dfrac{\partial\lambda}{\partial t} - c\nabla\lambda\cdot\mathbf{n} = 0; & \forall \mathbf{x}\in\mathbf{\Gamma}_2,\ \forall t\in[0,T]\,. \end{cases} \tag{2.37}$$

Finally, the equations governing the adjoint wavefield is

$$\begin{cases} \dfrac{1}{\kappa^\star}\dfrac{\partial^2\lambda}{\partial t^2} - \nabla\cdot\left(\dfrac{1}{\rho^\star}\nabla\lambda\right) = -\mathcal{R}^T\nabla_{\mathbf{d}_{\text{cal}}} h\left(\mathbf{d}_{\text{cal}} = \mathcal{R}p, \mathbf{m}; \mathbf{d}_{\text{obs}}\right); & \forall\mathbf{x}\in\mathbf{\Omega},\ \forall t\in[0,T] \\ \dfrac{\partial\lambda}{\partial t}(\mathbf{x},T) = 0 \text{ and } \quad \lambda(\mathbf{x},T) = 0; & \forall\mathbf{x}\in\mathbf{\Omega} \\ \lambda|_{\mathbf{\Gamma}_1} = \lambda(\mathbf{x},t) = 0; & \forall\mathbf{x}\in\mathbf{\Gamma}_1,\ \forall t\in[0,T] \\ \dfrac{\partial\lambda(\mathbf{x},t)}{\partial t} - \mathbf{c}^\star(\mathbf{x})\,\nabla\lambda(\mathbf{x},t)\cdot\mathbf{n}(\mathbf{x}) = 0, & \forall\mathbf{x}\in\mathbf{\Gamma}_2,\ \forall t\in[0,T]\,. \end{cases} \tag{2.38}$$

The adjoint state λ is solution of a backward problem where initial conditions (2.26) are replaced by the terminal condition (2.32). Homogeneous free surface boundary condition should be the same for the forward and adjoint wave field, whereas absorbing boundary conditions involving time derivatives needs to be properly time-reversed.

Introducing the change of variable $\tilde{t}\to T-t$, the adjoint problem can be rewritten in the same form as the forward problem with the auxiliary unknown:$\tilde{\lambda} = \lambda(\mathbf{x}, T-t)$ and the time-reverse source $-\mathcal{R}^T\nabla_{\mathbf{d}_{\text{cal}}} h(\mathbf{d}_{\text{cal}}; d_{\text{obs}})(T-t)$. The result needs then to be time-reverse at each point $\mathbf{x}$. This is a nice property that enable to solve the forward and the adjoint problem with the solver.

Note however that λ is not in general the physical wave field run backward in time because of the limited

sampling of the receivers, and instead is introduced purely out of computational convenience.

It is interesting to notice that changing the way to measure the distance between observed and modelled data (i.e. modifying the definition of h) only amounts to modify the source term in the adjoint equation.

2.5 A simplification: the constant density model

To go further, we will assume the following hypothesis

1. The density is assumed to be constant. More precisely, we assume that

$$\rho^\star(\mathbf{x}) = \rho_\mathbf{0}, \quad \kappa^\star(\mathbf{x}) = \rho_\mathbf{0}\,[\mathbf{c}^\star(\mathbf{x})]^2.$$

   In this situation, the only parameter is $\mathbf{c}^\star(\mathbf{x})$, the velocity law in Ω.

2. The admissible parameter set $\mathcal{M}$ is a convex subset of a finite dimensional vectorial space. Practically, we choose M scalar non negative functions $\psi_m(\mathbf{x})$, $\mathbf{x} \in \mathbf{\Omega}$ such that $\sum_m \psi_m(\mathbf{x}) = 1$, $\mathbf{x} \in \mathbf{\Omega}$ and we define ($0 < v^- < v^+$ are some given parameters)

$$\mathcal{M} = \left\{ \mathbf{c}^\star(\mathbf{x}) = \sum_{m=1}^{M} v_m\,\psi_m(\mathbf{x}),\ v_- \le v_m \le v_+,\ \forall m = 1,..,M \right\}. \tag{2.39}$$

   For example, given the partition $\mathbf{\Omega} = \bigoplus_m C_m$, function $\psi_m(\mathbf{x})$ can be 1 on C_m and 0 in $\mathbf{\Omega}\backslash C_m$. Other choices are of course possible.

3. Function $h(\mathbf{d}_{cal}, \mathbf{m}, \mathbf{d_{obs}})$ does not depends on $\mathbf{m} = \mathbf{c}^\star$:

$$h(\mathbf{d_{cal}}, \mathbf{m}, \mathbf{d_{obs}}) = \tilde{h}(\mathbf{d_{cal}}, \mathbf{d_{obs}})$$

   In this situation, (2.19) amounts to

$$\delta\mathcal{J} = \left\langle \lambda, [\nabla_{\mathbf{c}^\star}\mathcal{L}_p(\mathbf{c}^\star)\,\delta\mathbf{c}^\star]p \right\rangle_{\mathcal{W}}. \tag{2.40}$$

With these simplifications, the pressure $p(\mathbf{x}, t)$ is the solution of

$$\begin{cases} \dfrac{1}{[\mathbf{c}^\star(\mathbf{x})]^2}\dfrac{\partial^2 p(\mathbf{x},t)}{\partial t^2} - \Delta p(\mathbf{x},t) = \rho_0\, f_p(\mathbf{x},t); & \forall \mathbf{x} \in \mathbf{\Omega},\ \forall t \in [0,T] \\ \dfrac{\partial p}{\partial t}(\mathbf{x},0) = 0 \text{ and } \quad p(\mathbf{x},0) = 0; & \forall \mathbf{x} \in \mathbf{\Omega} \\ p(\mathbf{x},t) = 0; & \forall \mathbf{x} \in \mathbf{\Gamma}_1,\ \forall t \in [0,T] \\ \dfrac{\partial p(\mathbf{x},t)}{\partial t} + \mathbf{c}^\star(\mathbf{x})\,\nabla p(\mathbf{x},t)\cdot\mathbf{n}(\mathbf{x}) = 0, & \forall \mathbf{x} \in \mathbf{\Gamma}_2,\ \forall t \in [0,T]\,. \end{cases} \tag{2.41}$$

It is possible to show that when $f_p(\mathbf{x}, t)$ is in $\mathcal{W}$ and $\mathbf{c}^\star$ is in $\mathcal{M}$ as given in (2.39) then p is in $C^1([0,T], L^2(\mathbf{\Omega}))\cap$

$C^0([0,T],H^1(\mathbf{\Omega}))$. Now, we define the source term for the adjoint equation as

$$S_{\rm ad}(\mathbf{x},t)) = -\mathcal{R}^T\nabla_{\mathbf{d}_{\rm cal}}\tilde{h}\left(\mathbf{d}_{\rm cal}=\mathcal{R}p,\mathbf{d}_{\rm obs}\right)(\mathbf{x},t), \tag{2.42}$$

or, more explicitely,

$$S_{\rm ad}(\mathbf{x},t) = -\sum_r \delta(\mathbf{x}-\mathbf{x}_r)\nabla_{\mathbf{d}_{\rm cal}}\tilde{h}\left(\mathbf{d}_{\rm cal}=\mathcal{R}p,\mathbf{d}_{\rm obs}\right)(\mathbf{x}_r,t). \tag{2.43}$$

The adjoint state is solution of

$$\begin{cases} \dfrac{1}{[\mathbf{c}^\star(\mathbf{x})]^2}\dfrac{\partial^2\lambda(\mathbf{x},t)}{\partial t^2} - \Delta\lambda(\mathbf{x},t) = S_{\rm ad}(\mathbf{x},t) & \forall\mathbf{x}\in\mathbf{\Omega},\ \forall t\in[0,T] \\ \dfrac{\partial\lambda}{\partial t}(\mathbf{x},T)=0 \text{ and } \quad \lambda(\mathbf{x},T)=0; & \forall\mathbf{x}\in\mathbf{\Omega} \\ \lambda(\mathbf{x},t)=0; & \forall\mathbf{x}\in\mathbf{\Gamma}_1,\ \forall t\in[0,T] \\ \dfrac{\partial\lambda(\mathbf{x},t)}{\partial t} - \mathbf{c}^\star(\mathbf{x})\,\nabla\lambda(\mathbf{x},t)\cdot\mathbf{n}(\mathbf{x})=0, & \forall\mathbf{x}\in\mathbf{\Gamma}_2,\ \forall t\in[0,T]\,. \end{cases} \tag{2.44}$$

Now, let us define the M-vector $\mathbf{v}=(v_1,v_2,...,v_M)^T$ and $p(\mathbf{x},t;\mathbf{v})$, $\lambda(\mathbf{x},t;\mathbf{v})$ as $p(x,t)$, and $\lambda(x,t)$, the solutions of (2.41)-(2.44) with $\mathbf{c}^\star(\mathbf{x})=\sum_m v_m\psi_m(\mathbf{x})\in\mathcal{M}$; the functional is now a simple function of vector $\mathbf{v}$ and is given by

$$\tilde{J}(\mathbf{v}) = \tilde{h}\left(\mathcal{R}p(\mathbf{x},t;\mathbf{v});\mathbf{d}_{\rm obs}\right) = \tilde{h}\left((p(\mathbf{x}_1,t;\mathbf{v}),p(\mathbf{x}_2,t;\mathbf{v}),\ldots,p(\mathbf{x}_R,t;\mathbf{v}));\,\mathbf{d}_{\rm obs}\right),$$

and we have, according to (2.40),

$$\delta\tilde{J}(\mathbf{v}) = \int_0^T\int_{\mathbf{\Omega}}\lambda(\mathbf{x},t)\,\left([\nabla_{\mathbf{c}^\star}\mathcal{L}_p(\mathbf{c}^\star)\,\delta\mathbf{c}^\star]p(\mathbf{x},t)\right)\,d\mathbf{x}dt.$$

Since

$$\mathcal{L}_p(\mathbf{c}^\star) = \frac{1}{[\mathbf{c}^\star]^2}\frac{\partial^2}{\partial t^2} - \Delta + \delta_{\mathbf{\Gamma}_2}\left(\frac{1}{\mathbf{c}^\star}\frac{\partial}{\partial t} - n\cdot\nabla\right),$$

we have

$$[\nabla_{\mathbf{c}^\star}\mathcal{L}_p(\mathbf{c}^\star)\,\delta\mathbf{c}^\star] = -\frac{2\delta\mathbf{c}^\star}{[\mathbf{c}^\star]^3}\frac{\partial^2}{\partial t^2} - \delta_{\mathbf{\Gamma}_2}\left(\frac{\delta\mathbf{c}^\star}{[\mathbf{c}^\star]^2}\frac{\partial}{\partial t}\right),$$

or, when $\mathbf{c}^\star(\mathbf{x})=\sum_m v_m\psi_m(\mathbf{x})$,

$$[\nabla_{\mathbf{c}^\star}\mathcal{L}_p(\mathbf{c}^\star)\,\delta\mathbf{c}^\star] = -\sum_{m=1}^M\left(\frac{2\psi_m\delta v_m}{[\mathbf{c}^\star]^3}\frac{\partial^2}{\partial t^2} + \delta_{\Gamma_2}\frac{\psi_m\delta v_m}{[\mathbf{c}^\star]^2}\frac{\partial}{\partial t}\right),$$

so that

$$\delta\tilde{J}(\mathbf{v}) = \sum_m\left[-\int_0^T\left(\int_{\mathbf{\Omega}}\lambda(\mathbf{x},t)\,\frac{2\psi_m(\mathbf{x})}{[\mathbf{c}^\star(\mathbf{x})]^3}\,\frac{\partial^2p(\mathbf{x},t)}{\partial t^2}d\mathbf{x} + \int_{\mathbf{\Gamma}_2}\lambda(\mathbf{x},t)\,\frac{\psi_m(\mathbf{x})}{[\mathbf{c}^\star(\mathbf{x})]^2}\,\frac{\partial p(\mathbf{x},t)}{\partial t}d\mathbf{x}\right)\,dt\right]\,\delta v_m,$$

and finally, the gradient of $\tilde{J}$ at point $\mathbf{v}$ is given by

$$\frac{\partial \tilde{J}}{\partial v_m}(\mathbf{v}) = -\int_0^T \left(\int_{\mathbf{\Omega}} \lambda(\mathbf{x},t)\, \frac{2\psi_m(\mathbf{x})}{[\mathbf{c}^\star(\mathbf{x})]^3}\, \frac{\partial^2 p(\mathbf{x},\mathbf{t})}{\partial t^2} d\mathbf{x} + \int_{\mathbf{\Gamma}_2} \lambda(\mathbf{x},t)\, \frac{\psi_m(\mathbf{x})}{[\mathbf{c}^\star(\mathbf{x})]^2}\, \frac{\partial p(\mathbf{x},\mathbf{t})}{\partial t} d\mathbf{x} \right) dt. \tag{2.45}$$

2.6 Discretisation of the problem

The discretisation of the wave equation in heterogeneous medium is a classical topic and has been investigated for a long time. Here, we adopt the simplest method but others are possible. The starting point of the problem is to rewrite the wave problems (2.41) and (2.44) in a weak form, i.e.

$$\textit{Find } p,\ \lambda \textit{ in } C^1([0,T], H^1(\mathbf{\Omega})) \cap C^2([0,T], L^2(\mathbf{\Omega})) \textit{ such that for all } p^t,\ \lambda^t \textit{ in } H^1(\mathbf{\Omega}),$$

$$\begin{cases} m\left(\dfrac{d^2p}{dt^2}, p^t; \mathbf{c}^\star\right) + b\left(\dfrac{dp}{dt}, p^t; \mathbf{c}^\star\right) + k(p, p^t) = S_p(p^t, t) \\ \\ m\left(\dfrac{d^2\lambda_h}{dt^2}, \lambda^t; \mathbf{c}^\star\right) - b\left(\dfrac{d\lambda_h}{dt}, \lambda^t; \mathbf{c}^\star\right) + k(\lambda_h, \lambda^t) = S_\lambda(\lambda^t, t) \\ \\ p(\cdot, t=0) = 0, \quad \dfrac{dp}{dt}(\cdot, t=0) = 0, \quad \lambda(\cdot, t=T) = 0, \quad \dfrac{d\lambda}{dt}(\cdot, t=T) = 0. \end{cases} \tag{2.46}$$

In this formulation, appears three bilinear forms and two linear forms :

- The mass bilinear form
$$m(u, u^t; \mathbf{c}^\star) = \int_{\mathbf{\Omega}} \frac{u(\mathbf{x})\, u^t(\mathbf{x})}{[\mathbf{c}^*(\mathbf{x})]^2}\, d\mathbf{x}.$$

- The stiffness bilinear form
$$k(u, u^t) = \int_{\mathbf{\Omega}} \vec{\nabla} u(\mathbf{x}) \cdot \vec{\nabla} u^t(\mathbf{x})\, d\mathbf{x}.$$

- The boundary mass bilinear form
$$b(u, u^t; \mathbf{c}^\star) = \int_{\Gamma_2} \frac{u_{|\mathbf{\Gamma}_2}(\mathbf{x})\, u^t_{|\mathbf{\Gamma}_2}(\mathbf{x})}{\mathbf{c}^*(\mathbf{x})}\, d\mathbf{x}.$$

- The two linear forms associated to the second terms
$$S_p(p^t, t) = \rho_0 \int_\Omega f_p(\mathbf{x}, t)\, p^t(\mathbf{x})\, d\mathbf{x}, \qquad S_\lambda(\lambda^t, t) = \int_\Omega S_{ad}(\mathbf{x}, t)\, \lambda^t(\mathbf{x})\, d\mathbf{x}.$$

This weak formulation is obtained by multiplying each wave equation by a test function then integrating the result over $\mathbf{\Omega}$ and finally by using the Green's Theorem as well as the boundary conditions.

Remark: When the sources are ponctual, this formulation makes no sense since $H^1(\mathbf{\Omega})$ contains functions which are not continuous on $\mathbf{\Omega}$. This problem can be avoided by regularizing the Dirac mass, i.e. replacing $\delta(\mathbf{x} - \mathbf{x}_o)$ by some smooth positive function $\delta_r(\mathbf{x} - \mathbf{x}_o)$ supported by a small neighborrow of $\mathbf{x}_0$ such that $\int_{\mathbf{\Omega}} \delta_r(\mathbf{x} - \mathbf{x}_o) = 1$.

2.6.1 Semi discretisation in space

We assume that $\mathbf{\Omega}$ is a square in the plane, Γ_1 is its top horizontal line and Γ_2 is composed of the three other sides of the square boundary. We mesh $\mathbf{\Omega}$ with a regular grid of step h. The set of the nodes of this grid is denoted by $\tilde{\mathcal{N}}$ and the associated node is x_n; in order to distinguish between interior nodes and nodes located on the boundaries we set

$$\tilde{\mathcal{N}} = \mathcal{N} \oplus \mathcal{N}_{\Gamma_1}, \qquad \mathcal{N} = \mathcal{N}_0 \oplus \mathcal{N}_{\Gamma_2}.$$

The grid is decomposed in squared cells of side h; the set of all cells is denoted by $\mathcal{C}$: $\Omega = \bigoplus_{C\in\mathcal{C}} C$; at each node $n \in \mathcal{N}$ we associate $\mathcal{C}[n]$ the set of cells whose one vertex is node n. This set is composed of four cells for $n \in \mathcal{N}_0$, two or one cell(s) for $n \in \mathcal{N}_{\Gamma_2}$; we define $N_n^h(\mathbf{x})$, $n \in \mathcal{N}$ the only continuous function whose restriction on each cell is linear and such that $N_n^h(\mathbf{x}_n) = 1$, $N_n^h(\mathbf{x}_{n'}) = 0$, $n \neq n' \in \mathcal{N}'$; its easy to see that the support of $N_n^h(\mathbf{x})$ is $\cup_{C\in\mathcal{C}[n]} C$. The semi discretization of problems (2.41) and (2.44) consists in looking for approximate solutions in $V_h(\mathbf{\Omega}) \subset H^1(\mathbf{\Omega})$, the vector space spanned by the family of functions $\left\{N_n^h(\mathbf{x}),\, n \in \mathcal{N}\right\}$

$$p(\mathbf{x},t) \simeq p_h(\mathbf{x},t) = \sum_{n\in\mathcal{N}} p_n(t) N_n^h(\mathbf{x}), \quad \lambda(\mathbf{x},t) \simeq \lambda_h(\mathbf{x},t) = \sum_{n\in\mathcal{N}} \lambda_n(t) N_n^h(\mathbf{x});$$

p_h and λ_h are defined as the solution of

$$\textit{Find } p_h,\ \lambda_h \textit{ in } C^2([0,T], V_h(\mathbf{\Omega})) \textit{ such that for all } N_n^h, \quad n \in \mathcal{N}$$

$$\begin{cases} m_h\left(\dfrac{d^2p_h}{dt^2}, N_n^h; \mathbf{c}^\star\right) + b_h\left(\dfrac{dp_h}{dt}, N_n^h; \mathbf{c}^\star\right) + k(p_h, N_n^h) = S_p(N_n^h, t) \\ \\ m_h\left(\dfrac{d^2\lambda_h}{dt^2}, N_n^h; \mathbf{c}^\star\right) - b_h\left(\dfrac{d\lambda_h}{dt}, N_n^h; \mathbf{c}^\star\right) + k(\lambda_h, N_n^h) = S_{\mathrm{ad}}(N_n^h, t) \\ \\ p_h(\cdot, t=0) = 0, \quad \dfrac{dp_h}{dt}(\cdot, t=0) = 0, \quad \lambda_h(\cdot, t=T) = 0, \quad \dfrac{d\lambda_h}{dt}(\cdot, t=T) = 0. \end{cases}$$

To obtain this semi-discrete problem, we made the substitutions $p \to p_h$ and $\lambda \to \lambda_h$ and restricted the tests functions to element of $V_h(\mathbf{\Omega})$. The linear and bilinear forms have been approximated (the substrict $_h$) by using a quadrature rule on each cell; more precisely, we assume that the restriction of $\mathbf{c}^\star$ is constant on each cell

$$\mathbf{c}^\star(\mathbf{x}) = \sum_{C\in\mathcal{C}} v_c\, 1_C(\mathbf{x}).$$

We split the integrals according to

$$\int_\Omega \frac{u(\mathbf{x})\, N_n^h(\mathbf{x})}{[\mathbf{c}^\star(\mathbf{x})]^2}\, d\mathbf{x} = \sum_{C\in\mathcal{C}(n)} \int_C \frac{u(\mathbf{x})\, N_n^h(\mathbf{x})}{[\mathbf{c}^\star(\mathbf{x})]^2}\, d\mathbf{x} = \sum_{C\in\mathcal{C}(n)} \frac{1}{v_c^2} \int_C u(\mathbf{x})\, N_n^h(\mathbf{x})\, d\mathbf{x}.$$

$$\int_{\mathbf{\Gamma_2}} \frac{u_{|\Gamma_2}(\mathbf{x})\, N_n^h(\mathbf{x})}{[\mathbf{c}^\star(\mathbf{x})]}\, d\mathbf{x} = \sum_{C\in\mathcal{C}(n)} \int_{\Gamma_2\cap\partial C} \frac{u_{|\Gamma_2}(\mathbf{x})\, N_n^h(\mathbf{x})}{[\mathbf{c}^\star(\mathbf{x})]}\, d\mathbf{x} = \sum_{C\in\mathcal{C}(n)} \frac{1}{v_c} \int_{\Gamma_2\cap\partial C} u_{|\Gamma_2}(\mathbf{x})\, N_n^h(\mathbf{x})\, d\mathbf{x},$$

then we use the quadature rule ($\mathcal{V}(\mathcal{C})$ is the set of the four vertexes of cell C and $\mathcal{V}(\partial\mathcal{C})$ is the set of the two vertexes of $\Sigma_C = \partial C \cap \Gamma_2$ when Σ_C is not empty)

$$\int_C u(\mathbf{x})\, N_n^h(\mathbf{x})\, d\mathbf{x} \simeq \frac{h^2}{4} \sum_{\mathbf{x}_m \in \mathcal{V}(C)} u(\mathbf{x}_m)\, N_n^h(\mathbf{x}_m) = \frac{h^2}{4} u(\mathbf{x}_n),$$

$$\int_{\Gamma_2 \cap \partial C} u_{|\Gamma_2}(\mathbf{x})\, N_n^h(\mathbf{x})\, d\mathbf{x} \simeq \frac{h}{2} \sum_{\mathbf{x}_m \in \mathcal{V}(\partial C)} u_{|\Gamma_2}(\mathbf{x}_m)\, N_n^h(\mathbf{x}_m) = \frac{h}{2} u_{|\Gamma_2}(\mathbf{x}_n).$$

With this process, we obtain

$$m_h(u_h, N_n^h; \mathbf{c}^\star) = \left(\frac{h^2}{4} \sum_{C \in \mathcal{C}(n)} \frac{1}{v_c^2}\right) u_h(\mathbf{x}_n), \quad b_h(u_h, N_n^h; \mathbf{c}^\star) = \left(\frac{h}{2} \sum_{C \in \mathcal{C}(n)} \frac{1}{v_c}\right) [u_h]_{|\Gamma_2}(\mathbf{x}_n).$$

This approximation method is known as *Mass Lumping*; it is essential to obtain *in fine* a time explicit scheme. Now, we define the two diagonal mass matrices (δ_n^m is the Kronecker symbol)

$$\mathbb{M}(\mathbf{v})_{n,m} = \delta_n^m \left(\frac{h^2}{4} \sum_{C \in \mathcal{C}(n)} \frac{1}{v_c^2}\right), \quad \mathbb{B}(\mathbf{v})_{n,m} = \begin{cases} \delta_n^m \left(\frac{h}{2} \sum_{C \in \mathcal{C}(n)} \frac{1}{v_c}\right), & \text{if } \mathbf{x}_n \in \mathcal{N}_{\Gamma_2} \\ 0 & \text{if } \mathbf{x}_n \in \mathcal{N}_0 \end{cases} \tag{2.47}$$

and the stiffness matrix

$$\mathbb{K}_{n,m} = \int_{\mathbf{\Omega}} \vec{\nabla} N_n^h(\mathbf{x}) \cdot \vec{\nabla} N_m^h(\mathbf{x})\, d\mathbf{x},$$

which corresponds for n or m in $\mathcal{N}_0$ (interior nodes) to a nine points stencil for h^2 times the Laplacian. Note that the coefficients of matrix $\mathbb{K}$ are bounded by a pure constant, we deduce that there exists a pure non negative number κ such that

$$(\mathbb{K}U, U) \leq \kappa\, (U, U). \tag{2.48}$$

These definitions being made, the semi discrete problem can be put in the following matricial form

$$\textit{Find } P(t) = (p_n(t))_{n \in \mathcal{N}} \textit{ and } \Lambda(t) = (\lambda_n(t))_{n \in \mathcal{N}} \textit{ in } C^2([0,T], \mathbb{R}^{\#\mathcal{N}}) \textit{ such that}$$

$$\begin{cases} \mathbb{M}(\mathbf{v}) \dfrac{d^2 P(t)}{dt^2} + \mathbb{B}(\mathbf{v}) \dfrac{dP(t)}{dt} + \mathbb{K} P(t) = \tilde{S}_p(t) \\ \mathbb{M}(\mathbf{v}) \dfrac{d^2 \Lambda(t)}{dt^2} - \mathbb{B}(\mathbf{v}) \dfrac{d\Lambda(t)}{dt} + \mathbb{K} \Lambda(t) = \tilde{S}_{ad}(t) \\ P(t=0) = 0, \quad \dfrac{dP}{dt}(t=0) = 0, \quad \Lambda(t=T) = 0, \quad \dfrac{d\Lambda}{dt}(t=T) = 0. \end{cases}$$

2.6.2 Discretisation in time

We pick a time step Δt; our aim is to compute approximations of vectors $P(t)$ and $\Lambda(t)$ at discrete instant $k\Delta t$

$$P^k \simeq P(k\Delta t), \quad \Lambda^k \simeq \Lambda(k\Delta t), \quad k = 0, ..., K,\ K\Delta t = T.$$

For that, we introduce the finite difference operators

$$[\partial^2_{\Delta t^2} U]^k = \frac{U^{k+1} - 2U^k + U^{k-1}}{\Delta t^2}, \qquad [\partial_{2\Delta t} U]^k = \frac{U^{k+1} - U^{k-1}}{2\Delta t},$$

and define the totally discretised problem

$$\textit{Find } P^k \textit{ and } \Lambda^k \textit{ in } \mathbb{R}^{\#\mathcal{N}},\ k = 0, ..., K \textit{ such that for all } k > 0,$$

$$\begin{cases} \mathbb{M}(\mathbf{v})[\partial_{\Delta t^2} P]^k + \mathbb{B}(\mathbf{v})[\partial_{2\Delta t} P]^k + \mathbb{K} P^k = \tilde{S}_p^k = \tilde{S}_p(t = k\Delta t) \\ \mathbb{M}(\mathbf{v})[\partial_{\Delta t^2} \Lambda]^k - \mathbb{B}(\mathbf{v})[\partial_{2\Delta t} \Lambda]^k + \mathbb{K} \Lambda^k = \tilde{S}_{ad}(t = k\Delta t) \\ P^0 = 0, \quad P^1 = 0, \quad \Lambda^K = 0, \quad \Lambda^{K-1} = 0. \end{cases}$$

Note that computing P^{k+1} (resp. Λ^{k-1}) from P^k, P^{k-1} (resp. Λ^{k+1}, Λ^k) involves the inversion of matrix $\mathbb{M}(\mathbf{v}) + \frac{\Delta t}{2}\mathbb{B}(\mathbf{v})$; this matrix being diagonal (thanks to Mass Lumping) the scheme is explicit.

2.6.3 Stability and CFL condition.

The stability of this scheme relies on a conservation of a pseudo energy. For simplicity, we drop the source term and consider instead initial conditions. The model problem is now

$$\textit{Find } U^k \textit{ in } \mathbb{R}^{\#\mathcal{N}},\ k = 0, ..., K \textit{ such that for all } k > 0,$$

$$\begin{cases} \mathbb{M}(\mathbf{v})[\partial_{\Delta t^2} U]^k + \mathbb{B}(\mathbf{v})[\partial_{2\Delta t} U]^k + \mathbb{K} U^k = 0 \\ \dfrac{1}{2}(U^1 + U^0) = u^{(0)}, \quad \dfrac{U^1 + U^0}{\Delta t} = u^{(1)}. \end{cases}$$

We take the scalar product of the equation in U^k by $[\partial_{2\Delta t} U]^k$, we obtain

$$\left(\mathbb{M}(\mathbf{v})[\partial_{\Delta t^2} U]^k,\, [\partial_{2\Delta t} U]^k\right) + \left(\mathbb{K} U^k, [\partial_{2\Delta t} U]^k\right) + \left(\mathbb{B}(\mathbf{v})[\partial_{2\Delta t} U]^k,\, [\partial_{2\Delta t} U]^k\right) = 0. \tag{2.49}$$

Let us define

$$V^{k+\frac{1}{2}} = [\partial_{\Delta t} U]^{k+\frac{1}{2}} = \frac{U^{k+1} - U^k}{\Delta t}, \qquad [\delta_{\Delta t} U]^{k+\frac{1}{2}} = \frac{U^{k+1} + U^k}{2},$$

it is straightforward to show that

$$\begin{aligned} \Delta t \left(\mathbb{M}(\mathbf{v})[\partial_{\Delta t^2} U]^k,\, [\partial_{2\Delta t} U]^k\right) &= \frac{1}{2}\left(\mathbb{M}(\mathbf{v})\left([\partial_{\Delta t} U]^{k+\frac{1}{2}} - [\partial_{\Delta t} U]^{k-\frac{1}{2}}\right),\, \left([\partial_{\Delta t} U]^{k+\frac{1}{2}} + [\partial_{\Delta t} U]^{k-\frac{1}{2}}\right)\right) \\ &= \frac{1}{2}\left(\mathbb{M}(\mathbf{v}) V^{k+\frac{1}{2}},\, V^{k+\frac{1}{2}}\right) - \frac{1}{2}\left(\mathbb{M}(\mathbf{v}) V^{k-\frac{1}{2}},\, V^{k-\frac{1}{2}}\right), \end{aligned}$$

the last line being obtained thanks to the symmetry of matrix. $\mathbb{M}(\mathbf{v})$. In the same way, we have

$$\Delta t \left(\mathbb{K} U^k, [\partial_{2\Delta t} U]^k\right) = \frac{1}{2}\left(\mathbb{K} U^k, U^{k+1}\right) - \frac{1}{2}\left(\mathbb{K} U^k, U^{k-1}\right).$$

Now, define

$$\begin{cases} \mathcal{U}_a^k & = \left(\mathbb{B}(\mathbf{v})[\partial_{2\Delta t}U]^k,\, [\partial_{2\Delta t}U]^k\right) \geq 0 \\ \mathcal{E}^{k+\frac{1}{2}} & = \frac{1}{2}\left(\mathbb{M}(\mathbf{v})V^{k-\frac{1}{2}},\, V^{k-\frac{1}{2}}\right) + \frac{1}{2}\left(\mathbb{K}U^k, U^{k+1}\right), \end{cases}$$

we get from (2.49),

$$\mathcal{E}^{k+\frac{1}{2}} - \mathcal{E}^{k-\frac{1}{2}} + \Delta t\,\mathcal{U}_a^k = 0,\ k \geq 1,$$

and by a simple summation

$$\mathcal{E}^{k+\frac{1}{2}} + \sum_{q=1}^{k} \Delta t\,\mathcal{U}_a^q = \mathcal{E}^{\frac{1}{2}}. \tag{2.50}$$

At this point, the problem is that the potential energy $\frac{1}{2}\left(\mathbb{K}U^k, U^{k+1}\right)$ might be a negative term since vector U is not taken at the same instant. However, we can rewrite

$$U^{k+1} = [\delta_{\Delta t}U]^{k+\frac{1}{2}} + \frac{\Delta t}{2}[\partial_{\Delta t}U]^{k+\frac{1}{2}}, \quad U^k = [\delta_{\Delta t}U]^{k+\frac{1}{2}} - \frac{\Delta t}{2}[\partial_{\Delta t}U]^{k+\frac{1}{2}},$$

and so (due to the symmetry of Matrix $\mathbb{K}$),

$$\left(\mathbb{K}U^k, U^{k+1}\right) = \left(\mathbb{K}[\delta_{\Delta t}U]^{k+\frac{1}{2}}, [\delta_{\Delta t}U]^{k+\frac{1}{2}}\right) - \frac{\Delta t^2}{4}\left(\mathbb{K}[\partial_{\Delta t}U]^{k+\frac{1}{2}}, [\partial_{\Delta t}U]^{k+\frac{1}{2}}\right).$$

The energy term can be rewritten

$$\mathcal{E}^{k+\frac{1}{2}} = \frac{1}{2}\left(\tilde{\mathbb{M}}(\mathbf{v})V^{k-\frac{1}{2}},\, V^{k-\frac{1}{2}}\right) + \frac{1}{2}\left(\mathbb{K}\delta_{\Delta t}U^{k+\frac{1}{2}}, \delta_{\Delta t}U^{k+\frac{1}{2}}\right), \qquad \tilde{\mathbb{M}}(\mathbf{v}) = \mathbb{M}(\mathbf{v}) - \frac{\Delta t^2}{4}\mathbb{K}.$$

Our claim is that if this matrix is definite positive, i.e. $\frac{\Delta t^2}{4}\mathbb{K} < \mathbb{M}(\mathbf{v})$, then our explicit scheme is stable. We can get a crude approximation of this condition using (2.48) :

$$\frac{\Delta t^2}{4}(\mathbb{K}U, U) \leq \kappa\,\frac{\Delta t^2 v_+^2}{4h^2}\,\frac{h^2}{v_+^2}(U, U) \leq \frac{\kappa}{4}\frac{\Delta t^2 v_+^2}{h^2}(\mathbb{M}(\mathbf{v})U, U),$$

with v^+ the maximum of the components of $\mathbf{v}$. Finally, there exists a pure constant $\alpha^2 \leq \frac{\kappa}{4}$ such that when

$$C = \alpha^2\,\frac{\Delta t^2 v_+^2}{h^2} < 1 \quad (\textbf{CFL condition}), \tag{2.51}$$

then $\mathcal{E}^{k+\frac{1}{2}}$ is always positive and consequently (see (2.50) and use $\mathcal{U}_a^q > 0$)

$$0 \leq \mathcal{E}^{k+\frac{1}{2}} \leq \mathcal{E}^{\frac{1}{2}}.$$

It is easy to see that this *a priori* estimation implies that for each node n

$$\left(U_n^{k+1} - U_n^k\right)^2 \leq \frac{4v_+^2\Delta t^2}{h^2}\mathcal{E}^{\frac{1}{2}} \leq \frac{4}{\alpha^2}\mathcal{E}^{\frac{1}{2}},$$

and U_n^k can only grow linearly with k : the scheme is stable.

The estimates for the discretized problem with a source term is more involved and we fail to derive a simple estimate involving some norm of the source term in this case.

2.6.4 Accuracy: the dispersion effect

Analysis of the precision of such schemes are usually done for infinite grids (i.e. $\mathbf{\Omega} = \mathbb{R}^2$) and constant velocity. The scheme is simply

$$\frac{h^2}{c^2}\left[\partial^2_{\Delta t^2} U\right]^k + \mathbb{K}U^k = 0.$$

The scheme being invariant by time and spatial translations, there exists particular solutions of the form (discrete plane wave solution)

$$U^k = U^{(0)}, e^{i\omega k \Delta t}, \quad \text{with} \quad \frac{1}{h^2}\,\mathbb{K}U^\star = \frac{4\sin^2(\omega\frac{\Delta t}{2})}{c^2\Delta t^2}\,U^\star.$$

$U^\star$ is an eigenvector of matrix $\mathbb{K}$. Let $x_n = ih\hat{x} + jh\hat{z}$, we look for eigenvectors in the form

$$U_n^\star = e^{i\vec{k}_h(\omega)\cdot x_n} = e^{i(k_x^h(\omega)ih + k_h^y(\omega)jh)}.$$

Plugging this expression, we obtain the *dispersion relation* of the scheme

$$\frac{4\sin^2(\omega\frac{\Delta t}{2})}{c^2\Delta t^2} = \frac{1}{h^2}\hat{K}(k_x^h(\omega)h, k_y^h(\omega)h)\,; \tag{2.52}$$

$\hat{K}$ is the symbol of matrix $\mathbb{K}$. We skip its exact definition, but only retain that due to both the centered nature and the consistency of the scheme, we have

$$\frac{1}{h^2}\hat{K}(k_x^h(\omega)h, k_y^h(\omega)^2) = [k_x^h(\omega)^2 + k_y^h(\omega))]\,\left(1 + \mathcal{O}((k_x^h(\omega)h)^2) + \mathcal{O}((k_y^h(\omega)h)^2)\right).$$

Now the Taylor development of $\frac{\sin x}{x}$ gives immediately

$$\frac{\omega^2}{c^2} = [k_x^h(\omega)^2 + k_y^h(\omega))^2]\,\left(1 + \mathcal{O}((k_x^h(\omega)h)^2) + \mathcal{O}((k_y^h(\omega)h)^2) + \mathcal{O}((\omega\Delta t)^2)\right),$$

and the relation of dispersion $\frac{\omega^2}{c^2(k_x^2+k_y^2)} = 1$ is asymptotically recovered: the scheme is consistent. However, there is a second order error in space and time and this error is the cause of what is called the *Pollution effect*; as the wave propagates in the medium, the errors accumulate and this, all the more since the duration of the wave increases. It is usually mentioned that 10 points per wavelengh is enough to get a "precise" result, let us say $\epsilon = 0.1\%$ for some h. As the matter of fact, this result has been obtained in the seventies, when the computers only allows for a propagation over roughly 6 wavelengh. As the errors are second order in time and space, a propagation along 4×6 wavelengh would induce an error of 2ϵ. To get the same error we need to divide the space step h (as well as the time step for a constant CFL) by $\sqrt{2}$.

The analysis of the error of the numerical schemes for wave equations in heterogeneous medium has not been studied a lot. Usually, a thumb rule consists of choosing F_c, a cut frequency of the seismic source and take n_λ points per wavelenghts at this frequency and for the slowest velocity in the medium. The formula is

$$h = \frac{v^-}{F_c n_\lambda}. \tag{2.53}$$

The choice of n_λ depends, as we said above, on the size of the domain of propagation and on the required accuracy.

Note that for our inverse problem, we do not have a precise knowledge of the values of the velocities involved. The simplest way to control the dispersion effect is to find their values into a given and not too large interval.

2.6.5 The discrete functional and its gradient

At this point we have a process to compute an approximation of $p_h(\mathbf{x}, t)$ at nodes $\mathbf{x}_n$ of grid and time $k\Delta t$. Our discrete observed data set $\mathcal{D}_{h,\Delta t}$ are seismograms recorded at n_r receivers located on the grid

$$\mathbf{d}_{\text{obs}} : \{\mathbf{d}_{\text{obs}}(\mathbf{x}_{n_r}, k\Delta t),\, n_r \in \mathcal{N}_r \subset \mathcal{N}, k \in K_r \subset \{1, ..., K\}\,.\}\,, \tag{2.54}$$

The discrete predicted observables are defined as

$$\mathbf{d}_{\text{cal}}(\mathbf{v}) : \left\{P(\mathbf{v})^k_{n_r} = \mathcal{R}_{h,\Delta t} P(\mathbf{v});\, n_r \in \mathcal{N}_r,\; k \in K_r\right\}, \tag{2.55}$$

where $\mathcal{R}_{h,\Delta t} : [\mathbb{R}^{\#\mathcal{N}}]^K \to \mathcal{D}_{h,\Delta t}$ is the discrete extraction operator.

$$\mathcal{R}_{h,\Delta t} : P = (P^k_n)_{k=1,..,K, n\in\mathcal{N}} \to (P^k_n)_{k\in K_r, n\in\mathcal{N}_r} \tag{2.56}$$

The discrete functional is now

$$\mathcal{J}(\mathbf{v}) = \tilde{h}_{h,\Delta t}\left(\mathbf{d}_{\text{cal}} = \mathcal{R}_{h,\Delta t} P(\mathbf{v}); \mathbf{d}_{\text{obs}}\right).$$

To compute the first variation of the discrete functional $\mathcal{J}$, we move parameter from $\mathbf{v}$ to $\mathbf{v}+\delta\mathbf{v}$, the solution $P = P(\mathbf{v})$ of the direct problem move to $P + \delta P$ and to the first order we have

$$\delta\mathcal{J}(\mathbf{v}) = \sum_{k\in K_r}\sum_{n_r\in\mathcal{N}_r} \nabla_{\text{cal}} h(x_{n_r}, k\Delta t)\delta P^k_{n_r} = \sum_{k=1}^{K}\left(\mathcal{R}^T_{h,\Delta t}\nabla_{d_{\text{cal}}} h(\mathcal{R}_{h,\Delta t}P(\mathbf{v}); \mathbf{d}_{\text{obs}}),\, \delta P^k\right). \tag{2.57}$$

Now, to the first order we have

$$\begin{cases} \mathbb{M}(\mathbf{v})[\partial_{\Delta t^2}\delta P]^k + \mathbb{B}(\mathbf{v})[\partial_{2\Delta t}\delta P]^k + \mathbb{K}\delta P^k = -\,\delta\mathbb{M}(\mathbf{v})[\partial_{\Delta t^2} P]^k - \delta\mathbb{B}(\mathbf{v})[\partial_{2\Delta t} P]^k \\ \\ \delta P^1 = \delta P^0 = 0. \end{cases}$$

We take the scalar product of each of these equations by Λ^k and sum over $k = 1, ..., K$. Thanks to time discrete integrations by parts and using the null initial and final conditions, we get $\Lambda^K = \Lambda^{K-1} = \delta P^1 = \delta P^0 = 0$

$$\begin{cases} \displaystyle\sum_{k=1}^{K} \left(\mathbb{M}(\mathbf{v})[\partial_{\Delta t^2} \delta P]^k, \Lambda^k \right) = \sum_{k=1}^{K} \left(\mathbb{M}(\mathbf{v})[\partial_{\Delta t^2} \Lambda]^k, \delta P^k \right) \\ \displaystyle\sum_{k=1}^{K} \left(\mathbb{B}(\mathbf{v})[\partial_{2\Delta t} \delta P]^k, \Lambda^k \right) = -\sum_{k=1}^{K} \left(\mathbb{B}(\mathbf{v})[\partial_{2\Delta t} \delta \Lambda]^k, \delta P^k \right) \\ \displaystyle\sum_{k=1}^{K} \left(\mathbb{K} \delta P^k, \Lambda^k \right) = \sum_{k=1}^{K} \left(\mathbb{K} \Lambda^k, \delta P^k \right). \end{cases}$$

Adding the three equations and using the equations satisfied by Λ^k and δP^k, we get

$$-\sum_{k=1}^{K} \left(\delta \mathbb{M}(\mathbf{v})[\partial_{\Delta t^2} P]^k, \Lambda^k \right) - \sum_{k=1}^{K} \left(\delta \mathbb{M}(\mathbf{v})[\partial_{\Delta t^2} P]^k, \Lambda^k \right) = \sum_{k=1}^{K} \left(S_{ad}^k, \delta P^k \right). \tag{2.58}$$

Returning back to (2.57), we get

$$\delta \mathcal{J}_{h,\Delta t} = \sum_{k=1}^{K} \left(\delta \mathbb{M}(\mathbf{v})[\partial_{\Delta t^2} P]^k, \Lambda^k \right) + \left(\delta \mathbb{B}(\mathbf{v})[\partial_{2\Delta t} P]^k, \Lambda^k \right), \tag{2.59}$$

as soon as

$$(S_{ad}^k)_n = -\nabla_{\text{cal}} h(x_{n_r}, k\Delta t) \text{ for } k \in K,\ n \in \mathcal{N}_r, \text{ and 0 elsewhere.}$$

From this we deduce the expression of the gradient of $\mathcal{J}_{h,\Delta t}$,

$$\frac{\partial \mathcal{J}_{h,\Delta t}}{\partial v_c} = \sum_{k=1}^{K} \left(\frac{\partial \mathbb{M}(\mathbf{v})}{\partial v_c} [\partial_{\Delta t^2} P]^k, \Lambda^k \right) + \left(\frac{\partial \mathbb{B}(\mathbf{v})}{\partial v_c} [\partial_{2\Delta t} P]^k, \Lambda^k \right). \tag{2.60}$$

What we have done here is to compute the exact gradient of the discrete functional. It is nowadays well known that discretizing directly the gradient of the continuous functional given in (2.45) by some other consistent scheme leads to problems of coherency when using a solver.

From the very expressions of both diagonal matrices $\mathbb{M}$ and $\mathbb{B}$ (see (2.47)), it is seen that the two scalar products in (2.60) involve only a finite number of terms at each time step $k\Delta t$. The remaining problem is that P^k is computed forward in time while Λ^k is computed backward in time. Storing all the P^k's is memory consuming (solving the scheme only requires to store two successive instants). To overcome this the P^k's are first computed for $k = 1, .., K$, then we restart from the two last instants P^K and P^{K-1} and recompute backward in time the P^k and Λ^k, $k = K, K-1..., 1$; in order to avoid instability problems, it is nevertheless necessary to store all the P_n^k at node $n \in \mathcal{N}_{\Gamma_2}$ (i.e. on the absorbing boundary) since the boundary condition is only absorbant in the forward direction.

2.7 Algorithms and Computational framework

We have now all the ingredients for designing the algorithm for our Full Waveform Inversion method. A simple FWI workflow can be summarised by the flow chart of Figure 2.1.

For a given seismic source and an initial model:

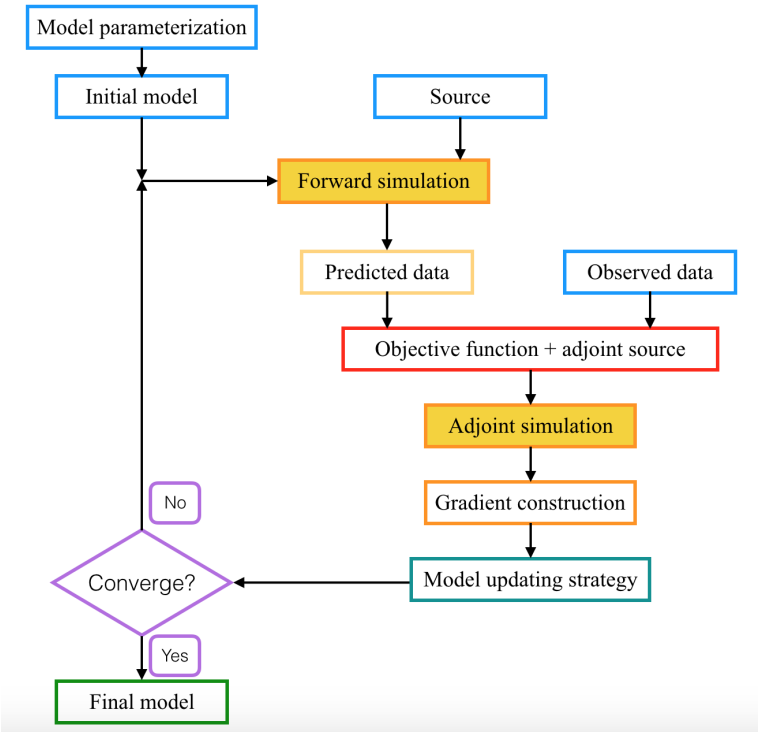


Figure 2.1: Flow chart of a simple full-waveform inversion workflow.

1. Solve numerically the forward problem defined by the state equations (2.1), together with appropriate initial and boundary conditions, to retrieve a physical realisation of the incident wavefield $\mathbf{u}^{\star}(\mathbf{m})$

2. Extract and map the physical realisation in the admissible state domain $\mathcal{W}$ to the receivers in the data domain $\mathcal{D}$ and compute the objective function (2.7)

3. Compute the gradient of the objective function with respect to the predicted observable and map it from the data domain to the admissible state domain to define the adjoint sources

4. Solve numerically the adjoint state equations (2.38), together with the complementary terminal conditions and boundary conditions, to retrieve the adjoint wave field

5. Compute the gradient of the misfit function with respect to the model parameters $\mathbf{m}$ using (2.19), which can be achieved by a local zero-lag cross correlation between the incident and the adjoint wave fields

6. Update the model parameters $\mathbf{m}$ using a gradient-based algorithm

7. Iterate until a prescribed convergence criterion based on the misfit function is achieved.

To implement this workflow, all the developments made during this thesis have been implemented within PySIT *Hewett et al.* (2013); *Hewett and Demanet* (2020)): the Python Seismic Imaging Toolbox. PySIT is an open Python-based platform that provides an efficient environment for incorporating new research on relatively short time scale and for accelerating the development and the prototyping of numerical methods in seismic imaging and full-waveform inversion.

In PySIT, the 2D acoustic wave forward and adjoint problems are solved using standard second-order finite differences in space and time on a uniform mesh in space together with a second-order explicit integration scheme in time with constant time step. Complex shifted frequency (CFS) PMLs are implemented following (*Sim*, 2010; *Grote and Sim*, 2011). Thus, there is slight differences from what we have presented in the previous subsections (first order absorbing conditions instead of `PML`, a different molecule for the discrete Laplacian,..) but all the ideas remain the same: we have a `CFL` condition, see (2.51), and the dispersion of the waves must be controlled using a large enough number of points per shortest wavelenghts (i.e. for the maximum frequency of the signal and the minimum velocity).

2.8 Least-squares misfit function and the cycle-skipping problem

Since the introduction of full-waveform inversion, the most commonly used objective function measures the L_2 norm difference between the synthetic observables and the observations.

As previously seen, in the data space $\mathcal{D}$, observations $\mathbf{d}_{\text{obs}}$ are common shot-gather of seismograms, often called seismic traces, recorded at n_r receivers for a source s:

$$\mathbf{d}^s_{\text{obs}} : \{\mathbf{d}^s_{\text{obs}}(\mathbf{x}_r, t),\ r = 1, \dots, n_r\}\ ,\ t \in [0, T] \tag{2.61}$$

where $\mathbf{x}_r$ denotes the spatial position vector of the receiver r

The predicted observables are defined accordingly for a physical realisation satisfying the state equation $\mathcal{L}[m]\mathbf{u} - \mathbf{f}^s$, together with the appropriate initial and boundary conditions

$$\mathbf{d}^s_{\text{cal}}(m) : \{\mathbf{d}^s_{\text{cal}}(\mathbf{x}_r, t; \mathbf{m}),\ r = 1, \dots, n_r\} = \mathcal{R}\mathbf{u}^s(\mathbf{x}, t; \mathbf{m});\ t \in [0, T] \tag{2.62}$$

where $\mathcal{R} : \mathcal{W} \to \mathcal{D}$ defined in (2.6) and maps the physical realisation onto the receiver positions $\mathbf{x}_r$.

The least-squares based objective function is defined as

$$\begin{aligned} \mathcal{J}(\mathbf{m}) = \tilde{h}(\mathbf{d}_{\text{cal}}(\mathbf{m}); \mathbf{d}_{\text{obs}}) &= \frac{1}{2} \sum_{s=1}^{n_s} \int_0^T |\mathbf{d}^s_{\text{cal}}(\mathbf{m}) - \mathbf{d}_{\text{obs}}|^2 \, dt \\ &= \frac{1}{2} \sum_{s=1}^{n_s} \int_0^T |\mathcal{R}\mathbf{u}^s(\mathbf{m}) - \mathbf{d}_{\text{obs}}|^2 \, dt \end{aligned} \tag{2.63}$$

where n_s is the number of seismic source used in the seismic survey.

Both seismic observations and predicted observables are oscillatory signals by nature. As such the least-squares based misfit function (2.63) is oscillatory and non convex in directions associated with wavenumber

components of the model that are longer than seismic wavelengths, *i.e.* longer than a half-period of the dominant inverted wavelet. For large kinematic errors, local optimisation algorithms will match incorrect phases between the predicted and observed signals, making the method prone to the so-called cycle-skipping effect, *i.e.* the inversion converges in a local minimum possibly far from the global minimum due to poor starting model. This can be illustrated in Figure2.2 with a very simple example where both the observed and predicted signals are sinusoidal signals of period T.

Apart from finding a good initial kinematic model, reformulating the FWI with alternative ways to measure the difference between predicted observable and observations to mitigate the non convexity with respect to time and space shifts has long been recognised in the FWI community as a proxy for the convexity of the FWI problem with respect to the model wave velocities (*Jannane et al.*, 1989).

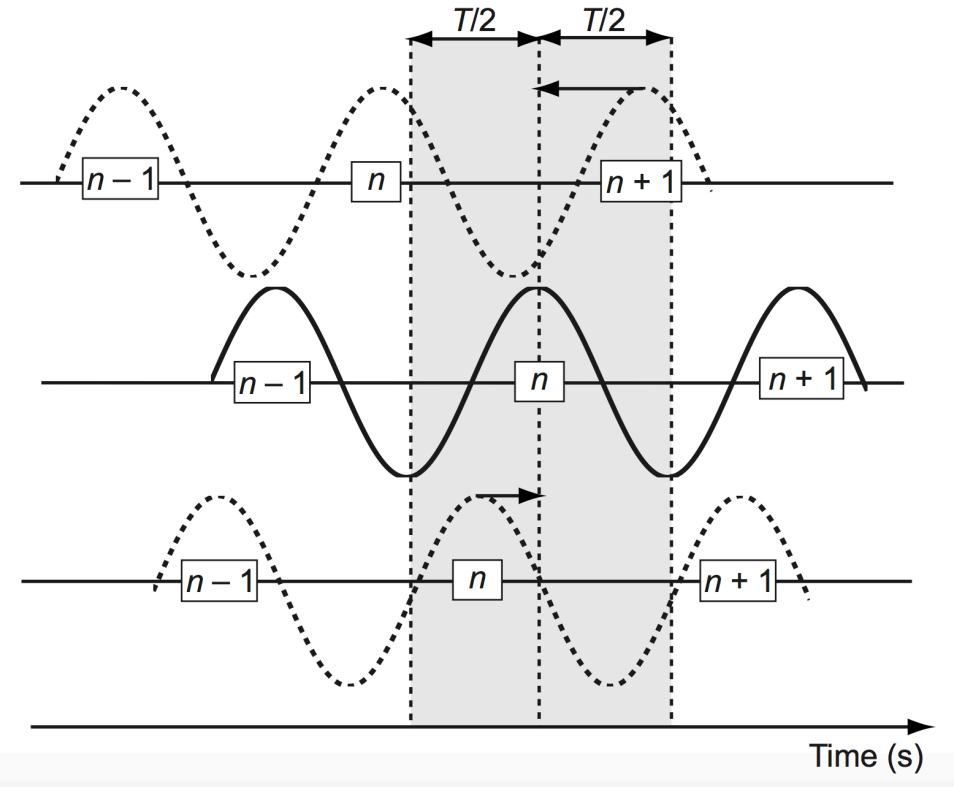


Figure 2.2: Simple example illustrating cycle-skipping artefact. The solid black line signal in the middle represents the observed sinusoidal signal with period T. The upper dashed line signal represents a predicted signal, which is simply the observed signal shifted by a time delay greater than, $T/2$. In this case the local optimisation problem will update the model so that the $n+1$-th phase of the predicted signal matches the n-th phase of the observed signal. The lower dashed line signal represents a predicted signal, which is the observed signal shifted by a time delay of less than $T/2$. In that case the local optimisation problem will rightly update the model such that the n-th phases of the observed and predicted signals match. From *Virieux and Operto* (2009).

The monotonicity of the optimal transport with respect to time-and-space shifts and dilations, together with improved robustness with respect to noise, has attracted the interest of the image processing and machine learnings. Today, there is a growing interest of the FWI community to extend the optimal transport in the context of seismic oscillatory signals for developing alternative continuous and differentiable objective functions. In the next section, a brief overview of the theoretical optimal transport framework of the optimal transport and of some of the current limitations in the context of FWI is provided.

Chapter 3

A brief overview of the Optimal Transport.

This chapter provides a very brief overview of optimal transport introducing the main notions needed in this thesis, comprehensive presentation of optimal Transport can be found for example in *Santambrogio* (2015) and *Peyré and Cuturi* (2019), and the references therein.

Optimal transport (OT), is an old problem originally formulated by *Monge* (1781) as finding the optimal way of moving soil from two places through a volume preserving transport in order to make embarkments. Assuming constant density and no mass creation or loss during the transport, the problem is to find a transport map that minimises the mechanical work , *i.e.* the mass $\mu(x)$ times the moving distance. The later being defined commonly as a L^p transport distance with $p \geq 1$.

3.1 Monge's problem

Let consider two probability measures: a "source" μ and a "target" ν that have a compact support $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}^d$, respectively, and the density defined with respect to the Lebesgue measure.

Mathematically, the problem is to find a *transport map* $\tilde{T}$ from X to Y such that

$$\mathrm{MP}(\mu,\nu): \qquad \tilde{T} = \arg\inf_{T\in\mathcal{M}} \int_X \|\mathbf{x} - T(\mathbf{x})\|^p \, d\mu(x), \tag{3.1}$$

where the displacement cost, *i.e. the ground cost*, is here defined as the L^p distance, while value function in (3.1) is called the *transport cost*.

Let denote $\mathcal{M} := \{T : X \to Y | T_{\#}\mu = \nu\}$ the set of all measure preserving maps between μ and ν with $T_{\#}\mu$ the pushforward of the measure μ, defined as:

$$[T_{\#}\mu](B) = \mu(T^{-1}(B)), \qquad \text{For any measurable subset } B \in Y$$

The Monge formulation (3.1) defines a non-convex optimisation problem with non linear constraints. However when, for example, μ and ν have densities, then the optimal transport map $T^\star$ exist a give a natural interpolation between the two measures. In particular for the L^p distance, the map $T_t(\mathbf{x}) = (1-t)\mathbf{x} + t\,T^\star(\mathbf{x})$ describes the path of particle $\mathbf{x}$ and furthermore the measure of μ pushed forward by T_t is the geodesic, *i.e.*

shortest path for the OT induced metric, between μ and ν.

If the map is smooth and the measures absolutely continuous, the constraint can be written as an equation for the Jacobian of T :

$$\mu(\mathbf{x}) = \nu\left(T(\mathbf{x})\right)\ \left|\det \nabla T(\mathbf{x})\right| \tag{3.2}$$

3.2 Monge Ampère and semi-discrete formulation

Since Brenier's theorem *Brenier* (1991) it is known that the quadratic cost OT problem, i.e. $p = 2$, has nice properties: under very general assumptions on the densities μ and ν, there exists a unique transport map, which is characterised as the gradient of a convex potential $\varphi(x) : \tilde{T}(\mathbf{x}) = \nabla\varphi(\mathbf{x})$.

Using this result in (3.2) yields the Monge-Ampère form of Monge problem :

$$\mu(\mathbf{x}) = \nu\left(\nabla\varphi(\mathbf{x})\right)\left|\det \nabla^2\varphi(\mathbf{x})\right| \qquad \nabla\varphi(X) \subset Y \tag{3.3}$$

where φ is a convex function.

When $\nu(\mathbf{x}) := \mu(\mathbf{x} - \tau)$, the convex potential is given analytically by $\varphi(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}^2\| - \tau \cdot \mathbf{x} + C$ (this is unique up to a constant). Plugging this in (3.1) the transport cost is $\|\tau\|^2$, i.e quadratic with respect to the shift τ. We will illustrate this property in chapter 5.

Several numerical methods have been proposed to discretise and solve this problem:

- A first class of methods (*Benamou et al.*, 2014; *Benamou and Duval*, 2018) is based on the discretisation of X using cartesian grids and optimised finite difference stencils. The measure μ is approximated as a discrete measure prescribed on the grid but ν requires a given analytic function or an oracle based on interpolation of ν, evaluated at $\nabla\varphi(\mathbf{x})$ in Y. This is the approach followed in particular by Engquist and collaborators (*Engquist and Froese*, 2014; *Engquist et al.*, 2016; *Yang et al.*, 2018) in the context of full-waveform inversion.

- A second class of methods (*Mérigot and Thibert*, 2021; *Lévy and Schwindt*, 2018) relies on a semi-discrete approach, in which (3.3) is solved from a dual point of view. The measure ν is assumed to be an arbitrary empirical measure, *i.e.* a weighted point cloud, and a continuous oracle is still needed but only for μ.

Both approaches lead to a non-linear set of equations that can be solved with a quasi-newton solver in $O(N \log N)$ operations, N being the number of points of the discrete measure, but with a problem-dependent constant that may be large.

3.2.1 Kantorovich Formulation

When both probability measures μ and ν have discrete support, *e.g.* as in numerical approximations, rearranging the mass using a one-to-one map T may become impossible, see Figure 3.1. For example, there

is no transport map T transforming a single Dirac measure into the sum of two Dirac measures.

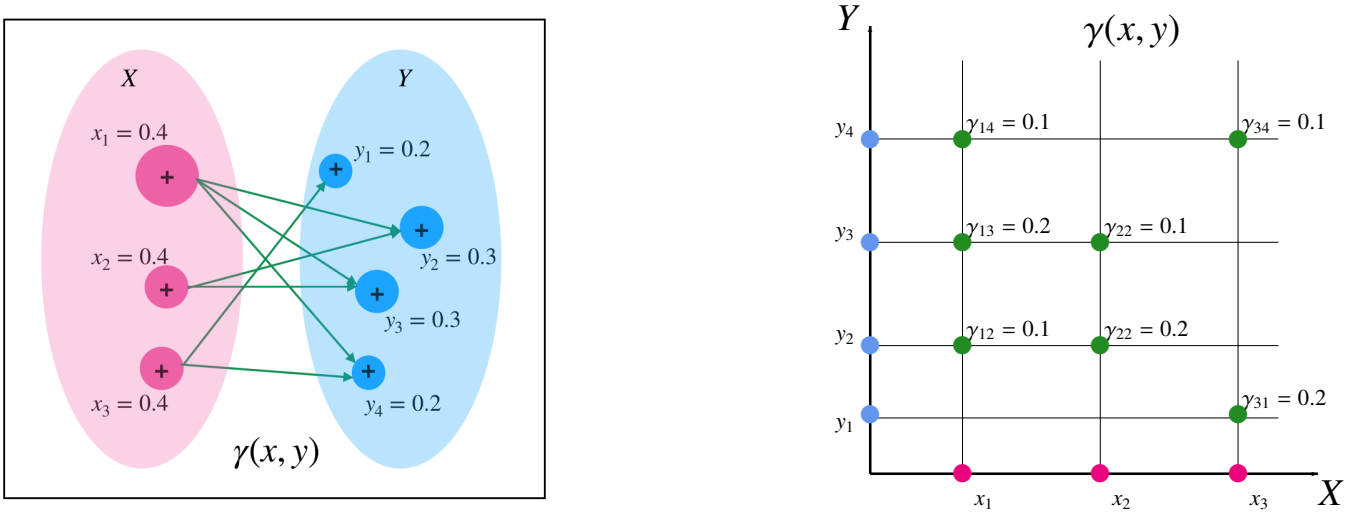


Figure 3.1: Kantorovich's transport plan between two discrete distributions of masses defined on spaces X and Y, allowing to distribute mass from the one location to multiple locations.

Kantorovich (1942) suggested to optimise instead a relaxed functional with respect to the *transport plans*:

$$\mathrm{KP}(\mu, \nu) = \inf_{\gamma \in \Pi(\mu,\nu)} \int_{X \times Y} \|\mathbf{x} - \mathbf{y}\|^p \; d\gamma(\mathbf{x}, \mathbf{y}). \tag{3.4}$$

The set of admissible transport plans

$$\Pi(\mu, \nu) = \{\gamma \in \mathcal{P}(X \times Y) \mid P^X_{\#}\gamma = \mu, \; P^Y_{\#}\gamma = \nu\} \tag{3.5}$$

is the set of probability measures on the product space $X \times Y$, also called "couplings", with prescribed marginal laws μ and ν, and $P^X : X \times Y \to X$ and $P^Y : X \times Y \to Y$ stand for the projections onto each space.

Transport plans still transport the μ-mass repartition to the ν-mass repartition but the mass at location $\mathbf{x}$ can be split and sent to different locations $\mathbf{y}$, see blueFigure 3.1 for illustration. The cost is again the total work, *i.e.* mass times ground cost.

For the L^p ground cost, $p > 1$, the Monge and the Kantorovich formulations are equivalent. They both define the same distance on the set of probability measures: the Wasserstein distance denoted $W_p^p(\mu, \nu)$,

$$W_p^p(\mu, \nu) = \inf_{\gamma \in \Pi(\mu,\nu)} \int_{X \times Y} \|\mathbf{x} - \mathbf{y}\|^p \; d\gamma(\mathbf{x}, \mathbf{y}) = \inf_{T \in \mathcal{M}} \int_X \|\mathbf{x} - T(\mathbf{x})\|^p \; \mu(\mathbf{x}) \, d\mathbf{x} \tag{3.6}$$

and, when μ has continuous densities, the support of any optimal transport plan $\gamma^\star$ is contained on the graph of a function $T^\star$. In particular this implies $\gamma^\star(A, B) = \mu\left(\{\mathbf{x} : \mathbf{x} \in A, T^\star(\mathbf{x}) \in B\}\right)$ and furthermore that the optimal plane defines a mapping between μ and ν.

The dual problem associated to (3.6) is formulated as

$$W_p^p(\mu,\nu) = \sup_{\substack{\phi\in C(X),\,\psi\in C(Y)\ s.t.\\ \phi(\mathbf{x})+\psi(\mathbf{y})\le\|\mathbf{x}-\mathbf{y}\|^p,\,\forall(\mathbf{x},\mathbf{y})\in X\times Y.}} \int_X \phi(\mathbf{x})\,d\mu(\mathbf{x}) + \int_Y \psi(\mathbf{y})\,d\nu(\mathbf{y}). \tag{3.7}$$

When freezing ν and considering the Wasserstein distance as a function of μ, the Wasserstein distance is the supremum of affine functions μ parameterised by (ϕ,ψ), and as such is convex with respect to μ. The dual formulation (3.7) leads to a convex, linear optimisation problem with linear constraints.

Under the primal form (3.6), the optimal transport is defined for probability measures, while the dual form (3.7) admits a notion of first variation over the space of general mesures, provided they have the same total mass, see *Santambrogio* (2015) section 7.2 for details.

3.2.2 Discrete Kantorovich problem

When X and Y are discrete, *i.e.* $X=\{\mathbf{x}_i, i=1,...,N\}, Y=\{\mathbf{y}_j, j=1,...,N\}$ where the $\mathbf{x}_i$ and $\mathbf{y}_j$ are in $\mathbb{R}^d$, μ and ν can be written as the sums of Dirac masses:

$$\mu(\mathbf{x}) = \sum_{i=1}^N \mu_i \delta_{\mathbf{x}_i} \quad \text{and} \quad \nu(\mathbf{y}) = \sum_{j=1}^N \nu_j \delta_{\mathbf{y}_j}.$$

The weights μ_i and ν_j have to satisfy the *positivity* and the *total mass balance* constraints:

$$\begin{cases} \mu_i > 0, \quad \nu_j > 0, \\ \sum_i \mu_i = 1\,, \quad \sum_j \nu_j = 1\,. \end{cases} \tag{3.8}$$

The set of admissible plans becomes:

$$\Pi(\mu,\nu) = \left\{ \gamma\in\mathcal{S}_{N,N} : \sum_{j=1}^N \gamma_{ij} = \mu_i, \quad \sum_{i=1}^N \gamma_{ij} = \nu_j, \quad \forall i,j \right\}$$

with

$$\mathcal{S}_{N,N} = \left\{ \gamma_{ij}\in\mathbb{R}_+^{N\times N} \ : \ \sum_{i=1}^N\sum_{j=1}^N \gamma_{ij} = 1 \right\}$$

The marginal constraints are given by summing over lines and columns the matrix γ.

Finally the optimisation problem becomes:

$$\min_{\gamma\in\Pi(\mu,\nu)} \sum_{i=1}^N\sum_{j=1}^N c_{ij}\gamma_{ij}, \quad c_{ij} = |x_i - y_j|^p \tag{3.9}$$

where c_{ij} is the ground cost between x_i and y_j.

Let $\langle \cdot, \cdot \rangle$ denote the L^2 scalar product on $\mathbb{R}^{N \times N}$, (3.9) can be written in short form,

$$\mathrm{OT}(\mu, \nu) = \min_{\gamma \in \Pi(\mu,\nu)} \langle \gamma, c \rangle. \tag{3.10}$$

This is a linear optimisation problem with N^2 unknowns and $2 \times N$ linear constraints. The best known linear programming methods have cubic complexity, see remark 3.3 in *Peyré and Cuturi* (2019), and cannot be practically used for realistic seismic applications.

The dual formulation associated to (3.7) yields

$$\mathrm{OT}(\mu, \nu) = \max_{\substack{u \in \mathbb{R}^N, v \in \mathbb{R}^N \\ u_i + v_j \leq c_{ij}}} \sum_i u_i \, \mu_i + \sum_j v_j \, \nu_j \tag{3.11}$$

In the context of this thesis, the discretisation (x_i, y_j) is assumed to be static, and defined by the receiver positions and by the time sampling of the acquisition, and (μ, ν) are vectors in $\mathbb{R}^N$ with coefficients $(\mu_i)s$ and $(\nu_j)s$.

The convexity of the map $\mu \in \mathbb{R}^N \mapsto \mathrm{OT}(\mu, \nu)$ again follows as the supremum of affine functions, and Danskin's theorem provides the gradient with respect to μ

$$\frac{\partial}{\partial \mu_i} \mathrm{OT}(\mu, \nu) = u_i^\star, \quad \frac{\partial}{\partial \nu_j} \mathrm{OT}(\mu, \nu) = v_j^\star, \ \forall \, i, j \tag{3.12}$$

where $(u^\star, v^\star)$ are the maximisers of (3.11).

The W_1 distance

The case $p = 1$ is special. The Kantorovitch metric is well defined but the Monge problem has no unique solution (*Kantorovich and Rubinshtein*, 1957; *Villani*, 2009).

Assuming $X = Y$, the dual formulation, a particular instance of the Kantorowitch-Rubinstein theorem, *e.g.* *Santambrogio* (2015) section 3.2.1, yields

$$W_1(\mu, \nu) \quad = \sup_{\mathrm{Lip}(\psi) \leq 1} \int_X \psi(\mathbf{x}) \, d(\mu - \nu). \tag{3.13}$$

where $\mathrm{Lip}(\psi)$ denotes the minimal Lipschitz constant for ψ.
The 1-Lipschitz functions ψ, for the ground cost associated to the ℓ^1 distance on $\mathbb{R}^d$, satisfy

$$\psi : X \to \mathbb{R}, \ |\psi(\mathbf{x}) - \psi(\mathbf{y})| \leq |\mathbf{x} - \mathbf{y}|, \quad \forall (\mathbf{x}, \mathbf{y}) \in X \times X$$

with

$$|\mathbf{x} - \mathbf{y}| = \sum_{i=1}^{d} |x_i - y_i|, \qquad \forall (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^d$$

The Kantorovitch problem can be generalised when the total mass between μ and ν is not conserved, leading

to the generalised $\tilde{W}_1$ distance

$$W_1(\mu,\nu) \quad = \sup_{\substack{\mathrm{Lip}(\psi)\leq 1 s.t. \\ \|\psi\|_\infty \leq C}} \int_X \psi(\mathbf{x})\, d(\mu-\nu). \tag{3.14}$$

which is a particular instance of the KR norm (*Bogachev*, 2007) defined in the space of Radon measures. This can also be seen as a generalisation of the L^1 norm (*Lellmann et al.*, 2014).

The generalised $\tilde{W}_1$ metric, also referred to as the KR distance, was first introduced in the context of full-waveform inversion by Métivier and collaborators (*Métivier et al.*, 2016a,b; *Métivier et al.*, 2016).

One advantage of the KR norm in the context of FWI is that there is no need to normalise the data to be positive and mass balanced. The KR norm has no direct connection with optimal transport when μ and ν are no more required to be probability measures.

The problem (3.14) leads to a convex optimisation problem under linear constraints that can be solved efficiently using proximal splitting methods (*Combettes and Pesquet*, 2011), reducing significantly the computational cost compared to the 2-Wasserstein metric.

3.2.3 Optimal Transport distances in the context of full-waveform inversion

Classical optimal transport defines distances W_p^p, $p>1$ on the set of probability measures, assuming

$$\mu>0,\ \nu>0, \qquad \text{and} \quad \int_X d\mu = \int_Y d\nu.$$

where μ and ν have for support $X\in\mathbb{R}^d$ and $Y\in\mathbb{R}^d$, respectively.

In the context of full-waveform inversion, the observables d_{ob} and d_{cal} are oscillatory signals in time, recorded at receiver positions over a time window. As such the positivity assumption breaks down and the total mass conservation needs to be carefully checked in practice.

Apart from data normalisation, there exists a significant amount of work from the mathematics community to generalise optimal transport to what is called the unbalanced optimal transport, either by penalising the marginal constraints (*Benamou*, 2003; *Piccoli and Rossi*, 2014, 2016; *Lombardi and Maitre*, 2015; *Gangbo et al.*, 2019) or by rigorously defining an "unbalanced" distance on the set of positive Radon measures (*Chizat et al.*, 2018b; *Chizar et al.*, 2018; *Kondratyev et al.*, 2016b; *Liero et al.*, 2018). The later will be considered in connexion with the entropic approximation of optimal transport and the Sinkhorn divergence formulation.

Optimal transport and oscillatory signals

In the mathematics community (*Ambrosio et al.*, 2011; *Mainini*, 2012) different approach have been explored to extend optimal transport to signed measure. One approach leads to split the predicted d_{cal} and the observed d_{obs} signals into a positive and a negative part and to recombine them as

$$\tilde{d}_{\mathrm{cal}} = d^+_{\mathrm{cal}} + d^-_{\mathrm{obs}} \qquad \tilde{d}_{\mathrm{obs}} = d^+_{\mathrm{obs}} + d^-_{\mathrm{cal}} \tag{3.15}$$

This amounts to compute the optimal transport distance between the positive and negative parts of $d_{\text{obs}} - d_{\text{cal}}$ leading to the objective function $h(d_{\text{cal}}, d_{\text{obs}} = W_2^2(\tilde{d}_{\text{cal}}, \tilde{d}_{\text{obs}})$. This type of transformation suffers however from a convexity loss with respect to translation and from a loss of sensibility as translation increases (*Métivier et al.*, 2018).

Another approach (*Engquist and Froese*, 2014; *Engquist et al.*, 2016) with regard to the positivity assumption is to split the predicted d_{cal} and the observed d_{obs} signals into a positive and a negative part, and to scale them by their respective total mass. The objective function $h(d_{\text{cal}}, d_{\text{obs}})$ is defined as the sum of the associated Wasserstein distances

$$h(d_{\text{cal}}, d_{\text{obs}}) = W_2^2\left(\frac{d_{\text{cal}}^+}{\langle d_{\text{cal}}^+ \rangle}, \frac{d_{\text{cal}}^-}{\langle d_{\text{cal}}^- \rangle}\right) + W_2^2\left(\frac{d_{\text{obs}}^+}{\langle d_{\text{obs}}^+ \rangle}, \frac{d_{\text{obs}}^-}{\langle d_{\text{obs}}^- \rangle}\right) \tag{3.16}$$

where $d^+ = \max\{d, 0\}$ and $d^- = \max\{-d, 0\}$.

This strategy assures convexity, but provides no guarantee that the positive part, respectively the negative part, of the predicted signal has the same total mass than the positive part, respectively the negative part, of the observed signal. Moreover such a transformation is not differentiable, which is problematic when computing the Fréchet derivatives of the objective function in the context of local gradient-based optimisation methods.

Signal transformation into probability distribution

Another approach to achieve data positivity and mass balance is to directly transform the seismic data into probability densities by linear or nonlinear scaling functions (*Qiu et al.*, 2017; *Yang and Engquist*, 2017, 2018; *Engquist and Yang*, 2021). The objective function is now defined as $h(d_{\text{cal}}, d_{\text{obs}}) = W_2^2(\tilde{d}_{\text{cal}}, \tilde{d}_{\text{obs}})$

A first type of transformation is the affine transformation

$$\tilde{d}_{\text{cal}} = \frac{d_{\text{cal}} + b}{\langle d_{\text{cal}} + b \rangle}, \quad \tilde{d}_{\text{obs}} = \frac{d_{\text{obs}} + b}{\langle d_{\text{obs}} + b \rangle}, \qquad b > 0 \tag{3.17}$$

where b is chosen such that $\tilde{d}_{\text{cal}}$ and $\tilde{d}_{\text{obs}}$ are both positive.

This kind of transformation is straightforward and ensures the total mass conservation. The convexity of the objective with respect to translation is however affected under such transformation as the constant b leads to the creation of artificial mass in the optimal transport.

A second class of transformation is the exponential transformation

$$\tilde{d}_{\text{cal}} = \frac{\exp(b\, d_{\text{cal}})}{\langle \exp(b\, d_{\text{cal}}) \rangle}, \quad \tilde{d}_{\text{obs}} = \frac{\exp(b\, d_{\text{obs}})}{\langle \exp(b\, d_{\text{obs}}) \rangle}, \qquad b > 0 \tag{3.18}$$

or the more stable Softplus version of the scaling,

$$\tilde{d}_{\text{cal}} = \frac{\log(\exp(b\, \tilde{d}_{\text{cal}}) + 1)}{\langle \log(\exp(b\, \tilde{d}_{\text{cal}}) + 1) \rangle}, \quad \tilde{d}_{\text{cal}} = \frac{\log(\exp(b\, \tilde{d}_{\text{cal}}) + 1)}{\langle \log(\exp(b\, \tilde{d}_{\text{cal}}) + 1) \rangle}, \qquad b > 0 \tag{3.19}$$

Under such transformations the influence of the high-positive values and of the high-negative values are not

the same. To mitigate this effect, it has been proposed to add in (3.18) and (3.19) the same transformation applied $-d_{\text{cal}}$ and $-d_{\text{obs}}$, leading to symmetric transformations.

Other type of transformations originally proposed in the context of FWI least-squares objective function by (*Liu et al.*, 2012; *Donno et al.*, 2013) are of relevance for optimal transport. They consider the normalised cumulative distributions associated to positive transforms of d_{cal} and d_{obs},

$$Q_{\text{cal}}(\mathbf{x}_r,t)=\frac{\int_0^t \tilde{d}_{\text{calc}}(\mathbf{x}_r,t)\,dt}{\int_0^T \tilde{d}_{\text{calc}}(\mathbf{x}_r,t)\,dt},\qquad Q_{\text{obs}}(\mathbf{x}_r,t)=\frac{\int_0^t \tilde{d}_{\text{obs}}(\mathbf{x}_r,t)\,dt}{\int_0^T \tilde{d}_{\text{obs}}(\mathbf{x}_r,t)\,dt} \tag{3.20}$$

where

$$\tilde{d}=\begin{cases}|d|\,, & [\text{absolute value of the signal}]\,,\\ d^2, & [\text{square of the signal}]\,,\\ \mathrm{E}(d), & [\text{enveloppe of the signal}]\end{cases} \tag{3.21}$$

Other strategies for full-waveform inversion

Seemingly uncorrelated with the previously mentioned strategies, Métivier et al (*Métivier et al.*, 2016a,b; *Métivier et al.*, 2016) developed an approach based on a generalised dual W_1 distance (3.14), see Section 3.2.2.

This is a particular instance of the KR norm defined on the space of signed measures (*Bogachev*, 2007; *Lellmann et al.*, 2014) that is lifted by adding a simple bound constraint on the dual potential to deal with unbalanced mass transport. As such there is no need to normalised the data to be positive and mass balanced.

The resulting objective function is differentiable and allows for simultaneous comparison of 2D and 3D shot gathers following the numerical strategy of *Métivier et al.* (2016a); *Métivier et al.* (2016). This strategy was shown in *Métivier et al.* (2018) to be equivalent to the one proposed by *Mainini* (2012) in the case of the 1-Wasserstein distance.

However the KR norm has no direct connection with optimal transport once we no longer require d_{obs} and d_{cal} to be probability measures (*Vershik*, 2013), and the convexity is not guaranteed with respect to large translation.

A more recent strategy introduced by *Métivier et al.* (2018); *Metivier et al.* (2019) is linked to a graph-space optimal transport distance, following the original developments of transport L^p distance for signal analysis by *Thorpe et al.* (2017).

After discretisation in time, the predicted d_{cal} and the observed d_{obs} signals are mapped into the 2D time-amplitude graph space as point clouds that are compared rather than the signals themselves. The graph-space optimal transport distance can be efficiently computed as a linear assignment problem.This signal graph representation ensures positivity.

This new objective function shows promising convexity properties with respect to translation and amplitude variation. This has been demonstrated for field data applications (*Górszczyk et al.*, 2020, 2021; *Pladys*

et al., 2021). However extension to 2D and 3D source gathers remain to be evaluated as it might increase significantly the computational cost.

Chapter 4

Entropic regularization of OT and its generalizations

A new class of numerical methods based on "entropic regularisation" has been introduced for OT computations in *Cuturi* (2013) (see also the book *Peyré and Cuturi* (2019)). It penalises $\mathrm{OT}(\mu, \nu)$ with the Shannon entropy of the transport plan. A small parameter, often denoted as ε, controls this penalisation.

This research topic has been very active since for two reasons: the method comes with a simple computational method called the Sinkhorn algorithm for which there are now countless variants and acceleration techniques; the class of problems it can applied to is very flexible and includes in particular the recent extension of OT to positive radon measure called "unbalanced" optimal transport (*Chizat et al.*, 2018a; *Chizar et al.*, 2018; *Liero et al.*, 2018). See section 4.2. We discuss its use in connection with positive/negative mass splitting in chapter 5.

Two problems arise when using entropic OT as a proxy for the classical OT distance or unbalanced OT distance. First, the numerical stability and cost of Sinkhorn algorithm deteriorates when ε goes to 0. Second, the entropic OT cost is not a mathematical distance on the set of probability measure anymore. A simple variant called Sinkhorn divergence (*Genevay et al.*, 2018; *Séjourné et al.*, 2019) fixes this issues. It is presented in section 4.3 and this is the misfit used in chapter 5.

This chapter follows *Vialard* (2019) and *Peyré and Cuturi* (2019), and adopts a discrete formulation of the Kantorovitch problem, see section 3.2.2, unless otherwise specified.

4.1 Entropic regularisation and Sinkhorn algorithm

4.1.1 Entropic regularisation

This section is presented in the discrete Kantorovich setting, section 3.2.2. The entropic regularisation of (3.4)-(3.5) has been introduced in the OT context by *Cuturi* (2013), the Shannon entropy is defined as

$$\mathrm{Ent}(\gamma) \overset{\text{def.}}{=} -\sum_{i,j} \gamma_{ij}(\log(\gamma_{ij}) - 1), \qquad \forall (i,j) \in [1,n] \times [1,m] \tag{4.1}$$

The Entropic regularised problem becomes

$$\mathrm{OT}_\varepsilon(\mu,\nu) = \min_{\gamma\in\Pi(\mu,\nu)} \langle\gamma, c\rangle - \varepsilon\,\mathrm{Ent}(\gamma), \tag{4.2}$$

where $\varepsilon > 0$ is a parameter, and c is the ground cost, with the associated pairwise cost matrix $c_{ij} = c(\mathbf{x}_i, \mathbf{x}_j)$ evaluated on the support of μ and ν. We will assume (as is the case in FWI) that μ and ν have the same support, and will restrict to the quadratic ground cost (i.e. $p = 2$)

$$c_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|^2 \tag{4.3}$$

Later, in the context of FWI, $\mathbf{x} = (t, x)$ will be a point in time t and offset x.

The problem (4.2) is an ε-strongly convex program with a unique optimal solution $\gamma^\star_\varepsilon$. The solution $\gamma^\star_\varepsilon$ converges to the solution with maximal entropy within the set of all optimal solutions of the Kantorovitch problem, *i.e.* $\mathrm{OT}_\varepsilon \xrightarrow{\varepsilon\to 0} \mathrm{OT}$.

Problem (4.2) can be reformulated as a projection onto the transport plans of the Gibbs kernel associated to the ground cost with respect to the Kullback-Leibler divergence

$$\begin{aligned}\mathrm{OT}_\varepsilon(\mu,\nu) &= \min_{\gamma\in\Pi(\mu,\nu)} \langle\gamma, c\rangle + \varepsilon\sum_{i,j}\gamma_{ij}(\log(\gamma_{ij}) - 1),\\ &= \min_{\gamma\in\Pi(\mu,\nu)} \varepsilon\sum_{i,j}\mathcal{KL}\left(\gamma_{ij} \mid K^\varepsilon_{ij}\right).\end{aligned} \tag{4.4}$$

where $K^\varepsilon_{ij} \overset{\text{def.}}{=} \exp\{-c_{ij}/\varepsilon\}$, is the Gibbs kernel associated to the ground cost, and the Kullback-Leibler divergence is defined as

$$\mathcal{KL}(s \mid t) \overset{\text{def.}}{=} \begin{cases} s\,log(\frac{s}{t}) - s & \text{for}(s,t) > 0,\\ +\infty & \text{for } t = 0\end{cases} \tag{4.5}$$

The Kullback–Leibler divergence or relative entropy takes two arguments s and t, and is a particular instance of statistical divergences that has nice analytical, and computational properties. As illustrated in Figure 4.1, it is strictly convex with its minimum at $s = t$, *e.g.* $t = 1$ in this case.

A key insight is that, as ε increases, the entropic part of the cost function favours mass spreading and smooths out small scale transport of the non-regularised OT. The optimal coupling becomes less and less sparse. While this has the effect of both accelerating computational algorithms and promoting faster convergence, it introduces a penalisation bias in the transport that tends to destroy the distance properties. In particular, $\mathrm{OT}_\varepsilon(\mu,\mu) > 0$ and μ is no more necessarily the minimum of $\nu \mapsto \mathrm{OT}_\varepsilon(\mu,\nu)$ as it may become cheaper to diffuse from μ than not moving. This is illustrated by the 1D example in Figure 4.3.

Remark on the reference measure In order to simplify the presentation of this chapter, we used the penalization $-Ent(\gamma_{ij}) = \mathcal{KL}(\gamma_{ij}|1)$. This is the relative entropy with respect to the uniform measure 1 (Lebesgue). A standard practice is to use $\mathcal{KL}(\gamma_{ij}|\mu_i\,\nu_j)$ which yields a smaller value for the the penalization

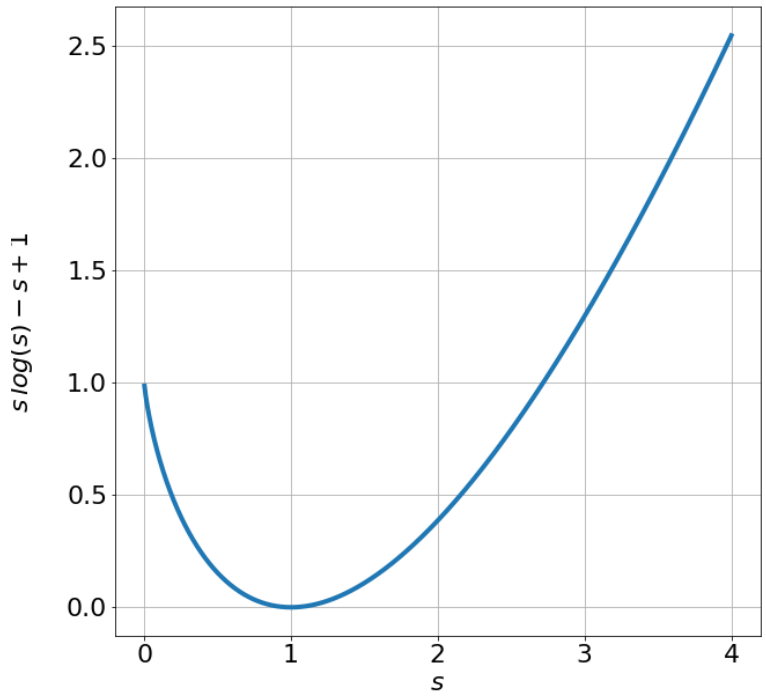


Figure 4.1: KL function. $s \mapsto \mathcal{KL}(s|1) + 1$.

and also a more accurate (in the sense of the non entropic problem) support of the entropic plan. All formulations and algorithms are easily extended to this more general case.

4.1.2 Sinkhorn algorithm

Derivation

Problem (4.2) is a (strictly) convex program with linear (marginal) constraints. The Entropy function $x \mapsto x\,\log(x)$ has infinite derivative at $x = 0$ and acts as an optimization barrier, lifting the positivity constraint on γ_{ij}. The solution of the problem (4.2) has a specific form, which can be parameterised using $n + m$ variables. That parameterisation is essentially dual as γ_{ij} has $n \times m$ variables but $n + m$ constraints.

Problem (4.2) can be recast, $\forall (i,j) \in [1,n] \times [1,m]$, as a saddle point problem:

$$\begin{aligned} \inf_{\gamma_{ij}} \sup_{u_i, v_j} L(\gamma_{ij}, u_i, v_j) := \gamma_{ij}\, c_{ij} \;+\; \varepsilon \sum_{i,j} \gamma_{ij}(\log(\gamma_{ij}) - 1) \\ -\; u_i \,(\sum_j \gamma_{ij} - \mu_i) - \; v_j \,(\sum_i \gamma_{ij} - \nu_j)\,, \end{aligned} \tag{4.6}$$

The Lagrangian L is convex in γ, and the first variation in γ_{ij} yields:

$$c_{ij} \;+\; \varepsilon \log(\gamma_{ij}) - u_i - v_j = 0, \qquad \forall (i,j) \in [1,n] \times [1,m] \tag{4.7}$$

where $\{u_i, i = 1, \ldots, n\}$ and $\{v_j, j = 1, \ldots, m\}$ are the Lagrangian multipliers associated to the n-constraints

$\sum_j \gamma_{ij} = \mu_i$ and the m-constraints $\sum_i \gamma_{ij} = \nu_j$, respectively.

The unique optimal coupling (plan) for entropic regularisation has the specific form

$$\gamma_{ij} = \exp\left\{\frac{u_i + v_j - c_{ij}}{\varepsilon}\right\} = a_i\, K_{ij}^{\varepsilon}\, b_j, \quad \forall (i,j) \in [1,n] \times [1,m] \tag{4.8}$$

where

$$K_{ij}^{\varepsilon} = \exp\left\{-\frac{c_{ij}}{\varepsilon}\right\}; \quad a_i = \exp\left\{\frac{u_i}{\varepsilon}\right\}; \quad b_j = \exp\left\{\frac{v_j}{\varepsilon}\right\}. \tag{4.9}$$

with $\mathbf{a}$ and $\mathbf{b}$ being non negative vectors.

Using the marginal constraints in $\Pi(\mu,\nu)$, it leads to the nonlinear system

$$a_i \left(\sum_{j=1}^{m} \mathbf{K}_{ij}^{\varepsilon} b_j\right) = \mu_i, \quad \text{and} \quad b_j \left(\sum_{i=1}^{n} \mathbf{K}_{ji}^{\varepsilon} a_i\right) = \nu_j \tag{4.10}$$

which involves two unknowns a_i and b_j

$$a_i = \frac{\mu_i}{\sum_{j=1}^{m} K_{ij}^{\varepsilon}\, b_j} \quad \text{and} \quad b_j = \frac{\nu_j}{\sum_{i=1}^{n} K_{ji}^{\varepsilon}\, a_i} \tag{4.11}$$

or in compact form ($\mathbf{K}^{\varepsilon} = (K_{ij}^{\varepsilon})_{(i,j)\in[1,n]\times[1,m]}$. is now a matrix)

$$\mathbf{a} \odot (\mathbf{K}^{\varepsilon}\, \mathbf{b}) = \boldsymbol{\mu}, \quad \text{and} \quad \mathbf{b} \odot \left(\mathbf{K}^{\varepsilon\, T}\, \mathbf{a}\right) = \boldsymbol{\nu} \tag{4.12}$$

where $\odot$ corresponds to the entry side multiplications of vectors, and the unknowns $\mathbf{a}$ and $\mathbf{b}$ are

$$\mathbf{a} = \boldsymbol{\mu} \oslash (\mathbf{K}^{\varepsilon}\mathbf{b}), \quad \text{and} \quad \mathbf{b} = \boldsymbol{\nu} \oslash (\mathbf{K}^{\varepsilon\, T}\mathbf{a}) \tag{4.13}$$

where $\oslash$ corresponds to the entry side divisions of vectors.

That problem is known in the numerical analysis community as the matrix scaling problem, see *Nemirovski and Rothblum* (1999) and references therein. An intuitive way is to solve it iteratively. The Sinkhorn algorithm (*Sinkhorn and Knopp*, 1967) constructs iteratively the solution

$$\gamma^{(l+1)} = \mathbf{a}^{(l+1)}\, \mathbf{K}^{\varepsilon}\, \mathbf{b}^{(l)}, \qquad \gamma_{ij}^{(l+1)} = a_i^{(l+1)}\, K_{ij}^{\varepsilon}\, b_j^{(l+1)}. \tag{4.14}$$

where $\mathbf{a}^{(l+1)} = \boldsymbol{\mu} \oslash (\mathbf{K}\mathbf{b}^{(l)})$ and $\mathbf{b}^{(l+1)} = \nu \oslash (\mathbf{K}^T\mathbf{a}^{(l+1)})$,

$$a_i^{(l+1)} = \frac{\mu_i}{\sum_{j=1}^{m} K_{ij}^{\varepsilon}\, b_j^{(l)}} \quad \text{and} \quad b_j^{(l+1)} = \frac{\nu_j}{\sum_{i=1}^{n} K_{ji}^{\varepsilon}\, b_i^{(l+1)}} \tag{4.15}$$

initialised with an arbitrary vector $\mathbf{b}^0 = \mathbb{1}_m$, where $\mathbb{1}_m$ is the m-vector of ones..

The convergence error at iteration l can be monitored using the residual in the Kullback-Leibler sense :

$$\mathrm{ERR}_{\mathcal{KL}}^{(l)} = \sum_{i=1}^{n} \mathcal{KL}\left(a_i^{(l+1)} \sum_{j=1}^{m} K_{ij}^{\varepsilon}\, b_j^{(l)} \,\middle|\, \mu_i\right) + \sum_{j=1}^{m} \mathcal{KL}\left(b_j^{(l)+1} \sum_{i=1}^{n} a_i^{(l+1)} K_{ji}^{\varepsilon} \,\middle|\, \nu_j\right). \tag{4.16}$$

Alternatively, the residual of the Kantorovich potentials in sup norm between successive iterations can also be used

$$\mathrm{ERR}_{\infty}^{(l)} = \varepsilon \left[\left\| \log\left(\frac{a_i^{(l+1)}}{a_i^{(l)}}\right) \right\|_{\infty} + \left\| \log\left(\frac{b_j^{(l+1)}}{b_j^{(l)}}\right) \right\|_{\infty} \right] \tag{4.17}$$

The Sinkhorn algorithm converges linearly, *e.g.* see *Peyré and Cuturi* (2019) Remark 4.13, but the asymptotic convergence rate deteriorates when $\varepsilon \to 0$. In the case of quadratic optimal transport (4.3) the algorithm behaves as $\mathcal{O}(1-\varepsilon)$. An example borrowed from *Peyré and Cuturi* (2019), and reproduced in Figure 4.2, illustrates this effect.

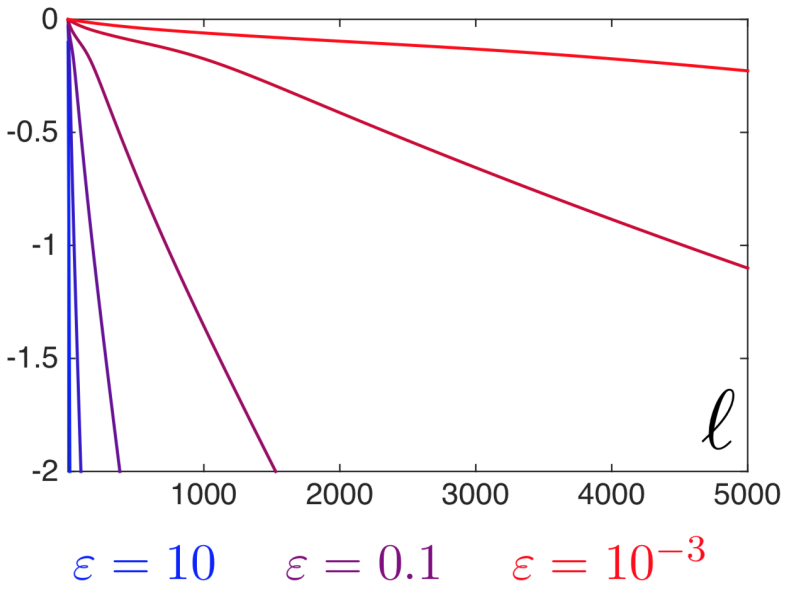


Figure 4.2: Influence of the regularization parameter ε on the convergence rate for (4.16) of Sinkhorn's algorithm (source: *Peyré and Cuturi* (2019))

It is also known, *e.g.* *Léonard* (2014) proposition (13), that when $\varepsilon \to 0$, the unique minimiser $\gamma_{\varepsilon}^{\star}$ of $\mathrm{OT}_{\varepsilon}(\mu,\nu)$ converges to the maximal entropy plan among the possible optimal transport plans of $\mathrm{OT}(\mu,\nu)$.

An example borrowed from *Benamou et al.* (2015), and reproduced Figure 4.3, illustrates the behaviour of the optimal transport plan when ε decreases. The optimal entropic plan $\gamma_{\varepsilon}^{\star}$ converges to the unregularized optimal transport plan solution of (3.10) and concentrates on the graph of the optimal transport map. The middle line shows how the mass is progressively shifted away from the diagonal during the Sinkhorn iterations for a fixed ε.

Numerical stability

When implementing and using the Sinkhorn algorithm it is important to keep in mind that the Kernel $\mathbf{K}^{\varepsilon}$ scales like $\exp -\tau^d/\varepsilon$ where τ is the scale of the ground transport and d the dimension of the physical space,

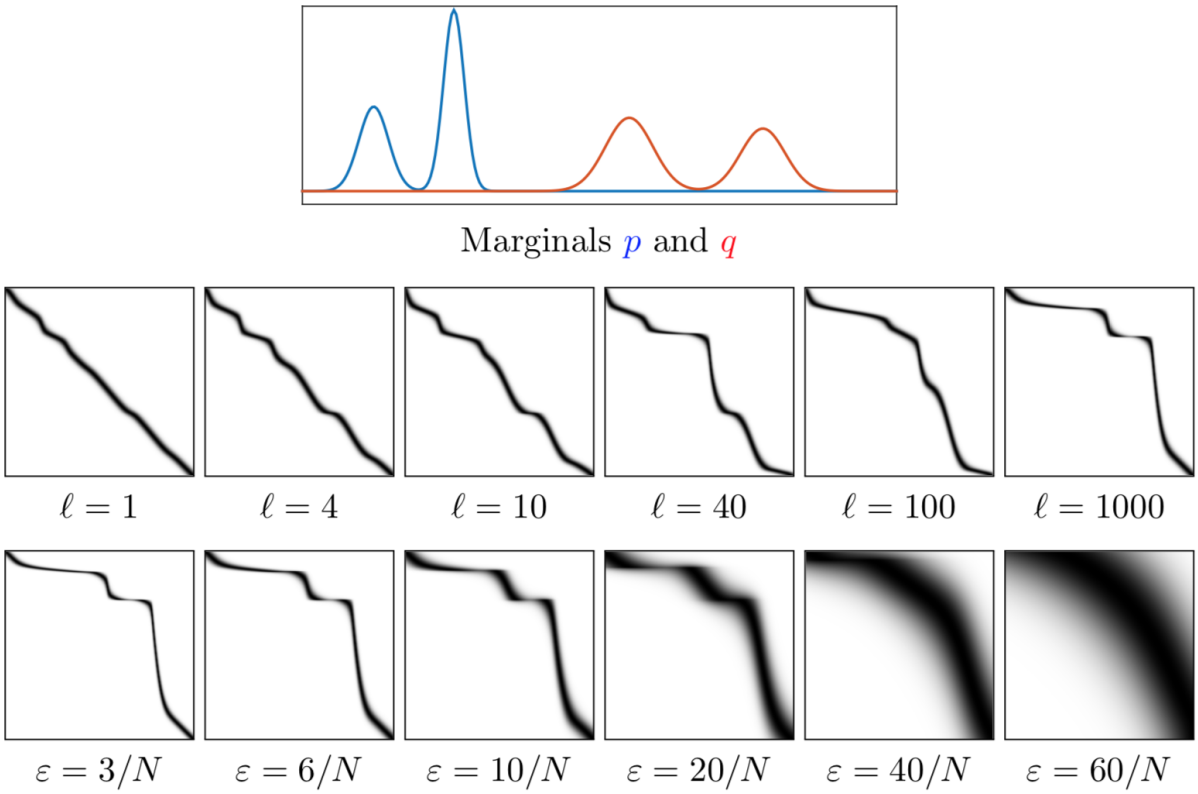


Figure 4.3: (top) The input densities p (blue) and q (curve) as functions of x and y. (center) Evolution of the couplings γ^l at iteration l of the Sinkhorn algorithm. (bottom) solution $\gamma_\varepsilon^\star$ of problem (4.4) for several values of ε. (source: *Benamou et al.* (2015)). The coupling are plotted in grey level in $x \times y$ space.

which depends on the data μ and ν. The numerical stability therefore depends on these parameters, and the convergence of the Sinkhorn algorithm deteriorates when $\varepsilon \to 0$.

The Sinkhorn algorithm often fail to terminate as soon as some of the elements of the kernel $\mathbf{K}^\varepsilon$ become too negligible for the machine precision. This can then result in a matrix product $\mathbf{K}^\varepsilon \mathbf{b}$ or $\mathbf{K}^{\varepsilon\,T}\mathbf{a}$ with very small entries with respect to the machine precision, resulting in a division by 0 numerically in the Sinkhorn update (4.15). Such issues can be partly resolved by carrying out computations in the log domain or using a technique called ε-scaling, *e.g.* *Schmitzer* (2019).

Cost

Denoting by N the number of points used in the discretisation of the marginal densities,*i.e.* $\{\mathbf{a}_1, \ldots, \mathbf{a}_N\}$ and $\{\mathbf{b}_1, \ldots, \mathbf{b}_N\}$ a naive Sinkhorn algorithm involves a $\mathcal{O}\left(n_{iter} \times (n \times m)^2\right)$ complexity. where n_{iter} is the number of Sinkhorn iterations to converge, which would be prohibitive in practical FWI applications. For small ε the sparsity of the Kernel can be used to reduce the cost, *e.g.* *Schmitzer* (2019).

Speed-up for separable kernels

As discussed in *Chizat* (2017) and in *Peyré and Cuturi* (2019) Remark 4.15, an important particular case, relevant in FWI applications, for which the complexity of each Sinkhorn iteration can be significantly reduced is when each index i and j considered in the ground cost-matrix c_{ij} can be described as a d-uple taken in

the cartesian product of d finite sets $[1, n_1], \ldots, [1, n_d]$.

$$i = (i_k)_{k=1}^d, j = (j_k)_{k=1}^d \in [1, n_1] \times \ldots \times [1, n_d]$$

When the ground cost is additive along the sub-indices, namely there exists d matrices $\mathbf{C}^1, \ldots, \mathbf{C}^d$, each of respective size $n_1 \times n_1, \ldots, n_d \times n_d$ such that

$$c_{ij} = \sum_{k=1}^{d} c^k_{i_k j_k},$$

then the kernel appearing in the Sinkhorn iterations has a separable multiplicative structure

$$K_{ij} = \Pi_{k=1}^d K^k_{i_k j_k}$$

leading to fast and exact matrix-vector multiplication, speeding-up the Sinkhorn algorithm.

Let illustrate the Kernel separability in the context of FWI. For 2D horizontal shot-gathers, the predicted $\mathbf{d}^{\text{cal}}(x_i, t_j) = d^{\text{cal}}_{ij}$ and the observed $\mathbf{d}^{\text{obs}}(x_j, t_j) = d^{\text{obs}}_{ij}$ are discrete realisations at the $i = 1, \ldots, n_r$ receiver spatial positions and the $j = 1, \ldots, n_t$ time samples, which define on a 2D grid in the space and time domain.

As such $\mathbf{d}^{\text{cal}}$ and $\mathbf{d}^{\text{obs}}$ are instantiated as $n_r \times n_t$ matrices, and as a consequence the Lagrange multiplier $\mathbf{u}$ and $\mathbf{v}$. For the quadratic cost, $c_{iljm} = |x_i - x_l|^2 + |t_j - t_m|^2$, the associated kernel matrix yields

$$K^{\varepsilon}_{ilmj} = \exp\left(-\frac{|x_i - x_l|^2 + |t_m - t_j|^2}{\varepsilon}\right) = K^{\mathbf{x}}_{il}\, K^t_{mj} \tag{4.18}$$

where $K^x_{il} = \exp\left(-|x_i - x_l|^2/\varepsilon\right)$ and $K^t_{mj} = \exp\left(-|t_m - t_j|^2/\varepsilon\right)$ are $n_r \times n_r$ and $n_t \times n_t$ convolution matrices defined on the receiver and time finite sets, respectively. In this case a multiplication by $\mathbf{K}^{\varepsilon}$ can be carried out more efficiently.

Such a separable multiplicative structure allows for a fast and exact multiplication by $\mathbf{K}$ when applying each 1-D convolution matrix along "slice" of the 2D shot-gather grid. Let rewrite $\mathbf{a}$ and $\mathbf{b}$ as $\mathbf{A}$ and $\mathbf{B}$, respectively, to emphasise the fact that the multipliers are reshaped as $n_r \times n_t$ matrices. Then computing $\mathbf{K}\,\mathbf{B}$, which would naively require $(n_r \times n_t)^2$ operations with a naive implementation, can be obtained by applying two 1-D convolutions separately, as

$$\left(\mathbf{K}^x \left(\mathbf{K}^t \mathbf{B}\right)^T\right)^T = \mathbf{K}^x\, \mathbf{B}\, \mathbf{K}^t$$

to recover a $n_r \times n_t$ matrix in $\left(n_r^2\, n_t + n_r\, n_t^2\right)$ operations instead of $n_r^2\, n_t^2$ operations.

The computation above is for $d = 2$, this is the time dimension plus a 1D offset line. For $d = 3$, *e.g.* a 2D dimension offset, one needs to apply these very same 1-D convolutions to each slice of $\mathbf{B}$, reshaped

as a tensor of suitable size. The relative gain is even higher at $\left(n_{r_1}^2\, n_t + n_{r_2}^2\, n_t + (n_{r_1} + n_{r_2})\, n_t^2\right)$ operations instead of $n_{r_1}^2\, n_{r_2}^2\, n_t^2$.

In order to get a simplified view of the tensorization gain, let us assume $n_t \simeq n_r := N^{1/d}$ where N is the total number of points in time $\times$ offset. Then the cost of one Sinkhorn iteration is $\mathcal{O}(N^{1+1/d})$ in place of $\mathcal{O}(N^2)$. Of course one still has to iterate Sinkhorn until convergence.

Acceleration by the Successive Over-Relaxation (SOR)

Since the Sinkhorn algorithm is a fixed-point algorithm, standard linear or even non linear relaxation schemes to enhance the conditioning of the fixed-point mapping near the solution and improve the linear convergence rate. A Successive Over-Relaxation (SOR) algorithm, with an acceleration parameter $\theta \in [1, 2]$ is presented in *Chizat* (2017), which yields.

$$\begin{cases} a_i^{(l+1)} = (a_i^{(l)})^{1-\theta} \left(\dfrac{\mu_i}{\sum_{j=1}^m K_{ij}^\varepsilon\, b_j^{(l)}} \right)^\theta \\ b_j^{(l+1)} = (b_j^{(l)})^{1-\theta} \left(\dfrac{\nu_j}{\sum_{i=1}^n K_{ji}^\varepsilon\, b_i^{(l+1)}} \right)^\theta \end{cases} \tag{4.19}$$

Numerical experiments in *Chizat* (2017) illustrate that the convergence rate can be improved by up to orders of magnitude when using SOR, and a detailed convergence analysis is provided, which explicit the best choice of the acceleration parameter θ. All the experiments in thesis have been performed for $\theta = 1.4$.

4.1.3 Computation of the Gradient

From (4.6), convex duality yields

$$\mathrm{OT}_\varepsilon(\mu_i, \nu_j) = \sup_{(u_i, v_j)} = \sum_i u_i\, \mu_i + \sum_j v_j\, \nu_j - \varepsilon \sum_{ij} a_i\, K_\varepsilon(x_i, y_j)\, b_j. \tag{4.20}$$

As in the non entropic case the dual formulation shows that the map $\mu \in \mathbb{R}^N \mapsto \mathrm{OT}_\varepsilon(\mu, \nu)$ is the supremum of linear functional, hence convex. The gradient is easily deduced as in (3.12), $\forall (i, j) \in [1, N] \times [1, N]$

$$\frac{\partial}{\partial \mu_i} \mathrm{OT}_\varepsilon(\mu, \nu) = u_i^\star = \varepsilon\, \log a_i^\star; \quad \frac{\partial}{\partial \nu_j} \mathrm{OT}_\varepsilon(\mu, \nu) = v_j^\star = \varepsilon\, \log b_j^\star \tag{4.21}$$

where $(u^\star, v^\star)$ are the maximisers of the entropic dual problem (4.20). .

4.2 Unbalanced Optimal Transport

In order to deal with seismic unbalanced data we want to drop the hard marginal constraints in $\Pi(\mu, \nu)$. Recently, a new metric on positive Radon measures has been independently introduced in *Chizat et al.* (2018b), *Kondratyev et al.* (2016b) and *Liero et al.* (2018).

It allows mass to be transported and also created or destroyed, but at a cost. The "good" additional cost turns out to be the Kullback-Leibler divergence between the data and the marginal of the plan. The Sinkhorn algorithm is very easily adapted to this distance that eliminates the total mass balance constraint.

4.2.1 Penalisation of the marginal constraints

The hard marginal constraints are replaced by the following soft penalisations:

$$\mathrm{OT}_\lambda(\mu,\nu) = \min_\gamma \sum_{i=1}^N \sum_{j=1}^N c_{ij}\gamma_{ij} + \lambda_1 \sum_{i=1}^N \mathcal{KL}\left(\sum_{j=1}^N \gamma_{ij} \,\middle|\, \mu_i\right) + \lambda_2 \sum_{j=1}^N \mathcal{KL}\left(\sum_{i=1}^N \gamma_{ij} \,\middle|\, \nu_j\right) \tag{4.22}$$

where $\mathcal{KL}$ is the Kullback-Leibler divergence defined in (4.5). The $\lambda_1 > 0$ and $\lambda_2 > 0$ are parameters quantifying the cost of unbalanced mass relaxation. If μ and ν are balanced, the classical optimal transport can be recovered by letting $\lambda_{1,2} \to +\infty$.

A nice interpretation of (4.22) is that, depending on the value of λ, available mass on both side will be either transported or just created or destroyed locally if it is cheaper through the $\mathcal{KL}$ penalization of the marginal constraint.

In the case of the quadratic ground cost (4.3), the value function OT_λ is called the Gaussian Hellinger-Kantorovich(GHK) distance, *Liero et al.* (2018).

4.2.2 Entropic regularised unbalanced OT problem

Entropic regularisation can also be applied to (4.22) as in (4.4), which yields:

$$\mathrm{OT}_{\varepsilon,\lambda}(\mu,\nu) = \min_{\gamma\in\Pi(\mu,\nu)} \varepsilon\,\mathcal{KL}\left(\gamma \,\middle|\, K_\varepsilon\right) + \lambda_1 \sum_{i=1}^N \mathcal{KL}\left(\sum_{j=1}^N \gamma_{ij} \,\middle|\, \mu_i\right) + \lambda_2 \sum_{j=1}^N \mathcal{KL}\left(\sum_{i=1}^N \gamma_{ij} \,\middle|\, \nu_j\right). \tag{4.23}$$

The dual formulation of the entropic regularised unbalanced optimal transport problem in discrete form follows again from convex duality

$$\begin{aligned} OT_{\varepsilon,\lambda}(\mu_i,\nu_j) = \sup_{(u_i,v_j)} & -\lambda_1 \sum_i \left(\exp\left\{-\frac{u_i}{\lambda_1}\right\} - 1\right) \mu_i - \lambda_2 \sum_j \left(\exp\left\{-\frac{v_j}{\lambda_2}\right\} - 1\right) \nu_j \\ & - \varepsilon \sum_{ij} a_i\, K_\varepsilon(x_i,y_j)\, b_j. \end{aligned} \tag{4.24}$$

As explained in *Chizat et al.* (2018b) the optimal plans are again in the form (4.8) and can be computed iteratively using the modified Sinkhorn iterations :

$$\begin{cases} a_i^{l+1} = \Big(\dfrac{\mu_i}{\sum_j (K_\varepsilon(x_i, y_j) b_j^l}\Big)^{\dfrac{\lambda_1}{\lambda_1 + \varepsilon}}, \ \forall i, \\ b_j^{l+1} = \Big(\dfrac{\nu_j}{\sum_i K_\varepsilon^T(x_i, y_j)\, a_i^{l+1}}\Big)^{\dfrac{\lambda_2}{\lambda_2 + \varepsilon}}, \ \forall j. \end{cases} \tag{4.25}$$

As before $a_i = \exp\{u_i/\varepsilon\}$ and $b_j = \exp\{v_j/\varepsilon\}$ and the Sinkhorn algorithm obeys to the same convergence properties. In practice we use the error (4.17) to monitor the convergence.

The dual formulation (4.24) shows again that the map $\mu_i \mapsto \mathrm{OT}_{\varepsilon,\lambda}(\mu_i, \nu_j)$ is concave. Applying Danskin's theorem once more the differentials of $\mathrm{OT}_{\varepsilon,\lambda}(\mu_i, \nu_j)$ with respect to μ_i and ν_i can be directly obtained as the differentials with respect to μ_i and ν_i of the right hand side of (4.24) for the optimal $(u_i^\star, v_i^\star)$s:

$$\begin{aligned} \frac{\partial}{\partial \mu_i} \mathrm{OT}_{\varepsilon,\lambda}(\mu, \nu) &= -\lambda_1 \left(\exp\left\{ \frac{-u_i^\star}{\lambda_1} \right\} - 1 \right) = -\lambda_1 \left(a_i^{-\varepsilon/\lambda_1} - 1 \right) \\ \frac{\partial}{\partial \nu_j} \mathrm{OT}_{\varepsilon,\lambda}(\mu, \nu) &= -\lambda_2 \left(\exp\left\{ \frac{-v_j^\star}{\lambda_1} \right\} - 1 \right) = -\lambda_2 \left(b_i^{-\varepsilon/\lambda_2} - 1 \right). \end{aligned} \tag{4.26}$$

4.3 Sinkhorn divergence

4.3.1 Sinkhorn divergence for balanced OT problems

As previously seen in Section 4.1, the entropic regularised optimal transport problem (4.4) can be efficiently solved with the Sinkhorn algorithm as long as the regularisation parameter ε is large enough. The larger the ε is, the faster the convergence of the iterative Sinkhorn algorithm. The entropic regularisation however introduces a bias increasing with ε.

In the continuous setting asymptotic results by *Conforti and Tamanini* (2021) and *Pal* (2019) on OT_ε when $\varepsilon \to 0$, reminding that $\mathrm{OT}_\varepsilon \to \mathrm{OT}$ when $\varepsilon \to 0$, offer a better understanding of the entropic bias

$$\mathrm{OT}_\varepsilon(\mu, \nu) - \mathrm{OT}(\mu, \nu) \simeq 2\,\varepsilon\sqrt{2\pi\,\varepsilon} + \varepsilon\,\left[\mathcal{KL}\left(\mu \mid \nu\right) + \mathcal{KL}\left(\nu \mid \nu\right)\right] + \mathcal{O}(\varepsilon^2). \tag{4.27}$$

This result holds for measures (μ, ν) which have smooth densities with respect to the Lebesgue measure, still denoted here as μ, ν by abuse of notation.

The definition of the $\mathcal{KL}$ divergence in (4.5) is also an abuse of notation,

$$\mathcal{KL}\left(\mu \mid \nu\right) \overset{\text{def.}}{=} \begin{cases} \displaystyle\int \mu(x) \left(\log\left\{ \frac{\mu(\mathbf{x})}{\nu(\mathbf{x})} \right\} - 1 \right) dx & \text{if } \mu \text{ absolutely continuous w.r.t } \nu. \\ +\infty & \text{otherwise} \end{cases} \tag{4.28}$$

It is worth to note that the densities μ and ν are decoupled in the first-order term. A simple idea to remove

these terms is to use the symmetric entropic OT between the marginals and to remove them, keeping in mind that $\mathrm{OT}(\mu,\mu)=0$

$$\mathrm{S}_\varepsilon(\mu,\nu)=\mathrm{OT}_\varepsilon(\mu,\nu)-\frac{1}{2}\left\{\mathrm{OT}_\varepsilon(\mu,\mu)+\mathrm{OT}_\varepsilon(\nu,\nu)\right\} \tag{4.29}$$

which is, formally at least, a second order approximation of $\mathrm{OT}(\mu,\nu)$ with respect to ε.

This formulation has been studied under the name "Sinkhorn divergence" in *Genevay et al.* (2018); *Feydy et al.* (2019). It does not define a distance but has nice properties. We immediately see that $\mathrm{S}_\varepsilon(\mu,\mu)=0$ for any $\varepsilon>0$, a property that was lost for OT_ε. It is also proven in *Séjourné et al.* (2019); *Feydy et al.* (2019) that it is symmetric in μ and ν, and remains positive and convex with respect to μ and ν. It also metrises the weak convergence of measures.

The Sinkhorn divergence (4.29) can be computed as well as its gradient just by applying the Sinkhorn algorithm three times. For small ε, The Sinkhorn divergence S_ε for the quadratic ground cost is therefore a better approximation of $\mathrm{OT}_0=W_2^2$ than OT_ε, inheriting the good properties of Entropic Optimal Transport: strict convexity, smoothness and the Sinkhorn computational algorithm.

4.3.2 Sinkhorn divergence for unbalanced OT problem

The Sinkhorn divergence has a natural extension to the unbalanced case (*Séjourné et al.*, 2019), yielding

$$\mathrm{S}_{\varepsilon,\lambda}(\mu,\nu)=\mathrm{OT}_{\varepsilon,\lambda}(\mu,\nu)-\frac{1}{2}\left\{\mathrm{OT}_{\varepsilon,\lambda}(\mu,\mu)+\mathrm{OT}_{\varepsilon,\lambda}(\nu,\nu)\right\}+\frac{\varepsilon}{2}\left(m(\mu)-m(\nu)\right)^2. \tag{4.30}$$

where $\mathrm{OT}_{\varepsilon,\lambda}(\mu,\nu)$ is the entropic regularised unbalanced optimal transport problem defined in (4.23), while $m(\mu)=\sum_i \mu_i$ and $m(\nu)=\sum_j \nu_j$ are the total mass of μ and ν, respectively, which can be different.

It is proved in *Séjourné et al.* (2019) that $\mathrm{S}_{\varepsilon,\lambda}(\mu,\nu)$ is again positive, definite, and convex. A comparison between $OT_{\varepsilon,\lambda}(\mu,\nu)$ with $S_{\varepsilon,\lambda}(\mu,\nu)$ is illustrated using $1D$ wavelets in the next chapter.

4.3.3 Computing the unbalanced Sinkhorn divergence and its gradient

In order to solve problem (4.30), let first look at $\mathrm{OT}_{\varepsilon,\lambda}(\mu,\nu)$. Following section 4.2.2, $\mathrm{OT}_{\varepsilon,\lambda}(\mu,\nu)$ can be solved as a fixed-point problem with the Sinkhorn iterations 4.25. For the symmetric case $\mathrm{OT}_{\varepsilon,\lambda}(\mu,\mu)$, the Sinkhorn algorithm simplifies *Séjourné et al.* (2019).

The dual problem becomes a concave maximisation problem, symmetric with respect to a unique unknown vector u_i:

$$OT_{\varepsilon,\lambda}(\mu_i,\mu_j)=\sup_{(u_i}-2\lambda\sum_i\left(\exp\left\{-\frac{u_i}{\lambda_1}\right\}-1\right)\mu_i-\varepsilon\sum_{ij}a_i\,K_\varepsilon(x_i,y_j)\,U_j. \tag{4.31}$$

where $a_i=\exp\left(-u_i/\varepsilon\right)$. At iteration $l+1$, the optimal vector a_i^{l+1} can be computed as

$$a_i^{l+1}=\left(\frac{\mu_i}{\sum_j K_\varepsilon(x_i,x_j)a_j^l}\right)^{\frac{\lambda}{\lambda+\varepsilon}} \tag{4.32}$$

As in (4.26), the first variation of $\mathrm{OT}_{\varepsilon,\lambda}(\mu,\mu)$ with respect to μ can be directly obtained from the dual formula (4.31):

$$\frac{\partial}{\partial \mu_i}\mathrm{OT}_{\varepsilon,\lambda}(\mu,\mu) = -2\,\lambda\left(a_i^{-\varepsilon/\lambda}-1\right). \tag{4.33}$$

The Sinkhorn divergence $\mathrm{S}_{\varepsilon,\lambda}$ can be computed using the Sinkhorn iterations (4.25) to solve the standard problem $\mathrm{OT}_{\varepsilon,\lambda}(\mu,\nu)$ and using the Sinkhorn iterations (4.32) to solve the symmetric problem $\mathrm{OT}_{\varepsilon,\lambda}(\mu,\mu)$ and $\mathrm{OT}_{\varepsilon,\lambda}(\nu,\nu)$.

The gradient of the unbalanced Sinkhorn divergence $\mathrm{S}_{\varepsilon,\lambda}(\mu_i,\nu_j)$ with respect to μ_i is finally given by:

$$\frac{\partial}{\partial \mu_i}\mathrm{S}_{\varepsilon,\lambda}(\mu,\nu) = \frac{\partial}{\partial \mu_i}\mathrm{OT}_{\varepsilon,\lambda}(\mu,\nu) - \frac{1}{2}\frac{\partial}{\partial \mu_i}\mathrm{OT}_{\varepsilon,\lambda}(\mu,\mu) + \varepsilon\,(m(\mu)-m(\nu)) \tag{4.34}$$

where the first term is given by (4.26) and the second term by (4.33).

4.4 Metric and domain scaling

4.4.1 A 2D illustration of metric scaling

In the context of FWI, the data are just pixelized images to be compared. Using a local misfit function (L_2 for instance) the relative size of the domain in time and offset does not impact the minimization. This is not the case when using a non local OT misfit.

When comparing seismograms, a natural question is how to measure the displacement of mass in time and offset space and how to choose the dimensions of along these two axis. This is equivalent to use a normalized domain $[0,1]^2$ and scaling the ground cost (4.3) ($\mathbf{x}=(t,x)$) We introduce a rescaled cost

$$c_{ijml} = \cos^2(\theta_M)\,|x_i - x_m|^2 + \sin^2(\theta_M)\,|t_j - t_l|^2 \tag{4.35}$$

depending on an angle $\theta_M \in]0,\pi/2[$. We are going to discuss how the translation of specific data depends on θ_M.

Let the observed data

$$\mu(t,x) = \delta_{\{\cos(\theta_D)\,x+\sin(\theta_D)\,t=0\}}$$

represent an idealized plane wave event. A single straight line carrying a weight of 1 and making an angle θ_D with the offset axis. The other part of the data (the simulation):

$$\nu_\tau(t,x) = \mu(t-\tau\sin(\theta_D), x-\tau\cos(\theta_D))$$

is a simple translation depending on a parameter τ in the direction $(\cos(\theta_D),\sin(\theta_D))$. This emulates the misfit associated with an erroneous constant background velocity and as explained in the introduction will cause cycle skipping when using L_2 misfit.

Assuming for simplicity that we are in a periodic box, the classical (non entropic) optimal transport map is a pure parallel translation of the support of μ. This is a consequence of Brenier's theorem which can be extended to ground costs formed as convex functions the vector $\mathbf{y}-\mathbf{x}$ such as (4.35) *Gangbo and McCann* (1996). The translation may however not occur in the normal direction $\theta_D + \frac{\pi}{2}$ because the shortest path now depends on the distance induced by (4.35) and the angle θ_M.

The OT misfit can therefore be reduced to computing the shortest path squared from the origin to the translated line. It is given by solving the convex program:

$$\begin{aligned} d_\tau^2 &= \underset{\substack{(x,t),\, s.t. \\ \cos(\theta_D)\, x + \sin(\theta_D)\, t = \tau}}{\operatorname{argmin}} \quad \frac{\cos(\theta_M)^2}{2}\,|x|^2 + \frac{\sin(\theta_M)^2}{2}\,|t|^2 \\ &= C\,\frac{\tau^2}{2} \end{aligned} \tag{4.36}$$

where

$$C = \frac{1}{\dfrac{\cos(\theta_D)^2}{\cos(\theta_M)^2} + \dfrac{\sin(\theta_D)^2}{\sin(\theta_M)^2}}. \tag{4.37}$$

Figure 4.4 shows the value of this parameter for different regimes (θ_D, θ_M).

The OT misfit remains quadratic in τ but C governs the strong convexity. We observe that the convexity modulus degenerates ($C \to 0$) only when the metric angle θ_M moves away from θ_D.

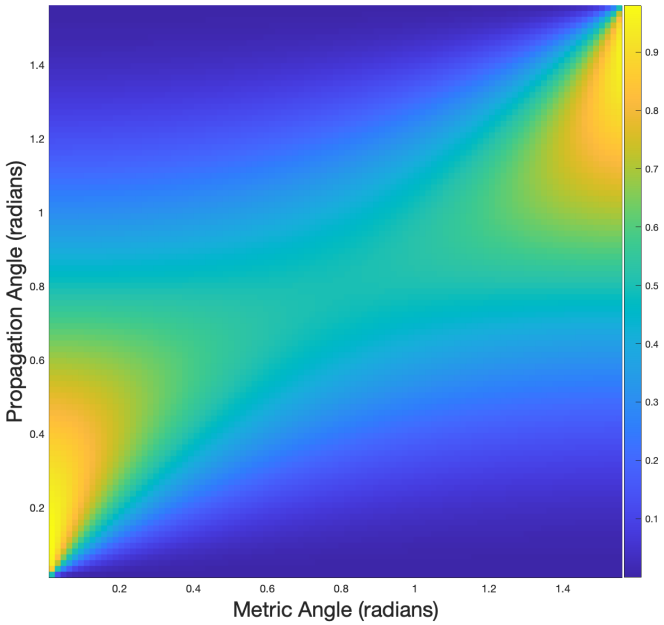


Figure 4.4: Colormap of C (4.37) as a function of θ_M (the Metric Angle) and θ_D (the Propagation Angle).

Real or synthetic complex seismic data may be interpreted a collection of local event possibly misplaced.

The discussion above raises the interesting but still untouched question of building a data dependent ground cost function where (α, β) depends on (t, x). One could even consider more general metrics such as Finsler metric for example *Benamou et al.* (2018).

4.4.2 On the (t, x) range of Entropic OT plans

In the entropic version of OT, the plan (4.8) is a diagonal scaling of the interaction Kernel which "governs" mass displacement. In its separable version (4.18) it becomes

$$K^{ijml} = e^{-\frac{|x_i - x_m|^2}{2\,\varepsilon}}\, e^{-\frac{|t_j - t_l|^2}{2\,\varepsilon}}. \tag{4.38}$$

Let us discuss the choice of the relative size of the domain in time and offset and their discretization. Let (T, N_t) and (X, N_x) be the maximum time, respectively offset, and number of points in time, respectively offset, used in the computation of the misfit. That gives a computational grid size $(dt, dx) = (\frac{T}{N_t}, \frac{X}{N_x})$. Setting $k_x\, dx = x_i - x_m$ and $k_t\, dt = t_j - t_l$ we rewrite (4.38):

$$K^{\varepsilon}_{ijml} = e^{-\frac{(k_x\, dx)^2}{2\,\varepsilon}}\, e^{-\frac{(k_t\, d_t)^2}{2\,\varepsilon}} = e^{-\frac{(k_x\, X)^2}{2\,\varepsilon\, N_x^2}}\, e^{-\frac{(k_t\, T)^2}{2\,\varepsilon\, N_t^2}} \tag{4.39}$$

Freezing ε and assuming (N_t, N_x) are given either by the data (number of lines) or though time sampling to limit the computational cost. The range of "horizontal" (in offset) or respectively "vertical" (in time) mass displacement is therefore governed by $(k_t\, T, k_x\, X)$. Setting the finite precision 0 at $\exp^{-36}$ the Kernel above will vanish when

$$k_t\, T = 6\,\sqrt{2\,\varepsilon}\, N_t \text{ or } k_x\, X = 6\,\sqrt{2\,\varepsilon}\, N_x.$$

Entropic Optimal Transport therefore allows to choose a smooth transport window of grid points in time and offset. It is possible to use (T, X) to control the width of this window. Let (k_t^{max}, k_x^{max}) be a prescribed *maximum allowed displacement* in terms of grid points and set

$$T = 6\,\frac{\sqrt{2\,\varepsilon}\, N_t}{k_t^{max}} \text{ or } X = 6\,\frac{\sqrt{2\,\varepsilon}\, N_x}{k_x^{max}}. \tag{4.40}$$

(also note that (T, X) can be interpreted as penalization of time/offset displacement in the ground cost as in the previous section).

Trace by trace misfit with entropic OT

We recall that the system solved by the Sinkhorn algorithm (4.11) is equivalent (in tensorized form (4.38)) to the marginal constraints:

$$\begin{cases} \sum_{ml} A_{ij}\, K^x_{im}\, K^t_{jl} B_{ml} = \mu_{ij} \;\; \forall\, i, j \\ \sum_{ij} A_{ij}\, K^x_{im}\, K^t_{jl} B_{ml} = \nu_{ml} \;\; \forall m, l \end{cases} \tag{4.41}$$

Setting $k_x^{max} = 1$ in particular corresponds to 1D "trace by trace" misfit. The offset part of the Kernel K_{im}^x will vanish except when $k_x = 0$, i.e. $x_i = x_m$, in (4.38) where it is just 1. The 2D marginal constraints (4.41) will therefore reduce *for all offsets* x_i to the 1D marginal constraints :

$$\begin{cases} \sum_l A_{ij}\, K_{jl}^t B_{il} = \mu_{ij} \;\; \forall\, j \\ \sum_j A_{ij}\, K_{jl}^t B_{il} = \nu_{il} \;\; \forall\, l \end{cases} \tag{4.42}$$

and therefore we solve the 1D trace by trace OT_ε problem. This remark naturally extends to unbalanced (soft marginal constraints) and Sinkhorn divergence variants.

Chapter 5

The unbalanced optimal transport misfit function

The general abstract form of the misfit to minimize (see (2.8)) is

$$\mathcal{J}(\mathbf{m}) = h\left(\mathbf{d}_{\text{cal}}(\mathbf{m}), \mathbf{d}_{\text{obs}}\right) \tag{5.1}$$

where m is the model we strive to reconstruct and $\mathbf{d}_{\text{obs}}$ is some given observation. The map $m \mapsto \mathbf{d}_{\text{cal}}(\mathbf{m})$ is the *forward model* and in this thesis we used a 2D constant-density acoustic finite difference wave modeling. We also limit ourselves to synthetic data sets that is $\mathbf{d}_{\text{obs}} := \mathbf{d}_{\text{obs}}(\mathbf{m}_{true})$, the observation is constructed with the same forward mode ling using a *true* model which is known a priori.

We recall that we have implemented on top of PySIT (`https://pysit.org/`) is an "an open source toolbox for seismic inversion and seismic imaging developed by Russell J. Hewett and Laurent Demanet at MIT". It was designed to allow for fast prototyping of the h component of the misfit function (5.1). Developers can implement their own misfit on top of the available forward models (in our case 2D constant-density acoustic finite difference wave modeling) and optimisation method (we have been using the Limited-memory BFGS method).

The rest of this chapter explains how we constructed the h function using the entropic OT costs and shows its behavior on simple parametric models.

5.1 The misfit function

In an attempt to preserve phase information, we follow the strategy of comparing separately the positive and negative parts ($f^+ = \max(f, 0)$ and $f^- = \max(-f, 0)$) of the observed and predicted seismic signal as proposed in (*Engquist and Froese*, 2014; *Engquist et al.*, 2016). This is to take care of the positivity constraint on the data. As explained in chapter 4 instead of normalizing and using the classic W_2^2 quadratic

Wasserstein distance, we can use the unbalanced OT modification and its various entropic approximations.

A first choice is to use (4.23)

$$h_{\mathrm{OT}_{\varepsilon,\lambda}}(\mathbf{d}_{\mathrm{cal}}, \mathbf{d}_{\mathrm{obs}}) \quad = \quad \mathrm{OT}_{\varepsilon,\lambda}\left(\mathbf{d}^+_{\mathrm{cal}}, \mathbf{d}^+_{\mathrm{obs}}\right)) + \mathrm{OT}_{\varepsilon,\lambda}\left(\mathbf{d}^-_{\mathrm{cal}}, \mathbf{d}^-_{\mathrm{obs}}\right) \tag{5.2}$$

but we will mostly focus on its "debiased" Sinkhorn divergence version (4.30)

$$h_{\mathrm{S}_{\varepsilon,\lambda}}\left(\mathbf{d}_{\mathrm{cal}}, \mathbf{d}_{\mathrm{obs}}\right) \quad = \quad \mathrm{S}_{\varepsilon,\lambda}\left(\mathbf{d}^+_{\mathrm{cal}}, \mathbf{d}^+_{\mathrm{obs}}\right) + \mathrm{S}_{\varepsilon,\lambda}\left(\mathbf{d}^-_{\mathrm{cal}}, \mathbf{d}^-_{\mathrm{obs}}\right). \tag{5.3}$$

This will not be used in this chapter but in order to ensure misfit function differentiability, we will split the signal, here denoted f, into a smooth positive $f^+_\delta = \mathrm{P}_\delta(f)$ and smooth negative $f^-_\delta = \mathrm{P}_\delta(-f)$ part, where P_δ is given by

$$\mathrm{P}_\delta(f) = \frac{1}{2}\left(f + \sqrt{f^2 + (\delta\, f_0)^2}\right) \tag{5.4}$$

The parameter δ is a small parameter that controls the smoothness of P_δ and f_0 carries the dimension of the signal (typically $f_0 = \|\mathbf{d}_{\mathrm{obs}}\|_\infty$.). This function is plotted in in Figure 5.1, for a simple function $f(t) = t$, $f_0 = 1$ and different values of δ.

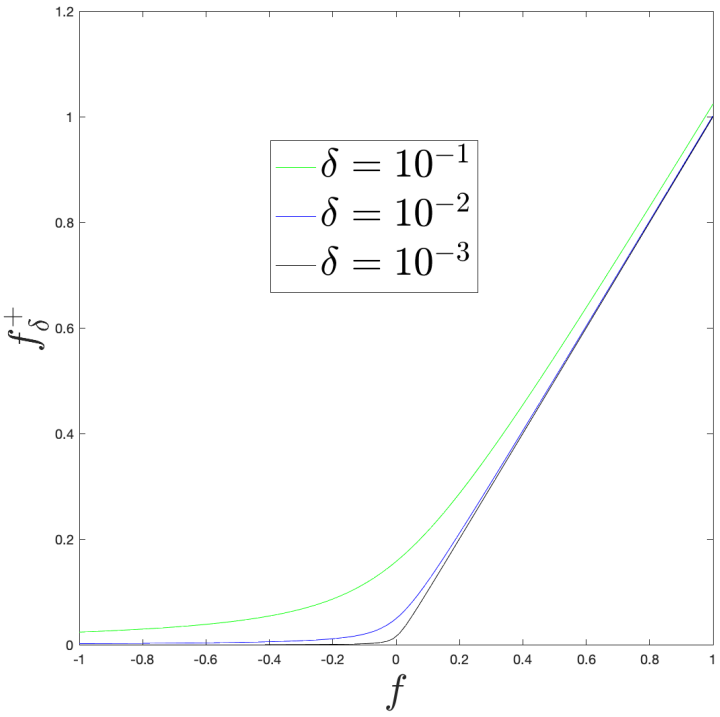


Figure 5.1: Illustration of the effect of δ for the transformation P_δ defined by (5.4), applied to the function $f(t) = t, \forall t \in [-1, 1]$, which changes sign at $t = 0$. As δ increases, $P_\delta(f)$ is getting smoother and smoother in the vicinity of $t = 0$, but increasing artificial mass is getting created that lead to possible bias in optimal transport.

In summary the misfit function, used for FWI in the last chapter, is

$$h_{\mathrm{S}_{\varepsilon,\lambda}}\left(\mathbf{d}_{\mathrm{cal}},\mathbf{d}_{\mathrm{obs}}\right) \quad = \quad \mathrm{S}_{\varepsilon,\lambda}\left((\mathbf{d}_{\mathrm{cal}})_{\delta}^{+},(\mathbf{d}_{\mathrm{obs}})_{\delta}^{+}\right)+\mathrm{S}_{\varepsilon,\lambda}\left((\mathbf{d}_{\mathrm{cal}})_{\delta}^{-},(\mathbf{d}_{\mathrm{obs}})_{\delta}^{-}\right). \tag{5.5}$$

and for completeness the adjoint-source to be used in the adjoint wave propagation (2.42) is given by

$$\begin{aligned}\frac{\partial}{\partial\mathbf{d}_{\mathrm{cal}}}h\left(\mathbf{d}_{\mathrm{cal}};\mathbf{d}_{\mathrm{obs}}\right)=&\frac{\partial}{\partial\mathbf{d}_{\mathrm{cal}}}\mathrm{S}_{\varepsilon,\lambda}\left((\mathbf{d}_{\mathrm{cal}})_{\delta}^{+},(\mathbf{d}_{\mathrm{obs}})_{\delta}^{+}\right)\ \frac{\partial}{\partial\mathbf{d}_{\mathrm{cal}}}\mathrm{P}_{\delta}\left(\mathbf{d}_{\mathrm{cal}}\right)\\&-\frac{\partial}{\partial\mathbf{d}_{\mathrm{cal}}}\mathrm{S}_{\varepsilon,\lambda}\left((\mathbf{d}_{\mathrm{cal}})_{\delta}^{-},(\mathbf{d}_{\mathrm{obs}})_{\delta}^{-}\right)\ \frac{\partial}{\partial\mathbf{d}_{\mathrm{cal}}}\mathrm{P}_{\delta}\left(-\mathbf{d}_{\mathrm{cal}}\right).\end{aligned} \tag{5.6}$$

where the differentials of the Sinkhorn divergence (4.30) with respect to $\mathbf{d}_{\mathrm{cal}}$ is given by (4.34).

Figure 5.2 shows that the smoothness of adjoint source, given by the pixel wise multiplication of the $\mathrm{S}_{\varepsilon,\lambda}$ gradient and the P_δ derivative depends on the parameter δ.

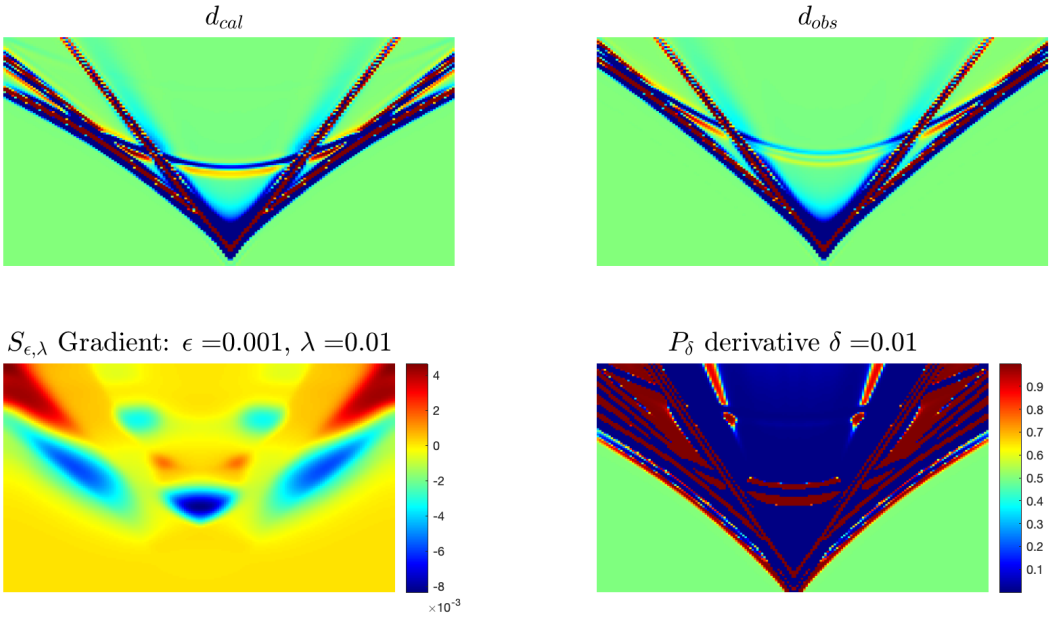


Figure 5.2: The top row are a three layer synthetic model recording (d_{obs}) and data (d_{cal}) for one of the intermediate models in the simulation. The bottom line show the $\mathrm{S}_{\varepsilon,\lambda}$ gradient for the positive part and the P_δ derivative of the signal needed to smooth the gradient.

5.2 1/2D parametric examples

In this section we skip the smoothing related to the sign splitting of the data. We will use misfits (5.2) and (5.3) and they will be referred to as $OT_{\varepsilon,\lambda}$ and $\mathrm{S}_{\varepsilon,\lambda}$. We will omit the λ when the data is balanced and there is no need for unbalanced OT.

5.2.1 Translation/shift of a 1D signal

We are considering two copies of the same Ricker wavelet, with $20Hz$ peak frequency, denoted $f(t)$, playing the role of the observed signal $d_{\text{obs}}(t)$, and $g(t)$, playing the role of the predicted signal $d_{\text{cal}}(t)$. The two wavelets are shifted in time such that $g(t;s) = f(t-s)$ see Figure 5.3. We will study the effect on the misfit of the variations in s the time shift.

We recall the classical result (see section 3.2): the squared Wasserstein distance W_2^2 is quadratic, hence convex with respect to the translation of identical data. Sign splitting will preserve the mass balance and the translation and as such will not change this property. This was already verified numerically in (*Engquist and Froese*, 2014). We here compare the behaviour of the L_2, OT_ε, and S_ε misfit functions, Figure 5.4. In this example, the data is balanced so we do not need the unbalanced version of the transport distance. For such a simple example there is not much difference between the entropic OT_ε and the Sinkhorn divergence but the debiasing effect discussed in section 4.3.1 is noticeable. We are very close to a perfect $s \mapsto 2\,s^2$ (the 2 factor comes from sign splitting). Sinkhorn divergence is known to be a good proxy for non entropic OT. In all experiments $\varepsilon = 10^{-2}$ and $\lambda = 10$.

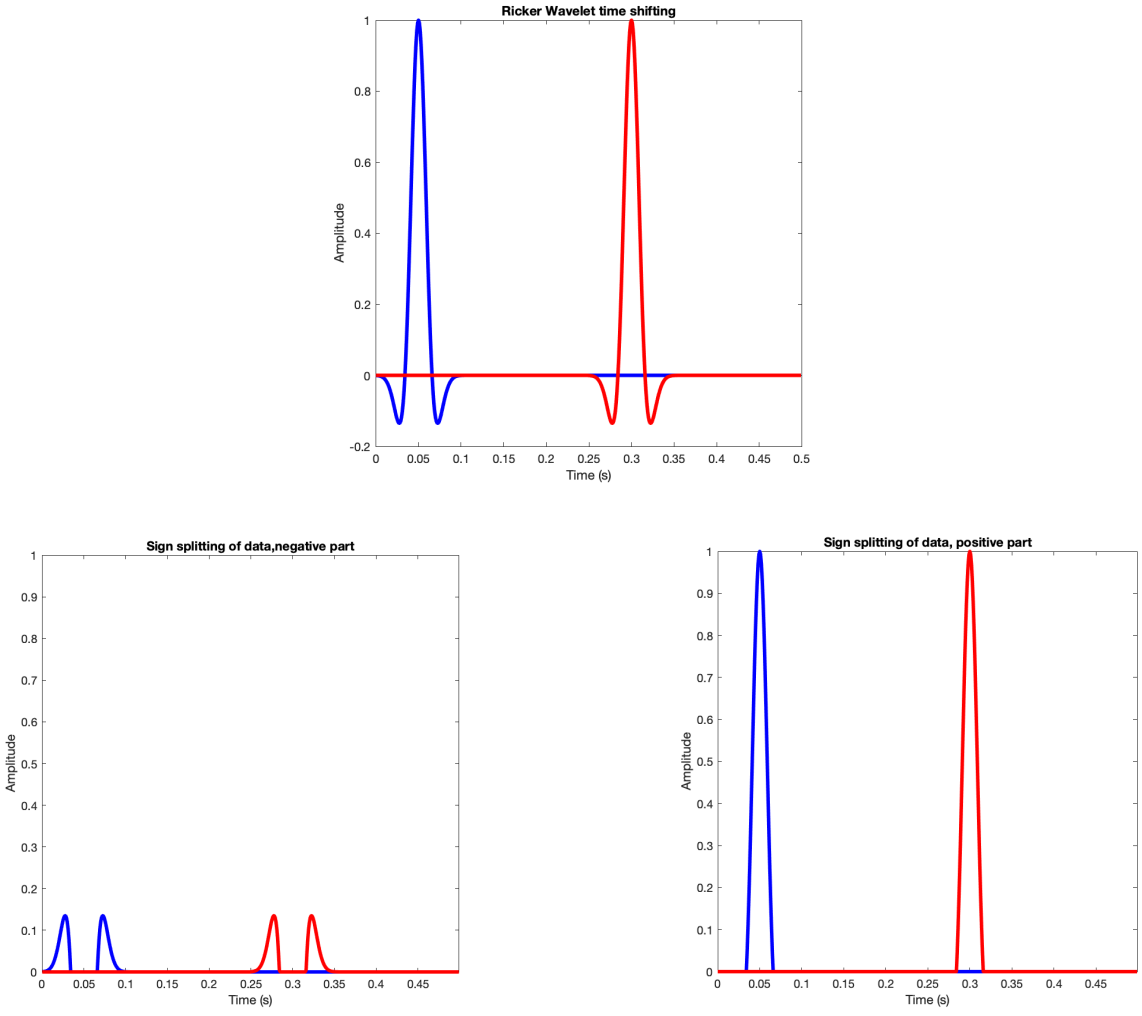


Figure 5.3: Top figure: Ricker wavelets $f(t)$ (solid blue line) and $g(t)$ (solid red line), the latter being a shifted version for a translation $s = 0.25$. Bottom Figures: the negative (bottom-left figure) and the positive (bottom-right figure) parts of the two Ricker wavelets. Also their negative and positive parts.

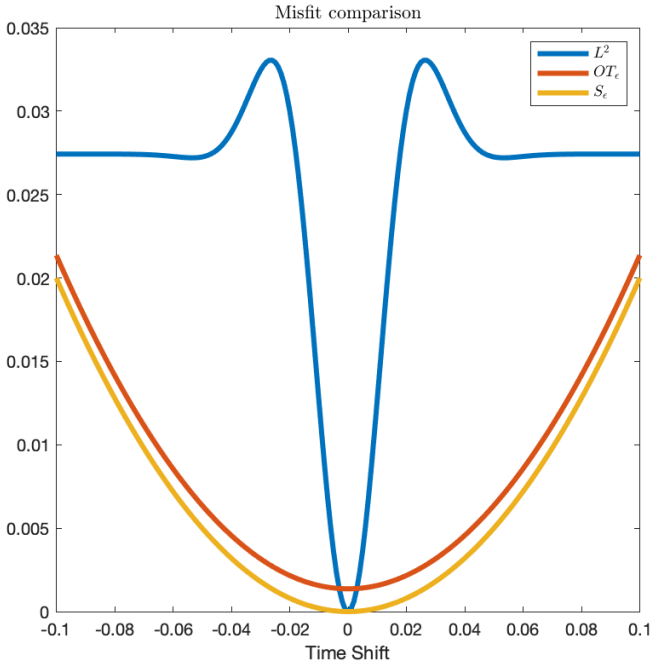


Figure 5.4: Misfit with respect to the shift s.

In the next experiment, Figure 5.6, we consider the misfit behaviour as a function of the time shift between the two Ricker wavelets when a Gaussian noise is added to one of the wavelet (linked for instance to the acquisition of d_{obs}), see Figure 5.5.

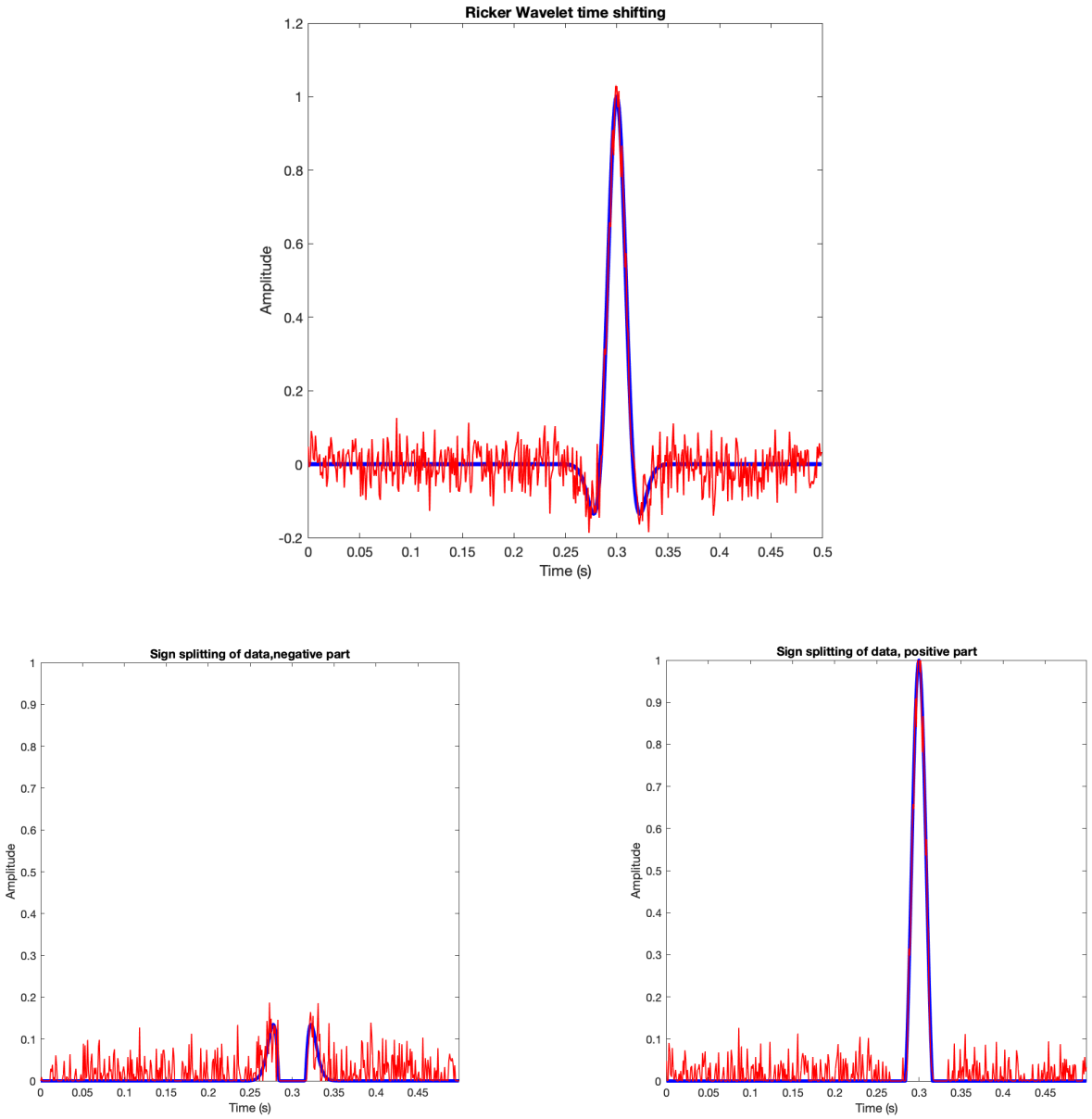


Figure 5.5: Top figure: Ricker wavelets $g(t)$ (solid blue line) and $f(t)$ (solid red line), with an added Gaussian noise. Bottom figures: the negative (bottom-left figure) and the positive (bottom-right figure) parts of the two Ricker wavelets.

We compare the L_2 and the unbalanced $OT_{\varepsilon,\lambda}$ and $\mathrm{S}_{\varepsilon,\lambda}$ misfit functions for an increasing level of Gaussian noise (1, 5, 10, 20 %). The data is now unbalanced because of the noise. We also consider the balanced S_ε after normalisation of the data (as in (*Engquist and Froese*, 2014; *Engquist et al.*, 2016)).

As expected L_2 global minima remains centred but more local minima appear due to the noise. The balanced (normalised) S_ε remains convex as expected but the global minimum is shifted up because of the noise. It is indeed built to be positive convex and vanish when comparing two identical measures, this never is the case with added noise. A constant is added for all the mass creation. The position of the minimum depicted by a black cross is biased because the noisy and normalized signal does not satisfy the initial Ricker symmetry anymore.

We recall that unbalanced OT remains a distance on radon measures, the marginal constraint on the data is penalised using the Kullback-Leibler divergence $\mathcal{KL}$ see (4.22) (we have used $\lambda_1 = \lambda_2 = \lambda$). Unbalanced

Sinkhorn Divergence $S_{\epsilon,\lambda}$ is also covex positive and corrects the unbalanced $OT_{\epsilon,\lambda}$ by removing the diagonal terms and therefore achieves a lower miminum. Again it cannot be zero as we are comparing different measures. The unbalanced version seems however to do a better job as the minimum is closer to the zero shift.

Finally, we remark that the noise is akin to add mass everywhere therefore impacting the transport. Because we use the quadratic ground cost (4.3), the transport will favor moving the nearby noise mass when possible instead of fetching the shifted Ricker. This effect increases with the noise level, reduces the range of the transport and the modulus of convexity of the misfit.

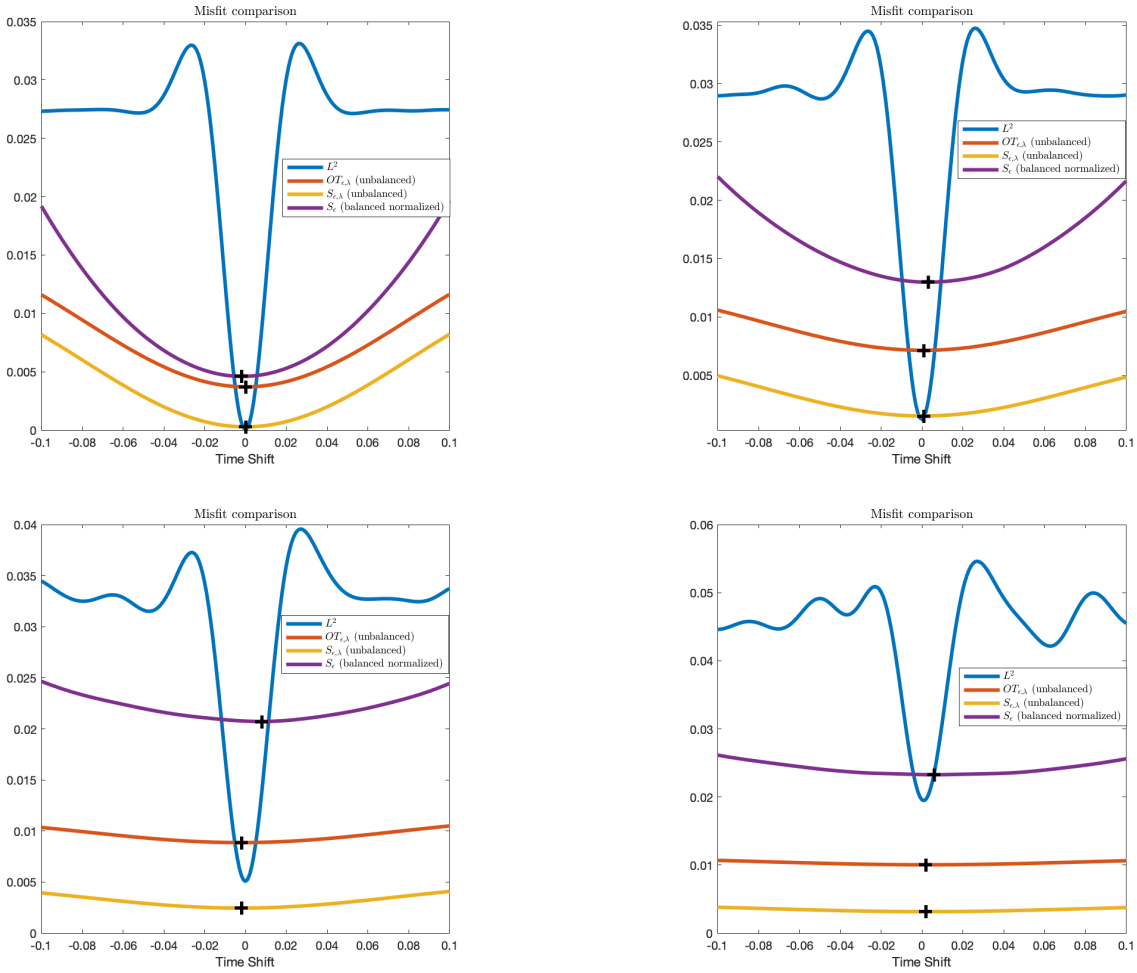


Figure 5.6: Comparison between the L^2 (solid blue curve), the balanced (normalised) $S_{\epsilon,\lambda}$ (solid purple curve), and the unbalanced $OT_{\epsilon,\lambda}$ (solid red curve) and $S_{\epsilon,\lambda}$ (solid yellow curve) misfit functions as a function of the time shift between the two Ricker wavelets for increasing Gaussian noise added to one of the wavelet. From top-left to bottom-right $\mathcal{N}(0, 0.01)$, $\mathcal{N}(0, 0.05)$, $\mathcal{N}(0, 0.1)$, $\mathcal{N}(0, 0.2)$: 1,5,10 and 20 %.

In figure 5.7 and or a fixed noise (5%) we explore the behavior of the different misfit for different values of ϵ and λ. In the first line ϵ is decreased, as expected $OT_{\epsilon,\lambda}$ and $S_{\epsilon,\lambda}$ (the second order approximation in ϵ of OT_λ converge). We also notice that normalizing the data induce a significant bias to the minimum. The second line, when the marginal constraint are locally relaxed, the relative diminution of the mass transport leads to degeneracy of the convexity property with respect to translation (the amount of mass beeing transported decrease).

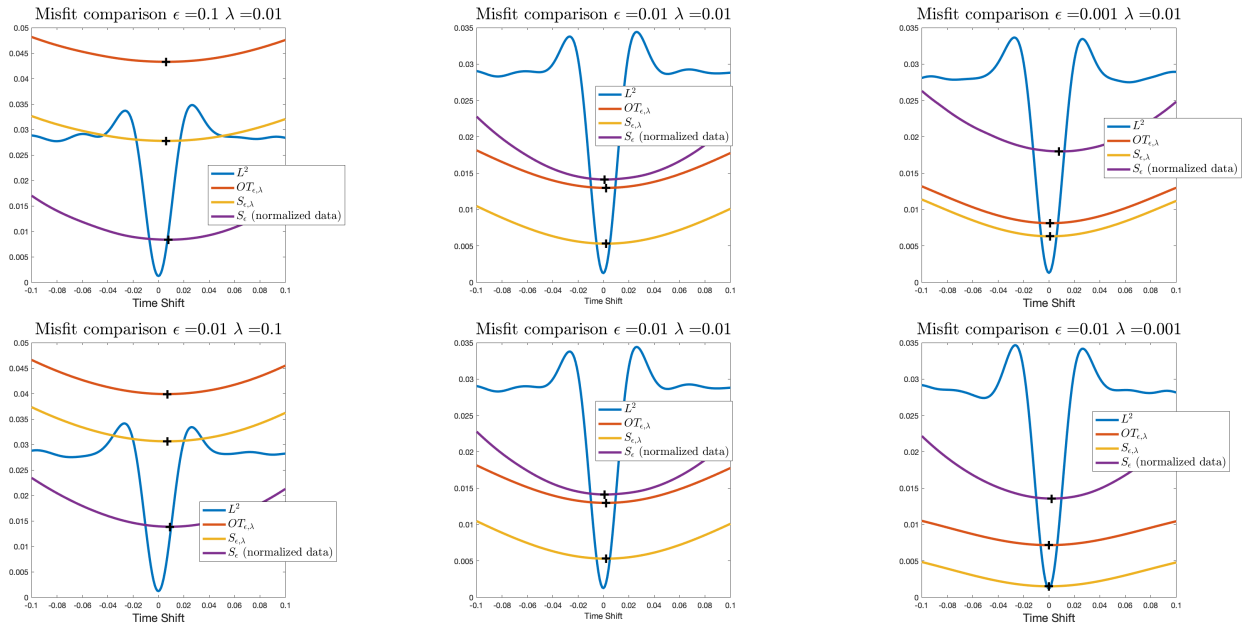


Figure 5.7: Comparison between the L^2 (solid blue curve), the balanced (normalised) $\mathrm{S}_{\epsilon,\lambda}$ (solid purple curve), and the unbalanced $\mathrm{OT}_{\epsilon,\lambda}$ (solid red curve) and $\mathrm{S}_{\epsilon,\lambda}$ (solid yellow curve) misfit functions as a function of the time shift between the two Ricker wavelets (one with added 5% noise) for decreasing ϵ and λ.

We perform the same experiment, Figure 5.9, when instead of adding noise to the recorded wavelet, we change its frequency content, see Figure 5.8. We roughly have the same observation as in figure 5.7. We notice that the normalisation approach does much better, probably because the support of the wavelet is localised.

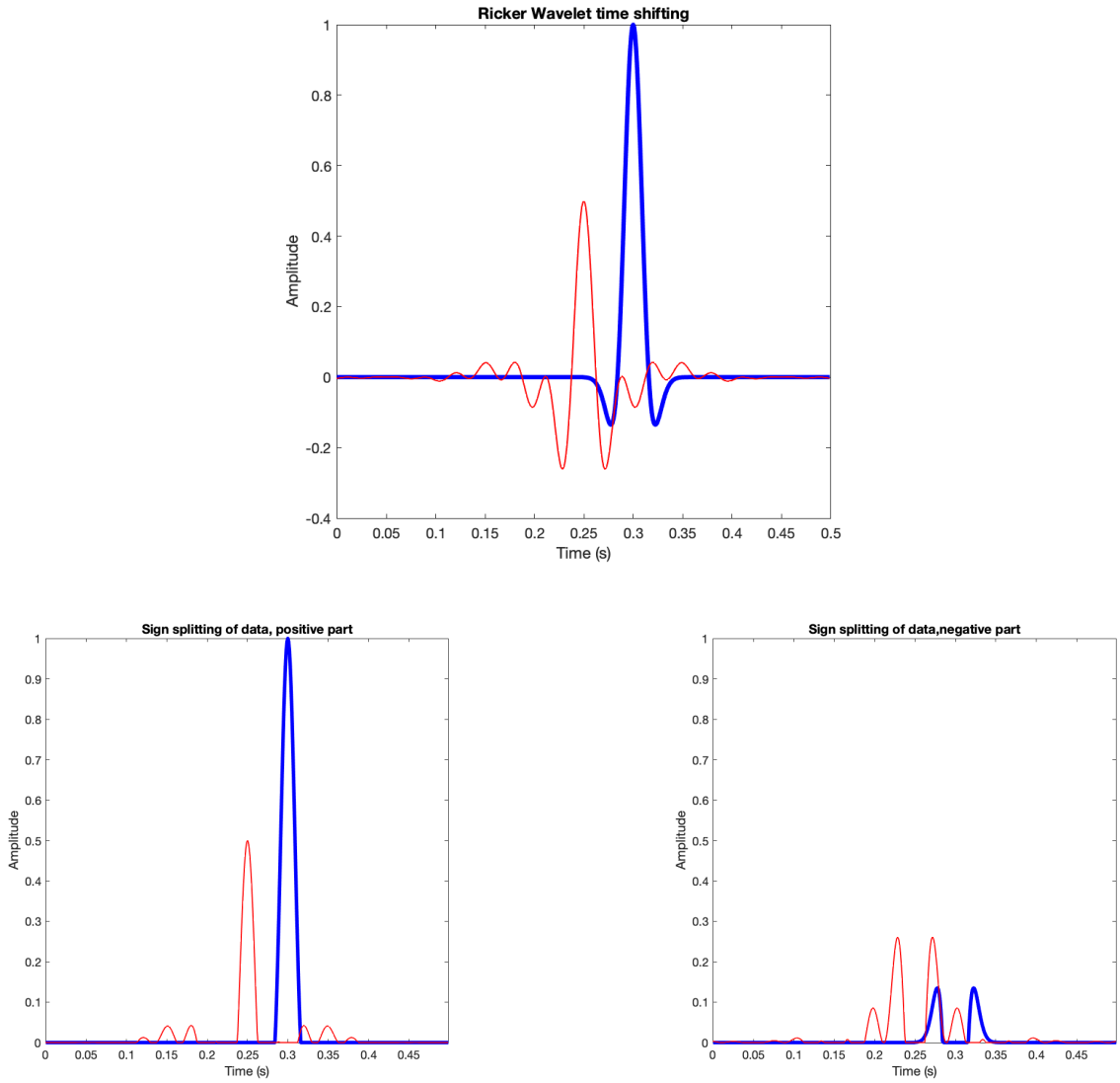


Figure 5.8: Top figure: Ricker wavelet $g(t)$ (solid blue line) and Ormsby Wavelet $f(t)$ (solid red line) with characteristic frequencies (5,10,30,35Hz) . Bottom figures: the negative (bottom-left figure) and the positive (bottom-right figure) parts of the two wavelets.

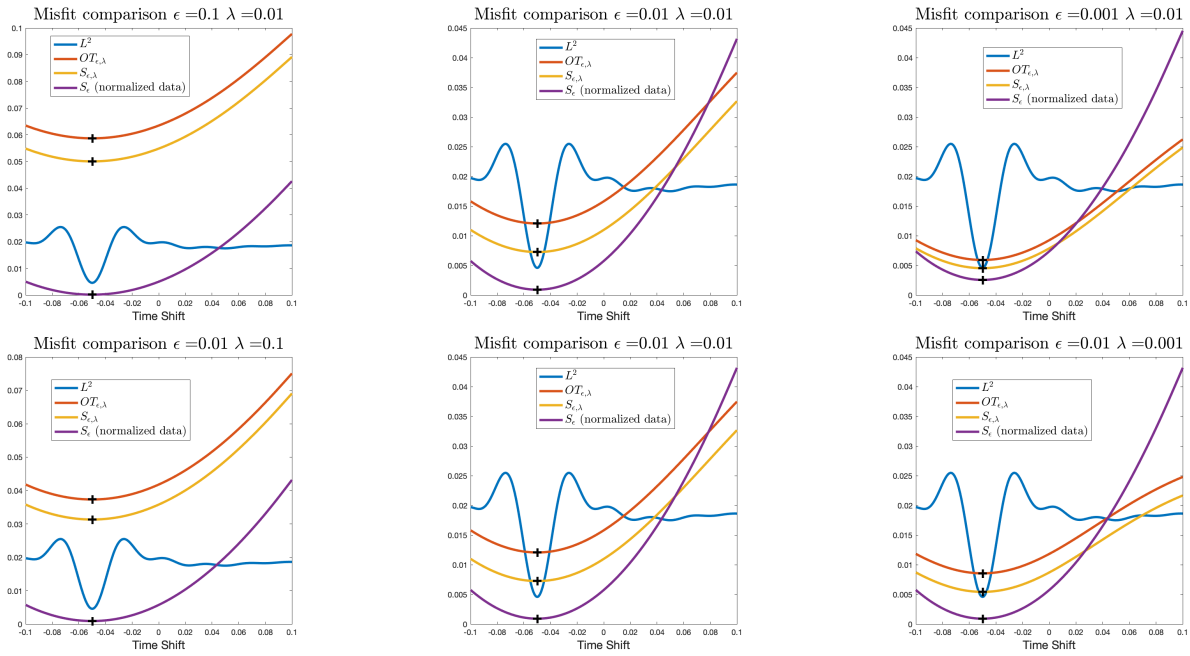


Figure 5.9: Comparison between the L^2 (solid blue curve), the balanced (normalised) $\mathrm{S}_{\epsilon,\lambda}$ (solid purple curve), and the unbalanced $\mathrm{OT}_{\epsilon,\lambda}$ (solid red curve) and $\mathrm{S}_{\epsilon,\lambda}$ (solid yellow curve) misfit functions as a function of the time shift between Ricker and Ormsby wavelets for decreasing ϵ and λ.

5.2.2 Linear 2D model

A simple two-parameter model, borrowed from *Métivier et al.* (2016a), allows the 2D representation of the L_2, $\mathrm{OT}_{\varepsilon,\lambda}$ and $\mathrm{S}_{\varepsilon,\lambda}$ misfit functions. We use sign splitting for the positivity for the transport misfits. Let consider a 2D-acoustic wave approximation model, with constant density, parameterised by the background acoustic velocity v_0 and a depth gradient α, such that

$$v = v_0 + \alpha z \tag{5.7}$$

The physical domain is a 2D rectangular domain 17 km long and 3.5 km deep. A single seismic source is located at $x = 8.45\,\mathrm{km}$ at a depth of 50 m. The radiated acoustic waves are recorded at 168 receivers located at the same depth position that the source and regularly deployed with a spacing of 100 m for $x \in [0.15, 16, 85]$ km. The source wavelet is a Ricker wavelet centred at 5 Hz.

The observed signals $\mathbf{d}_{\mathrm{obs}}(\mathbf{x}_r, t)$ correspond to a numerical physical realisation, using second-order finite difference stencils on a regular finite difference grid and a second-order leap-from time integration, for a model parameterised with $v_0^\star = 2\,\mathrm{km/s}$ and $\alpha^\star = 0.67\,\mathrm{s}^{-1}$.

The L_2, $\mathrm{OT}_{\varepsilon,\lambda}$ and $\mathrm{S}_{\varepsilon,\lambda}$ misfit functions are compared for predicted signals $\mathbf{d}_{\mathrm{cal}}(\mathbf{x}_r, t)$ corresponding to numerical physical realisations when exploring the parameter space $v_0 \in [1.75, 2.25]$ km/s and $\alpha \in [0.49, 0.91]$ s^{-1} with $\Delta v_0 = 0.0125\,\mathrm{km/s}$ and $\Delta\alpha = 0.015\,\mathrm{s}^{-1}$, leading to a discretised model space of 29×41 points. For the OT misfit functions, the entropic parameter is chosen as $\varepsilon = 0.01$.

The L_2 misfit function, in Figure 5.10, exhibits multiple local minima, and convergence to the global minimum with local gradient-descent methods requires to start with a good enough initial model, otherwise the optimisation problem will be trapped in a local minimum.

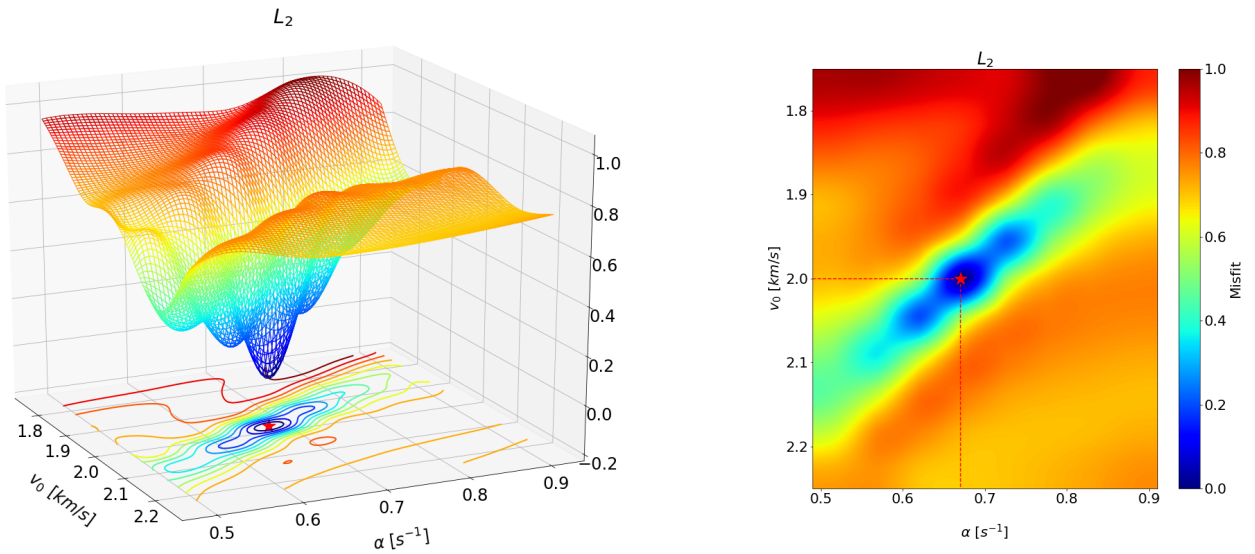


Figure 5.10: The L_2 misfit function as a function of the model parameters v_0 and α. The red star indicates the location of the global minimum, while in the right figure the intersection of the two dashed lines indicate the reference model $v_0^\star, \alpha^\star$ in the model space.

The $(\mathrm{OT}_{\varepsilon,\lambda})$ misfit function, in Figure5.11, exhibits better convexity property but poor sensitivity with a large basin of attraction toward the global minimum. As seen in the Figure 5.11, the location of the global minimum in the model space, indicated by the red star, does not recover the position of the reference model $v_0^\star, \alpha^\star$, indicated by the blue star in the left figure. This is the signature of a bias in the transport associated to the Entropic penalization: it may become cheaper (lower entropy) to diffuse from d_{cal} or d_{obs} than not moving them.

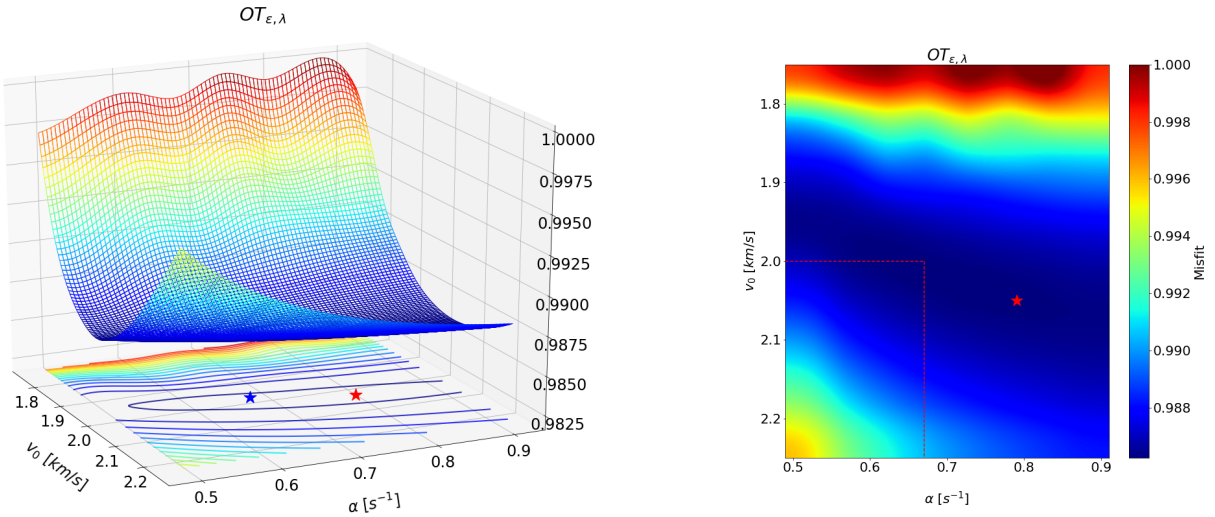


Figure 5.11: The $\mathrm{OT}_{\varepsilon,\lambda}$ misfit function as a function of the model parameters v_0 and α. The red star indicates the location of the global minimum. The blue star indicates the position of the reference model $v_0^\star, \alpha^\star$ in the left the figure, while in the right figure this position is indicated as the intersection of the two dashed lines.

The $\mathrm{S}_{\varepsilon,\lambda}$ misfit function, in Figure 5.12, exhibits nice convexity and sensitivity properties, together with a large basin of attraction toward the global minimum. It is worth to remind that while the mapping $d_{\mathrm{cal}} \mapsto \mathrm{S}_{\varepsilon,\lambda}(d_{\mathrm{obs}}, d_{\mathrm{cal}})$ is strictly convex, here this is the mapping $(v_0, \alpha) \mapsto \mathrm{S}_{\varepsilon,\lambda}(d_{\mathrm{obs}}, d_{\mathrm{cal}}(v_0, \alpha))$ that is actually plotted.

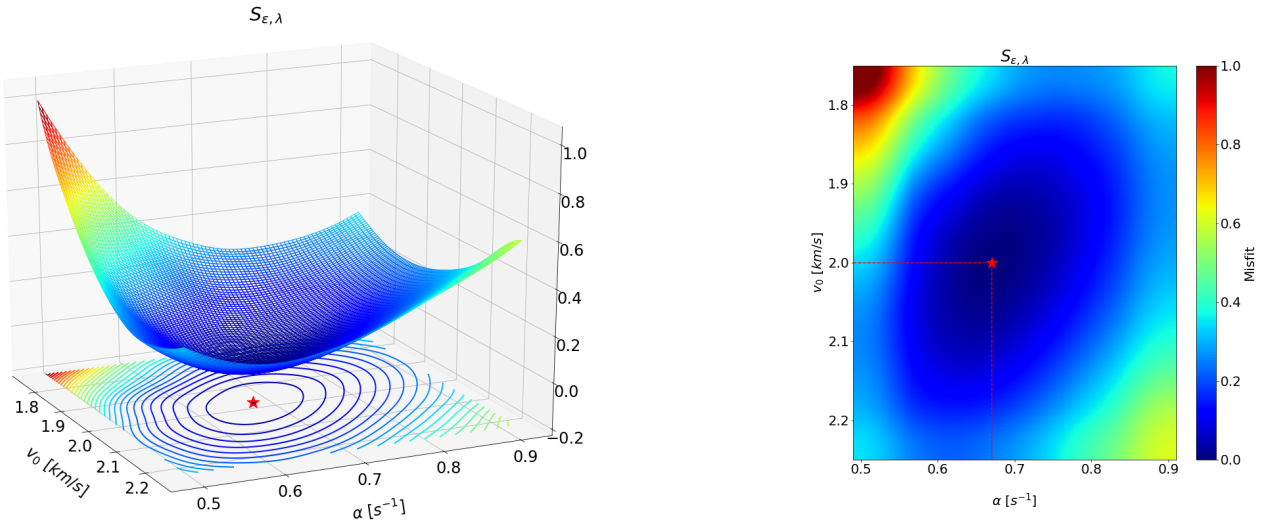


Figure 5.12: Misfit function based on the Sinkhorn divergence ($S_{\varepsilon,\lambda}$), with respect to v_0 and α. The red star is the location of the global minimum, while the position of the reference model $v_0^\star, \alpha^\star$ is indicated as the intersection of the two dashed lines.

Chapter 6

2D FWI numerical illustrations

6.1 Preliminaries

This section provides numerical illustrations for the use of the misfit function, see (5.6), based on the Sinkhorn divergence formulation of unbalanced, entropic optimal transport , see (4.30), through canonical 2D models.

The first model, *i.e.* the Camembert model, focuses on transmission and reflection configurations. The second model, *i.e.* layered model, focuses on simple reflection configurations. The third model, *i.e.* the Marmousi model, focus on a more realistic configuration.

All the numerical examples are performed using 2D constant-density acoustic wave modelling and local gradient-based optimisation as implemented in the PySIT platform. This is reviewed in section 2. The seimic source is assumed to be a point source with a Ricker wavelet for time function.

For all the numerical experiments, the main parameters values associated to the $S_{\varepsilon,\lambda}$-misfit function and the iterative Sinkhorn algorithm are summarised below

Parameter (used value)	Use
$\varepsilon = 10^{-2}$	Entropic regularisation, see section 4.1.
$\lambda_{1,2} = 10^{-1}$	$\mathcal{KL}$ unbalanced penalisation of the marginal constraints on d_{cal} and d_{obs}, see section 4.2.2.
$\delta = 10^{-3}$	P_δ-transform into positive/negative part of d_{cal} and d_{obs}, see section 5.1.
$\text{tol} = 10^{-09}$	Sinkhorn convergence error (4.17). With the choice of parameters above and an optimal transport grid size considered of the order 10^2 in space and 128 in time, the number of iterations to achieve this precision is always of the order of 10^2 iterations and never exceeded 10^3 iterations.
$n_t = 128$	Number of discretisation points in time used for the OT computation. This implies a sampling of the signals which discretization vary according to the model and n_r.
n_r problem-dependent	Number of discretisation points in offset for the OT computation. It also is by default the number of receivers used in the acquisition configuration and the FD discretization of the forward model.
$k_t^{max} = [25, 1]$	Transport window in time (in number of points) (see section 4.4.2), $k_t^{max} = 1$ corresponds to the trace by trace misfit.
$k_x^{max} = [25, 1]$	Transport window in space (in number of points) (see section 4.4.2).

6.2 Velocity anomaly inclusion

This is the classical circular velocity anomaly inclusion inclusion model in a homogeneous medium, and is inspired from *Pladys et al.* (2021). The physical model is defined on a 1000×1000 m square domain, and is parametrised by a homogeneous background acoustic wave velocity set to 1300m/s, and a circular inclusion of 100 m radius centred in space, with a homogeneous wave velocity 1700m/s ($\approx 30\%$ anomaly).

The computational domain includes additional perfectly matched layers along each of the boundaries. The constant-density, acoustic wave equation is solved in space on a regular grid of 256×256 points using a sixth-order, centred finite difference scheme and a second-order leap-frog integration scheme in time with constant time step satisfying the CFL condition, as implemented in PySIT.

The FWI inversion problem is solved using the l-BFGS method, as implemented in PySIT, and the number of FWI iterations is initially fixed to 25 iterations. Two homogeneous initial models are considered: initial model-A with acoustic wave velocity set to 1300 m/s; and initial model-B with acoustic wave velocity set to 1700 m/s.

Both the predicted $\mathbf{d}_{\text{cal}}$ and the observed $\mathbf{d}_{\text{obs}}$ data, *i.e.* the latter corresponding to the "true" model, are physical realisations obtained with the same wave propagation solver, and without considering additional noise to the data, *i.e.* the inversion crime setting.

Unless otherwise specified, the $\mathrm{S}_{\varepsilon,\lambda}$-misfit function is constructed as discussed in Chapter 5. The observed and predicted signals are decomposed into positive and negative parts using the P_{δ}-transforn in (5.4), and transported separately using the Sinkhorn divergence formulation of unbalanced optimal transport.

The numerical experiments focus on transmission and reflection configurations.

6.2.1 Transmission configuration

The acquisition system is composed of 48 sources evenly-spaced below the upper boundary of the physical domain, at 200m depth, and of 256 evenly-spaced receivers located 200m above the lower boundary of the physical domain. The source-time function is a Ricker wavelet with a central frequency of $f_c = 3$Hz, and the recording time is set to 2s.

Reconstructed models with the L_2 and the $\mathrm{S}_{\varepsilon,\lambda}$ misfit functions are shown in Figure 6.1 after only 25 iterations for both initial models at which the inversion has not yet fully converged, Figure 6.2

For the initial model-A both the L_2 and $\mathrm{S}_{\varepsilon,\lambda}$ based inversion reconstruct reasonably well the model after only 25 iterations. The L_2 based inversion recovers the peak-amplitude of the velocity anomaly, whereas the peak-amplitude is not yet fully recovered (≈ 1550m/s) after 25 iterations with the $\mathrm{S}_{\varepsilon,\lambda}$ based inversion.

For the initial model-B the L_2-based misfit function suffers from cycle-skipping, and the inversion problem gets trapped in a local minimum and fails reconstructing the inclusion while the $\mathrm{S}_{\varepsilon,\lambda}$-based inversion provides more reasonable results and shows less sensitivity with respect to the initial model. even though the the background velocity is not yet fully recovered after 25 iterations. The improved convexity of the $\mathrm{S}_{\varepsilon,\lambda}$-misfit function avoids the FWI problem to be trapped in local minima and enable the convergence toward the global minimum.

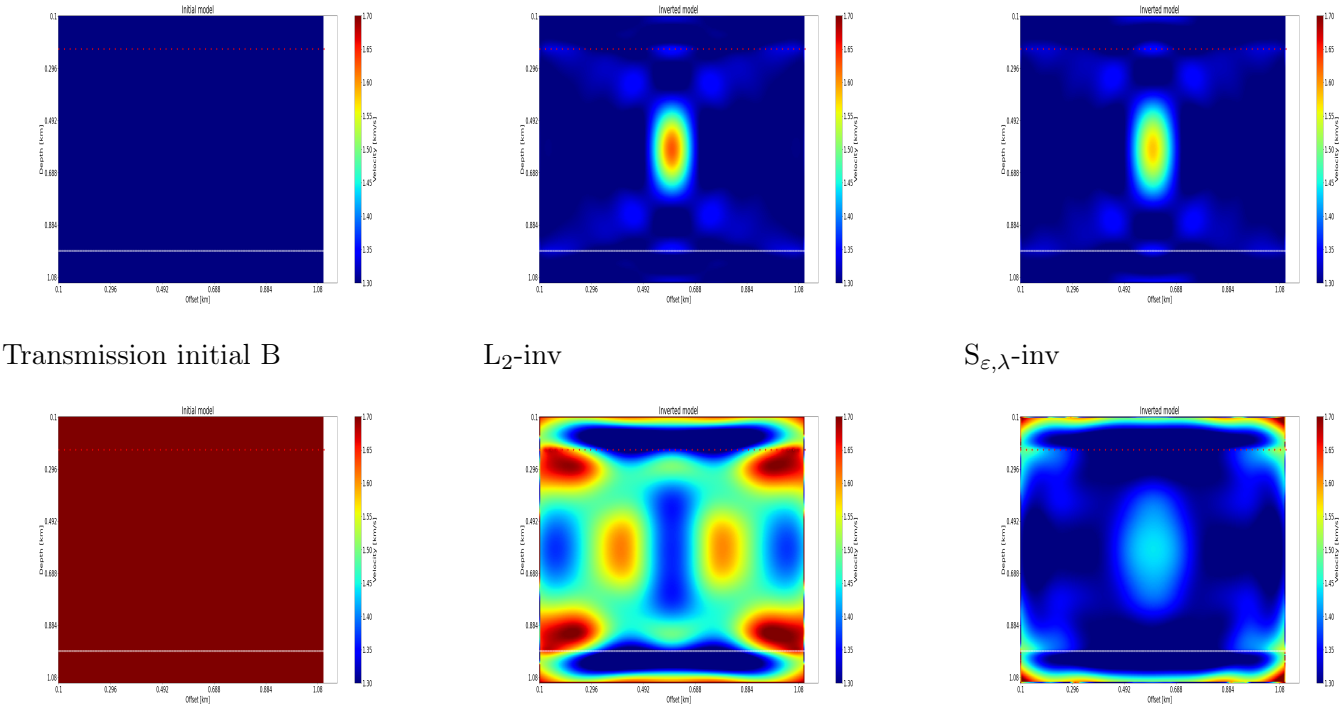


Figure 6.1: Numerical inversion after 25 iterations

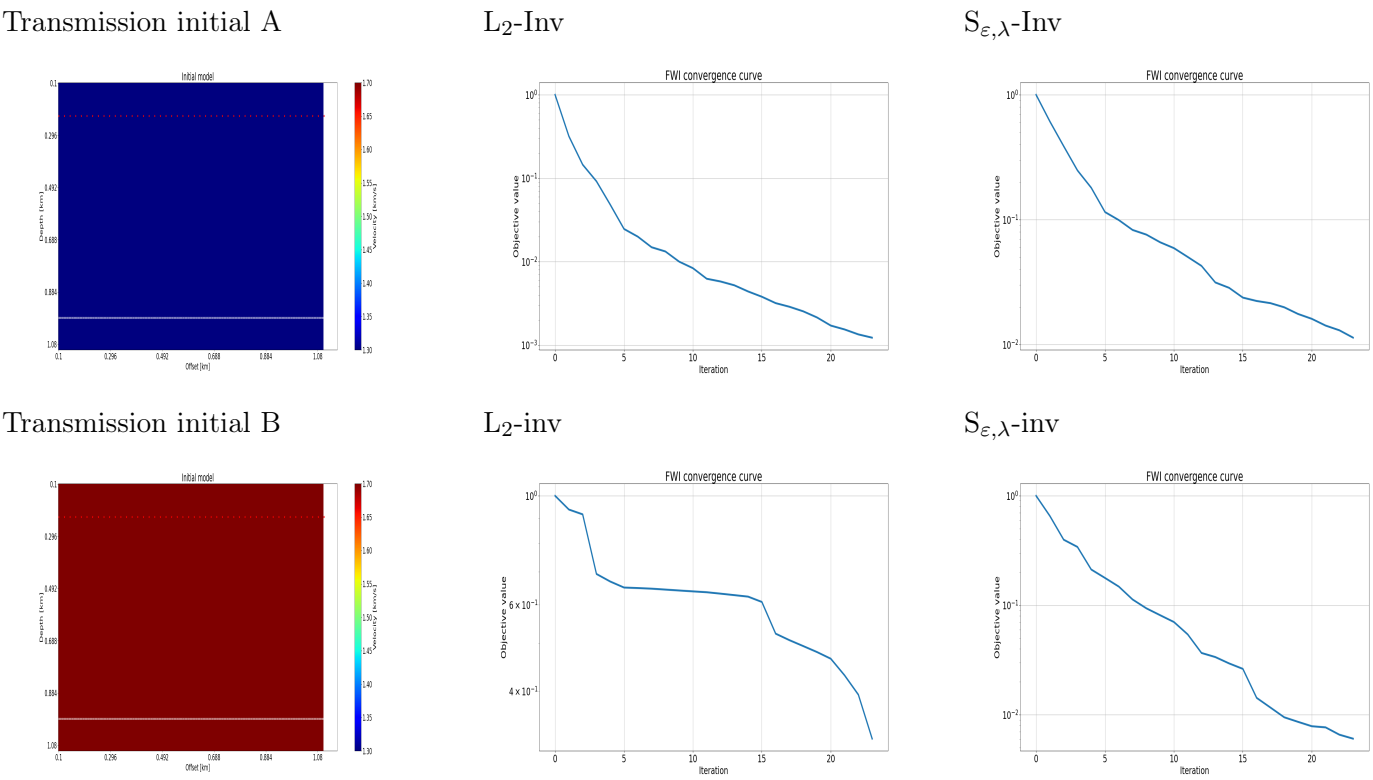


Figure 6.2: Inversion convergence

For initial model-B, the $S_{\varepsilon,\lambda}$ inversion is clearly improved after 100 iterations, as shown in Figure**??**. The peak amplitude of the velocity anomaly is now well recovered.

As expected, the resolution for both initial models is higher in the horizontal direction that in the vertical direction due to the acquisition geometry, with a vertical smoothing of the reconstructed velocity anomaly.

The shape of the inclusion is not fully recovered possibly due to limited offset and deficit in long-wavelength components of the transmitted waves and lack of reflected wave information.

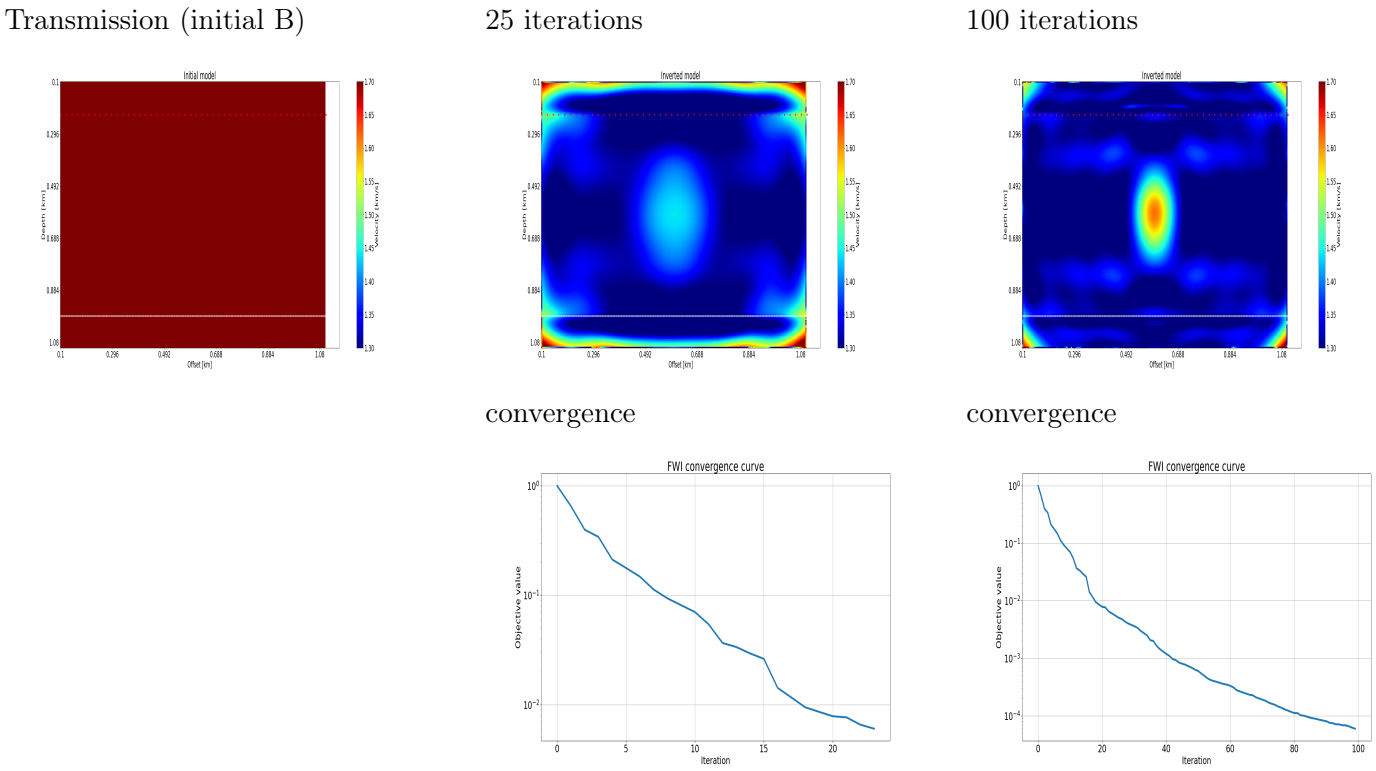


Figure 6.3: $S_{\varepsilon,\lambda}$ inversion: transmission (initial model B)

Initial model-A

Reconstructed models with the L_2 and the $S_{\varepsilon,\lambda}$ misfit functions are shown in Figure 6.4.

The L_2-misfit function suffers from cycle-skipping., and the inversion problem gets trapped in a local minimum and fails reconstructing the inclusion.

The $S_{\varepsilon,\lambda}$-misfit function provides reasonably accurate results, and the observed and predicted seismic traces are in phase after 25 iterations, as illustrated in Figure 6.5, even though the exact background velocity is not fully recovered.

As expected, the resolution is higher in the horizontal direction that in the vertical direction due to the acquisition geometry, leading to vertical smoothing of the reconstructed velocity anomaly.

The peak-amplitude of the velocity anomaly is not fully recovered after 25 iterations, see Figure 6.4 and Figure 6.6, as well as the exact background velocity. The shape of the inclusion is not fully recovered also possibly due to limited offset and relative deficit in long-wavelength components.

The initial and final predicted data, are shown in the time-receiver domain Figure 6.7. A diffraction pattern can be observed due to the relatively strong velocity anomaly in the inclusion.

The improved convexity of the $S_{\varepsilon,\lambda}$-misfit function avoids the FWI problem to be trapped in local minima and enable the convergence toward the global minimum.

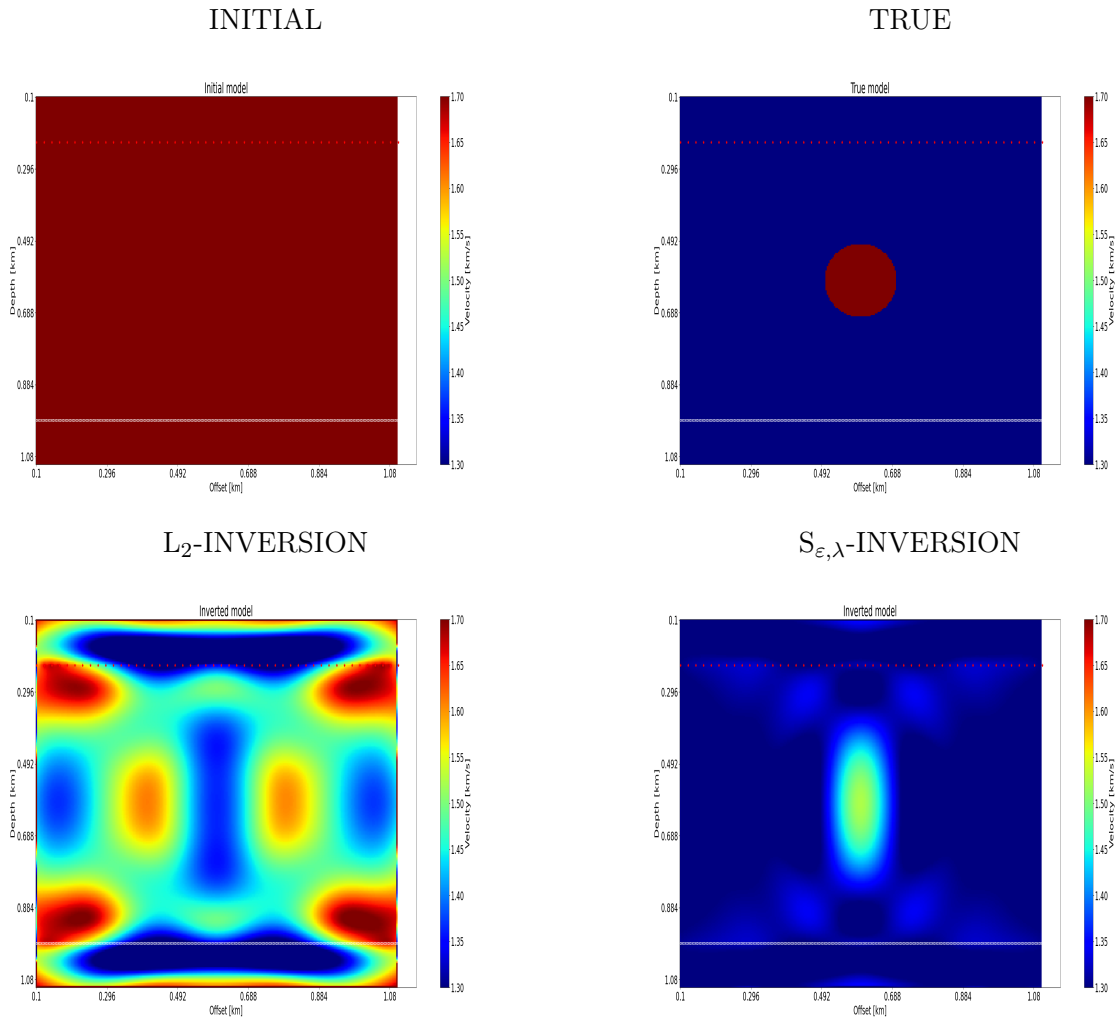


Figure 6.4: Final reconstructed models: initial model (upper-left figure) and targeted model (upper-right figure); FWI results with the L_2- misfit function (bottom-left figure) and the $S_{\varepsilon,\lambda}$-misfit function (bottom-right figure).

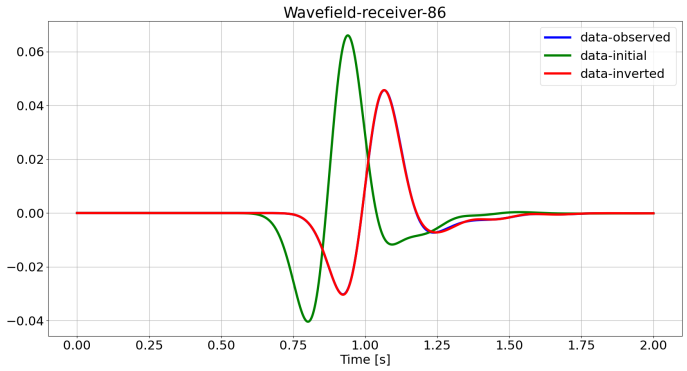


Figure 6.5: Seismic traces with the $S_{\varepsilon,\lambda}$-misfit function: initial (solid-green line) and predicted data (solid-red line) after 25 FWI iterations, together with observed data (solid-blue line).

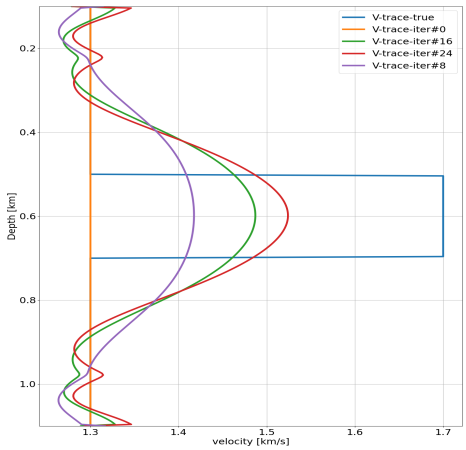


Figure 6.6: Velocity profile along a vertical section of the physical domain at a surface position $x = 600$ m as reconstructed by the $S_{\varepsilon,\lambda}$ based inversion.

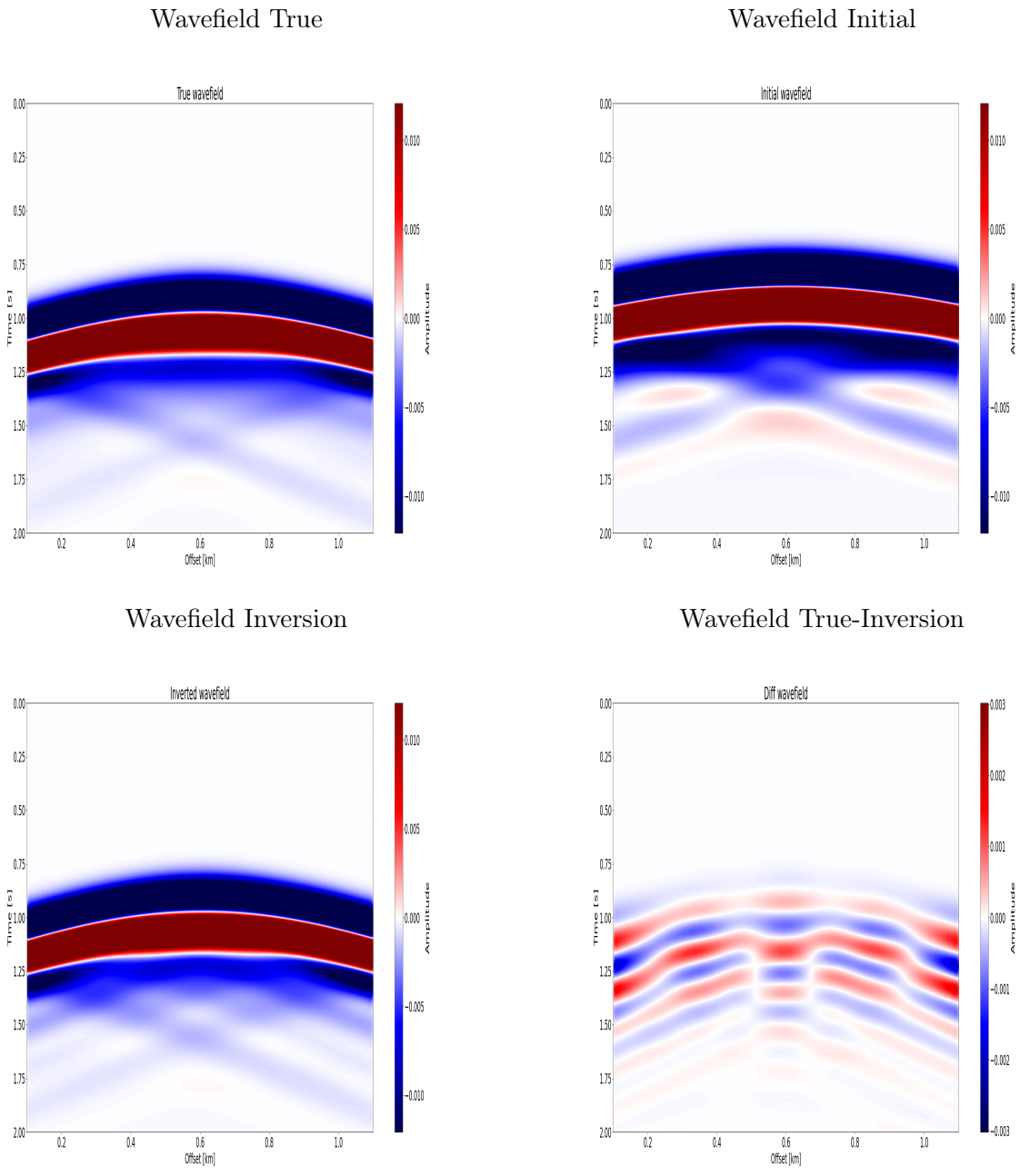


Figure 6.7: Wavefield in the time-receiver domain for the common shot gather associated to the centred source: "true" wavefield (upper-right figure); initial wavefield (upper-left figure); wavefield at the final FWI iteration (lower-right figure); difference between the "true" wavefield and the wavefield at the final FWI iteration (lower-left figure).

Initial model-B

Reconstructed models with the L_2 and the $S_{\varepsilon,\lambda}$ misfit functions are shown in Figure 6.8.

The L_2-misfit function now does not suffer from cycle-skipping, and both the L_2 and $S_{\varepsilon,\lambda}$ inversion reconstruct reasonably well the model. The observed and predicted seismic traces are in phase after 25 iterations, as illustrated for the $S_{\varepsilon,\lambda}$ based inversion in Figure 6.9.

The L_2 based inversion recovers the peak-amplitude of the inclusion velocity anomaly, whereas the peak-amplitude is not yet fully recovered (≈ 1550m/s) after 25 iterations with the $S_{\varepsilon,\lambda}$ based inversion.

The $S_{\varepsilon,\lambda}$ based inversion shows little sensitivity with respect to the initial model, see Figures 6.4 and 6.8.

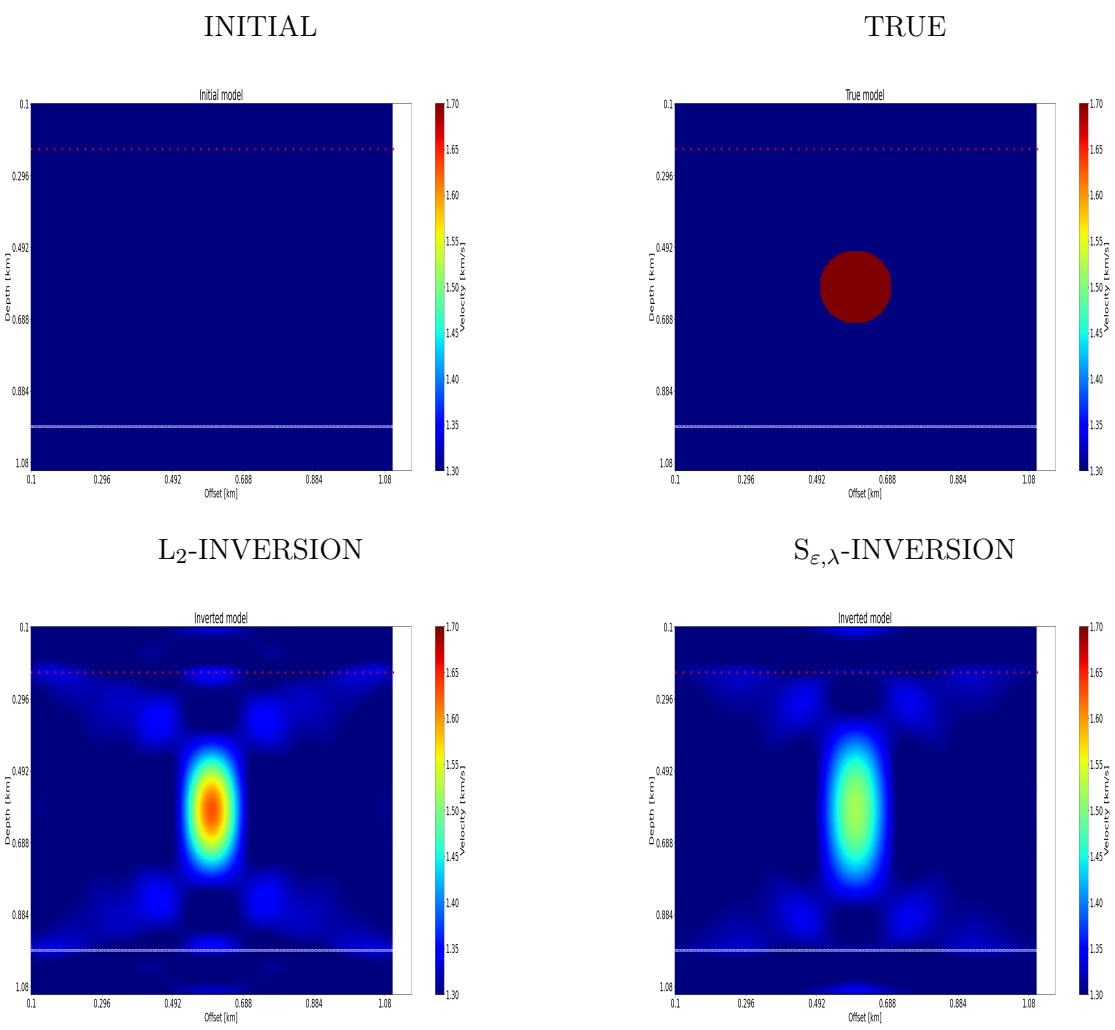


Figure 6.8: Final reconstructed models: initial model (upper-left figure) and targeted model (upper-right figure); FWI results with the L_2- misfit function (bottom-left figure) and the $S_{\varepsilon,\lambda}$-misfit function (bottom-right figure).

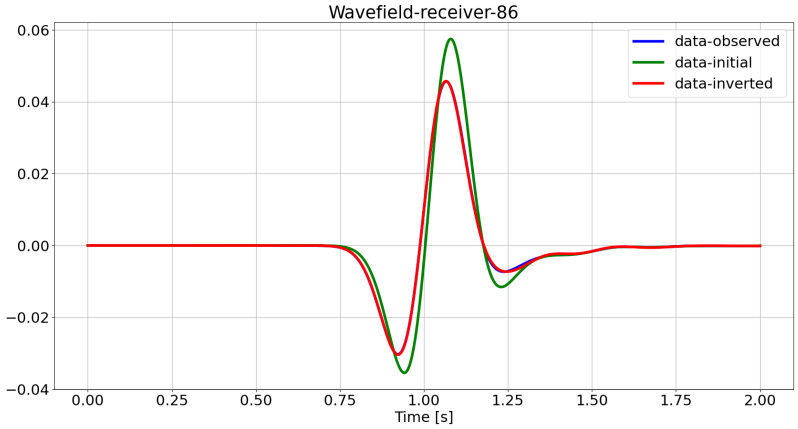


Figure 6.9: Seismic traces: initial (solid-green line) and predicted data (solid-red line) after 25 FWI iterations, together with observed data (solid-blue line).

With both the L_2 and $S_{\varepsilon,\lambda}$ based inversion, the shape of the inclusion is not fully reconstructed, due limited offset and relative deficit in long-wavelength components. The resolution in the horizontal direction is higher than in the vertical one, leading to vertical smoothing.

The L_2 and $S_{\varepsilon,\lambda}$ adjoint-sources are different, as shown in Figure 6.10, leading to different sensitivity kernels.

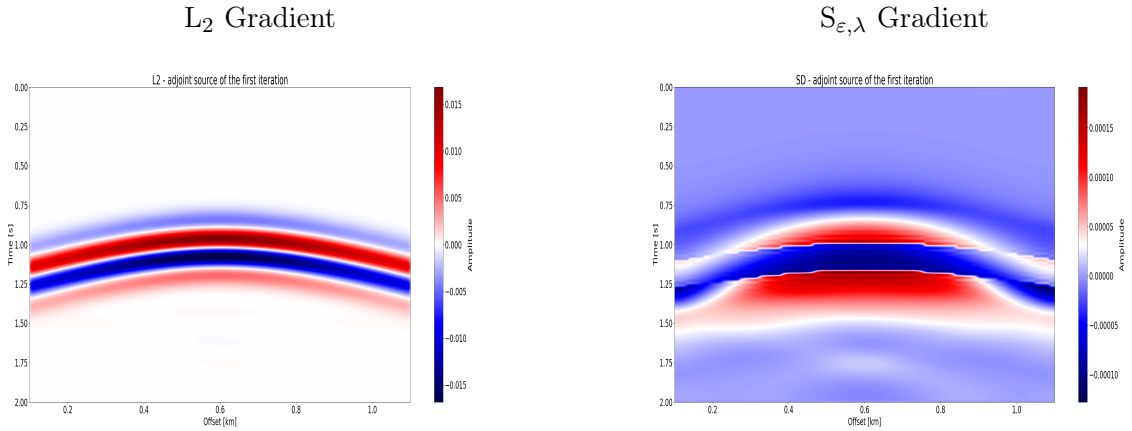


Figure 6.10: Ajoint-source in the time-receiver domain at the first FWI iteration.

The L_2 adjoint source is sparse and keeps memory of the Ricker wavelet due to the low kinematic errors associated with the initial model B.

The $S_{\varepsilon,\lambda}$ adjoint source is more complicated with the signature of a mass-spreading over the whole domain. This is primarily due to the penalisation of the marginals in the unbalanced Sinkhorn divergence formulation, and more specifically to the gradient of Kullback-Leibler divergence, and possibly to the separate 2D-transport of the positive and negative parts of the seismic signals.

The initial and final predicted data, in the time-receiver domain, are shown in Figure 6.11 together with the "true" data, *i.e.* the observed data. A diffraction pattern can be observed as well as a resolution deficit in the long wavelengths components, which precludes accurate reconstruction of the inclusion shape.

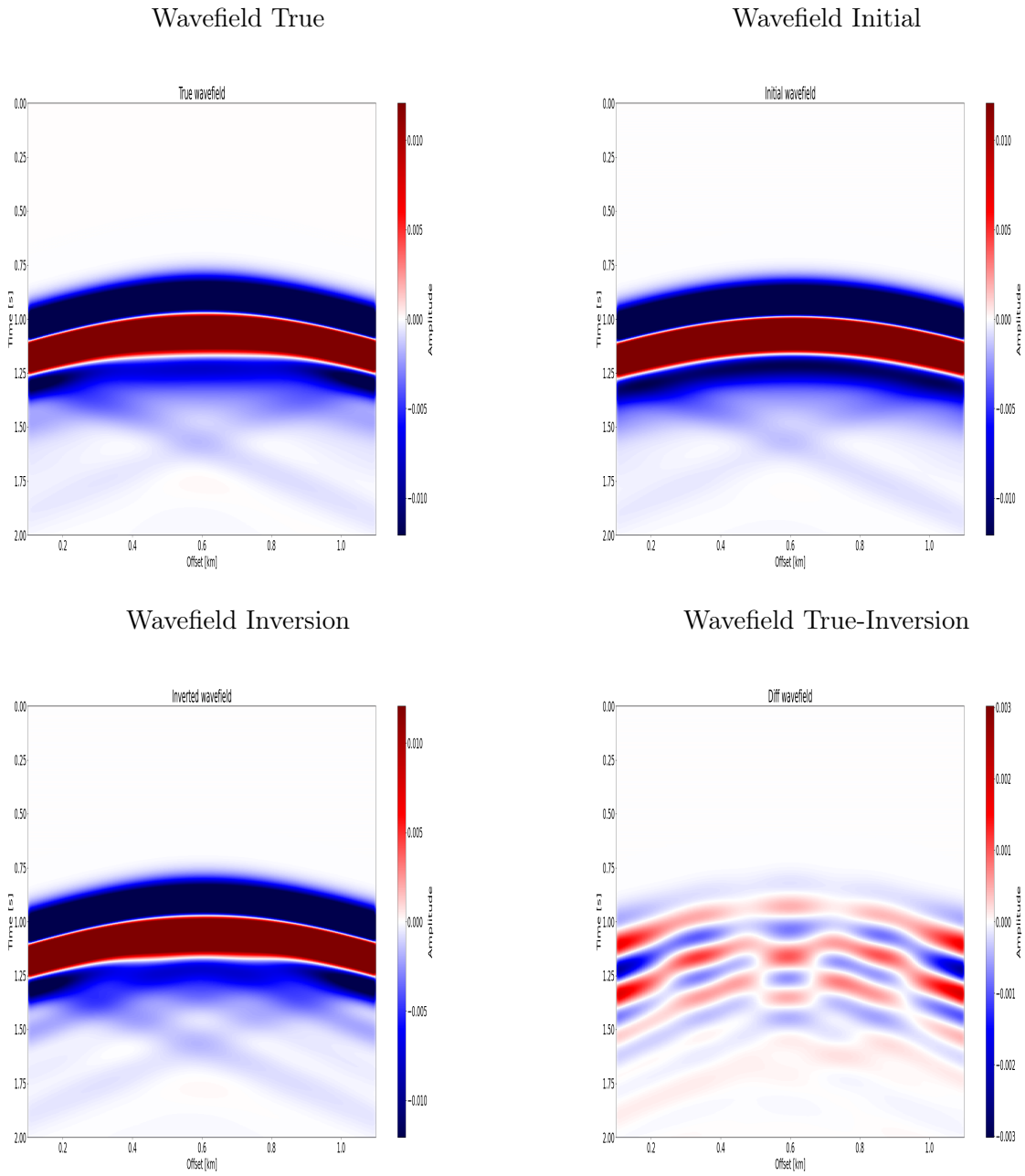


Figure 6.11: Wavefield in the time-receiver domain for the common shot gather associated to the centred source: "true" wavefield (upper-right figure); initial wavefield (upper-left figure); wavefield at the final FWI iteration (lower-right figure); difference between the "true" wavefield and the wavefield at the final FWI iteration (lower-left figure).

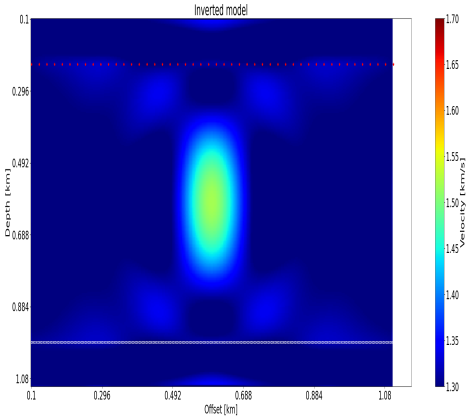
FWI results, using different implementations of the $S_{\varepsilon,\lambda}$-misfit function are shown in Figure 6.12 and Figure 6.13 for comparison.

A first implementation makes use of the squared-transformation, without normalisation, of the observed and predicted signals. With this signal transformation, the 2D $S_{\varepsilon,\lambda}$-misfit function leads to reasonably correct model reconstruction, and relatively improved estimation of the peak-amplitude of the velocity anomaly compared to the 2D $S_{\varepsilon,\lambda}$-misfit function based on the decomposition into positive and negative parts of the signals.

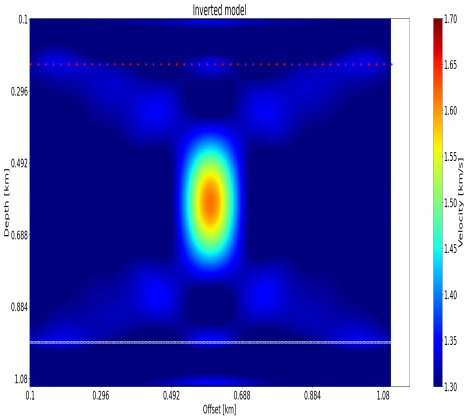
Another implementation makes use of the separability of the Sinkhorn kernel matrix and of appropriate scaling of the transport domain, see section 4.4.2. The $S_{\varepsilon,\lambda}$-misfit can be reduced to either 1-D unbalanced transport in time (trace-by-trace) or 1-D unbalanced transport in space (offset-by-offset).

Both leads to similar results that with the 2-D $S_{\varepsilon,\lambda}$-misfit function. The trace-by-trace 1D unbalanced transport in time appears to provide a better reconstruction of the peak-amplitude of the velocity anomaly, see for instance Figure 6.13.

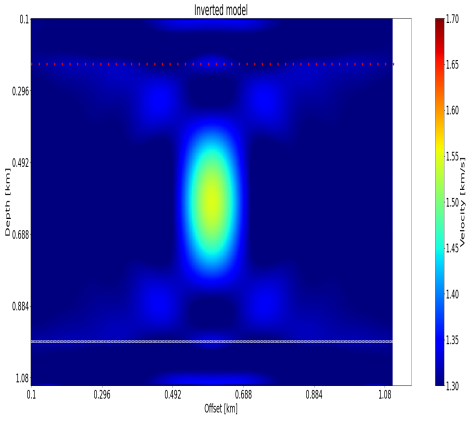
$S_{\varepsilon,\lambda}$ (2D - P_δ)



$S_{\varepsilon,\lambda}$ (2D-Squared Amp.)



$S_{\varepsilon,\lambda}$ (1D-Offset-P_δ)



$S_{\varepsilon,\lambda}$ (1D-Time-P_δ)

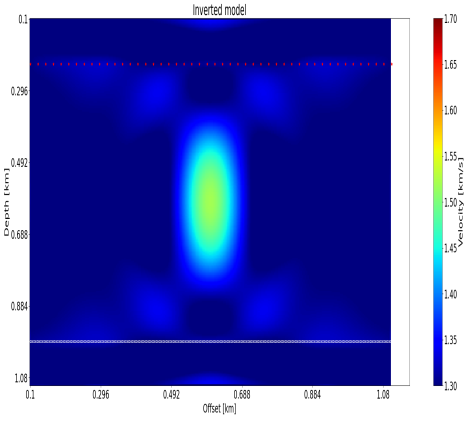


Figure 6.12: Reconstructed models for different implementation strategies of the $S_{\varepsilon,\lambda}$-misfit function: 2-D transport after decomposition of the signals into poistive and negative parts (upper-left figure); Squared-transform of the signals before 2-D transport (upper-left figure); 1-D transport in space (offset-by-offset) after decomposition of the signals into positive and negative parts (lower-left figure); 1-D transport in tile (trace-by-trace) (in time) after decomposition of the signals into positive and negative parts (lower-right figure). The later two implementations make use of the separability of the Gibbs matrix in the Sinkhorn algorithm and of the appropriate scaling of the transport domain.

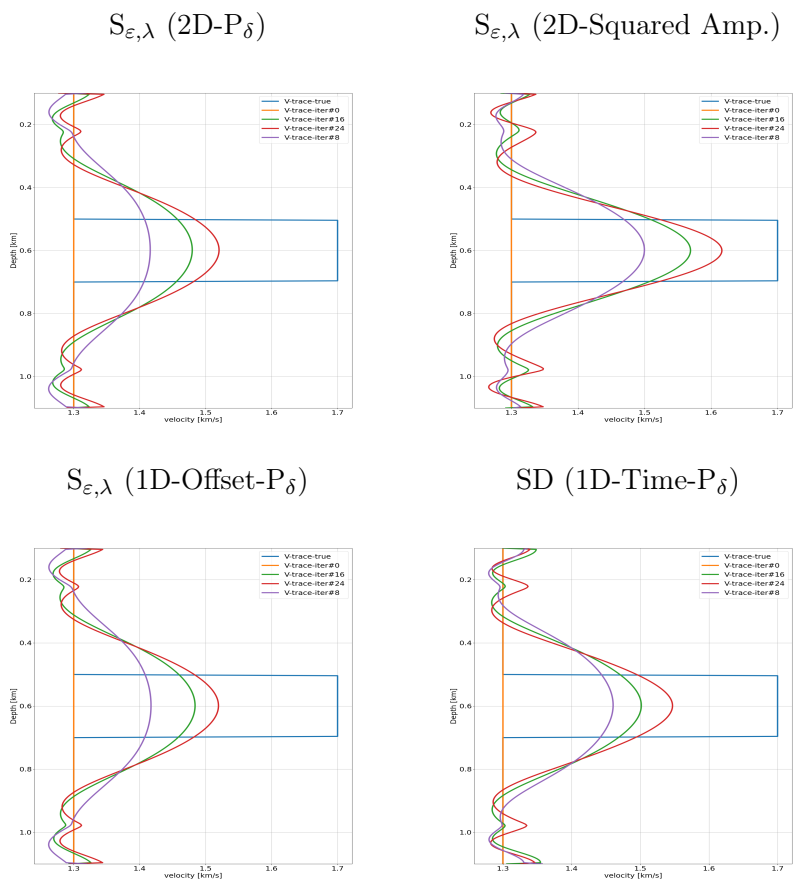


Figure 6.13: Velocity profile along a vertical section of the physical domain for different FWI iterations and different implementations of the $S_{\varepsilon,\lambda}$-misfit function.

6.2.2 Reflection configuration

The acquisition system is now composed of 48 evenly-spaced sources along the upper boundary of the physical domain, at 200m depth, and of 256 evenly-spaced receivers located at the same depth position that the sources. The source-time function of the point sources is a Ricker wavelet with a central frequency of $f_c = 3$Hz, and the recording time is set to 2s.

Initial model-A

Reconstructed models with L_2 and $S_{\varepsilon,\lambda}$ based inversion are shown in Figure 6.14.

The L_2-misfit function suffers, as expected, from cycle-skipping, and the inversion problem gets trapped in a local minimum and fails reconstructing the model, while the improved convexity of the $S_{\varepsilon,\lambda}$-misfit function avoids this pittfall.

The $S_{\varepsilon,\lambda}$ based inversion provides a meaningful model reconstruction after 25 iterations, and the data are in phase, see Figure 6.15.

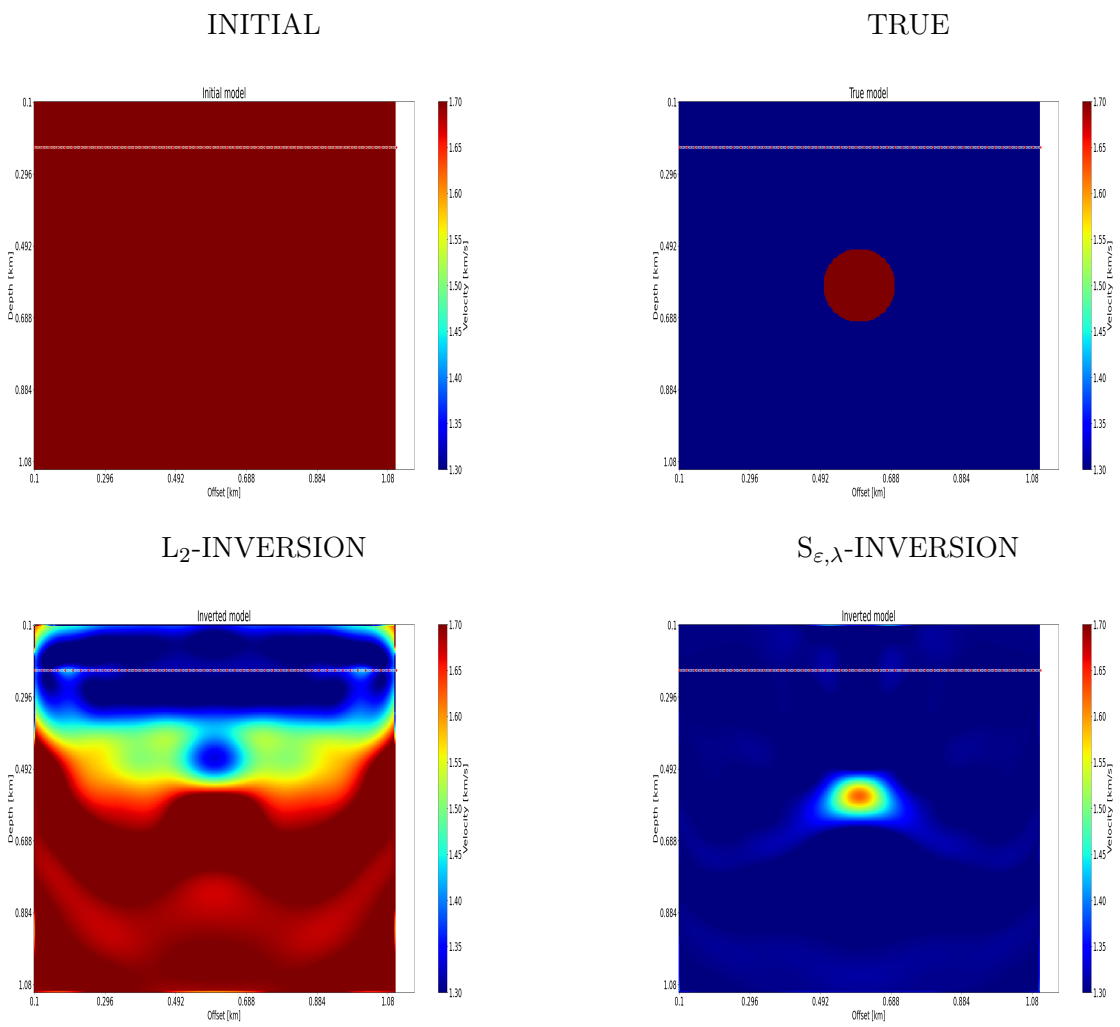


Figure 6.14: Final reconstructed models: initial model (upper-left figure) and targeted model (upper-right figure); FWI results with the L_2- misfit function (bottom-left figure) and the $S_{\varepsilon,\lambda}$-misfit function (bottom-right figure).

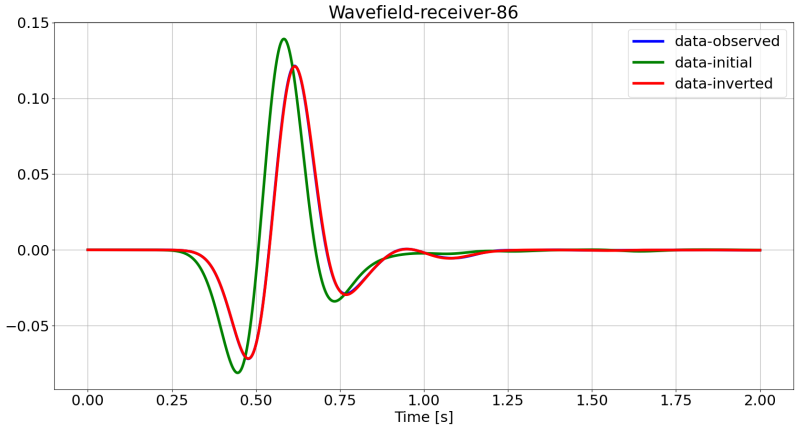


Figure 6.15: Seismic traces: initial (solid-green line) and predicted data (solid-red line) after 25 FWI iterations, together with observed data (solid-blue line).

The peak-amplitude of the velocity anomaly is reasonably well retrieved, see Figure 6.17. The upper interface and horizontal extension of the inclusion is meaningful, whereas the lower inclusion boundary is not properly resolved and appears as "melted" because reflections from the lower inclusion boundary have not been sufficiently recorded due to the limited offset.

The initial and final predicted data are shown in the time-receiver domain Figure 6.18. Reflection phase associated to the lower boundary of the inclusion, and the weaker multiple reflections behind, are poorly resolved. The background velocity also is not exactly retrieved.

The $S_{\varepsilon,\lambda}$ based inversion exhibits a monotonous convergence rate and the misfit keeps continuing to decrease after 25 iterations, see Figure 6.16.

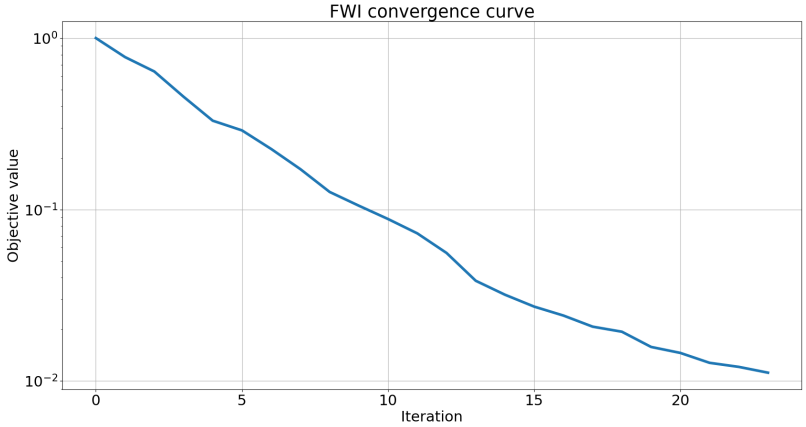


Figure 6.16: Convergence rate for the $S_{\varepsilon,\lambda}$ based inverstuin

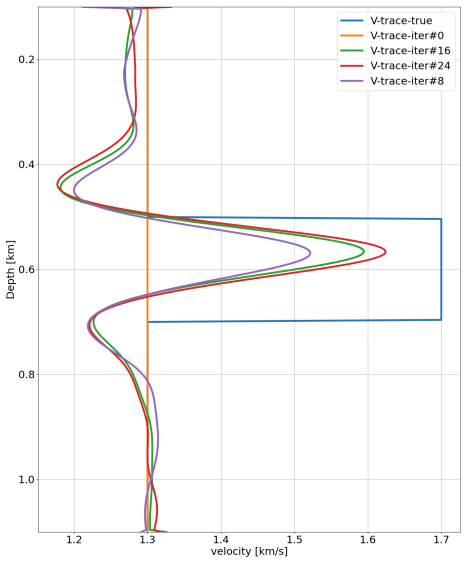


Figure 6.17: Velocity profile along a vertical section of the physical domain for $S_{\varepsilon,\lambda}$.

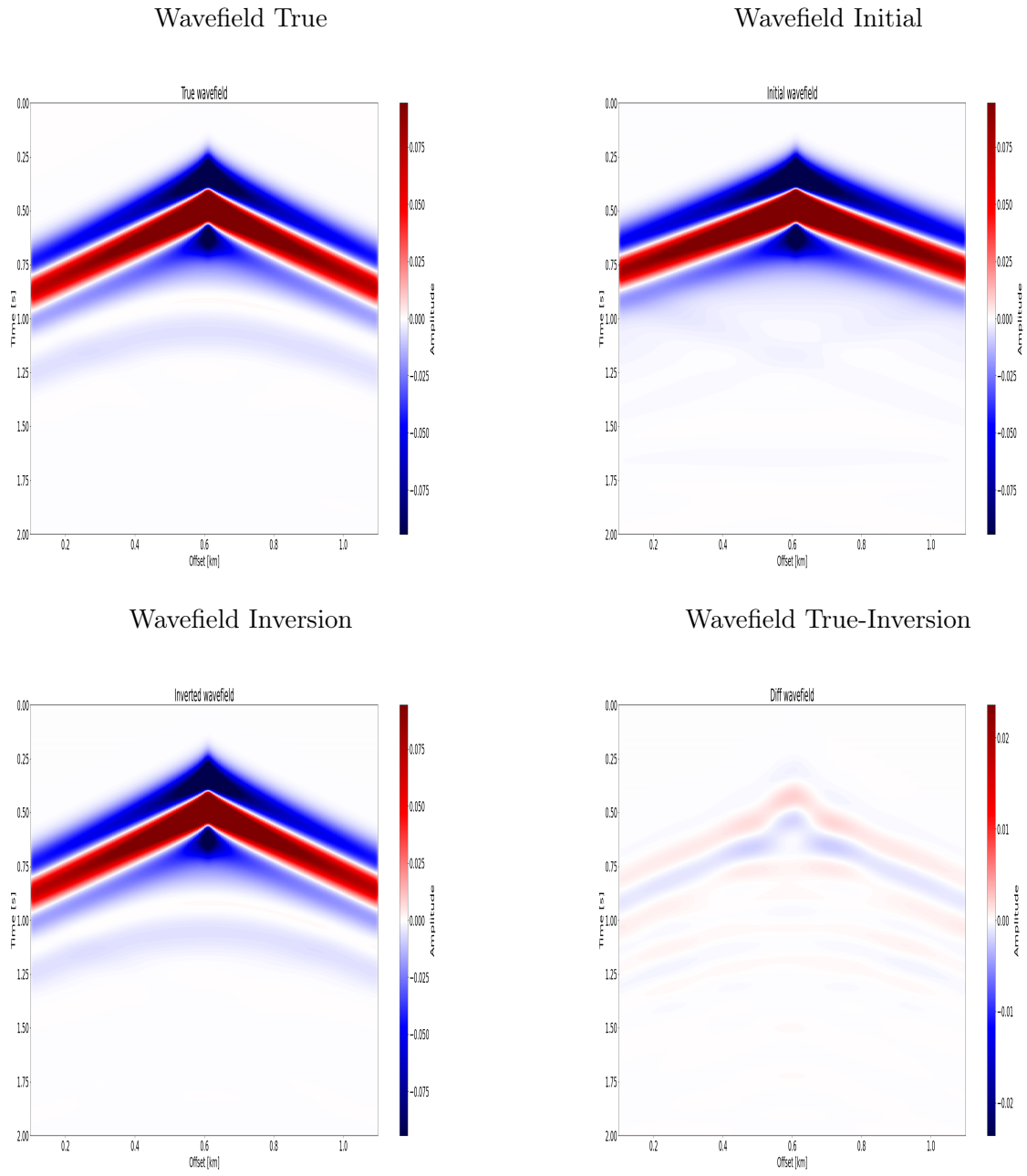


Figure 6.18: Wavefield in the time-receiver domain for the common shot gather associated to the centred source: "true" wavefield (upper-right figure); initial wavefield (upper-left figure); wavefield at the final FWI iteration (lower-right figure); difference between the "true" wavefield and the wavefield at the final FWI iteration (lower-left figure).

Initial model B

Reconstructed models after 25 iterations with the L_2 and the $S_{\varepsilon,\lambda}$ based inversion are shown in Figure 6.19.

The L_2-misfit function now does not suffer from cycle-skipping, and both the $L2$ and $S_{\varepsilon,\lambda}$ based inversion reconstruct meaningful models. The observed and predicted data are in phase, see Figure 6.9 for the $S_{\varepsilon,\lambda}$-misfit function.

The peak-amplitude of the velocity anomaly is not yet fully recovered after 25 iterations with the L_2-misfit function, whereas with the $S_{\varepsilon,\lambda}$-misfit function a relatively more accurate peak-amplitude is recovered, see Figures 6.19 and 6.21.

In such a control case, the L_2 based inversion should recover the peak-amplitude of the inclusion anomaly

when letting the inversion iterate to convergence.

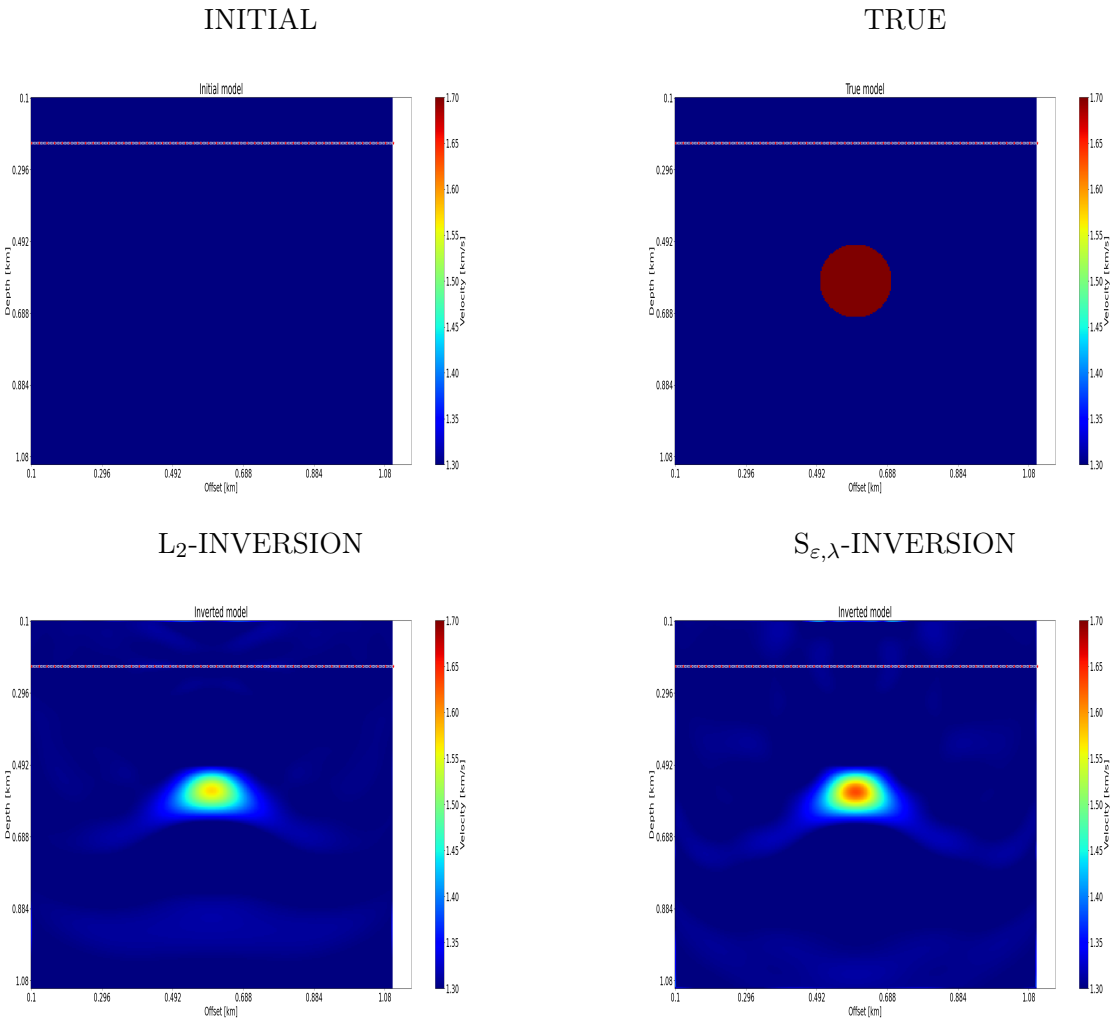


Figure 6.19: Final reconstructed models: initial model (upper-left figure) and targeted model (upper-right figure); FWI results with the L_2- misfit function (bottom-left figure) and the $S_{\varepsilon,\lambda}$-misfit function (bottom-right figure).

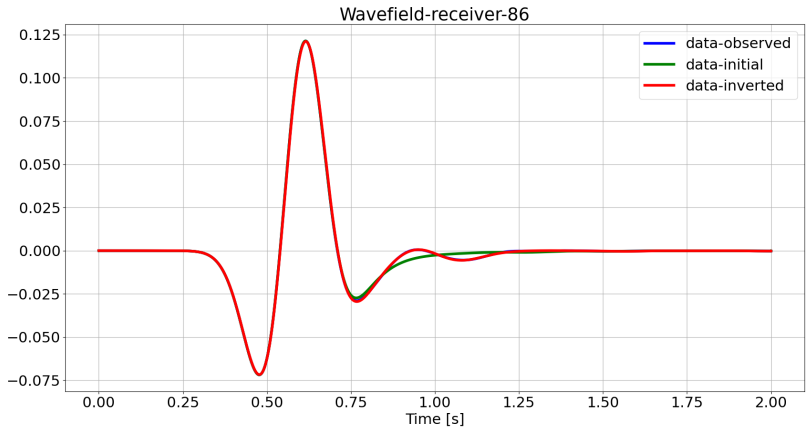


Figure 6.20: Seismic traces: initial (solid-green line) and predicted data (solid-red line) after 25 FWI iterations, together with observed data (solid-blue line).

In both cases, the position of the upper interface and the extension of the inclusion are meaningful. The lower interface of the inclusion is not recovered and appears as melted due to limited offset.

The L_2 and $S_{\varepsilon,\lambda}$ adjoint sources are again quite different, as shown in Figure 6.22.

L_2-misfit $S_{\varepsilon,\lambda}$-misfi

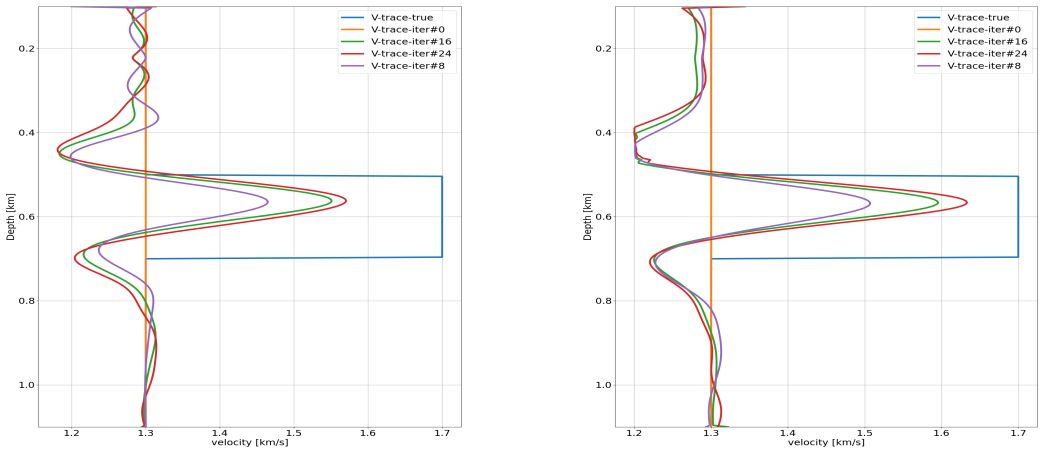


Figure 6.21: Velocity profile along a vertical section of the physical domain at surface position $x = 600$ m at different inversion iterations with the L_2 and the $S_{\varepsilon,\lambda}$ misfit functions.

L_2 Gradient $S_{\varepsilon,\lambda}$ Gradient

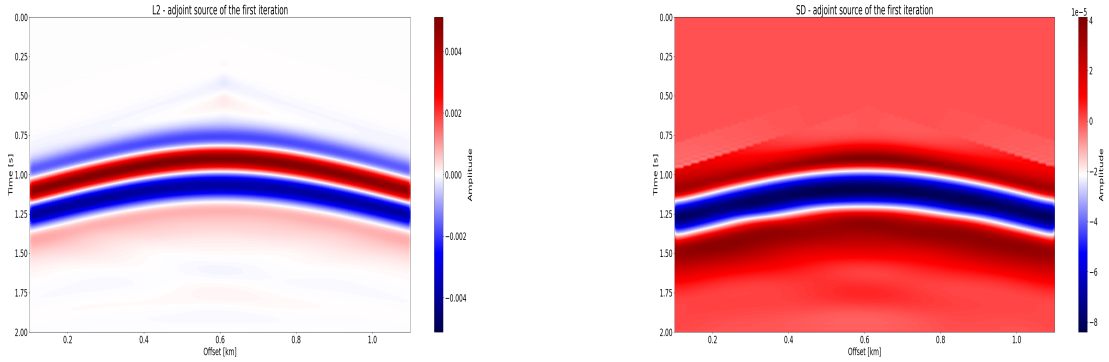


Figure 6.22: Ajoint-source in the time-receiver domain at the first FWI iteration.

The L_2 adjoint source is rather sparse and keeps memory of the Ricker wavelet. In the $S_{\varepsilon,\lambda}$ adjoint source the signature of a mass-spreading over the whole domain can be seen, the origin of which is related to the gradient of the Kullback-Leibler divergence in the Sinkhorn divergence formulation of unbalanced entropic transport, and possibly to the separate transport of the positive and the negative parts of the seismic signals.

The initial and final predicted data are shown in the time-receiver domain Figure 6.23, which again show poor resolution at the lower inclusion interface, and tonly a partial reconstruction of the background velocity model.

Convergence rate of the L_2 and $S_{\varepsilon,\lambda}$ based inversion is shown in Figure 6.26. In both cases the convergence rate is monotic and the misfit distance keep continuing to decrease after 25 iterations.

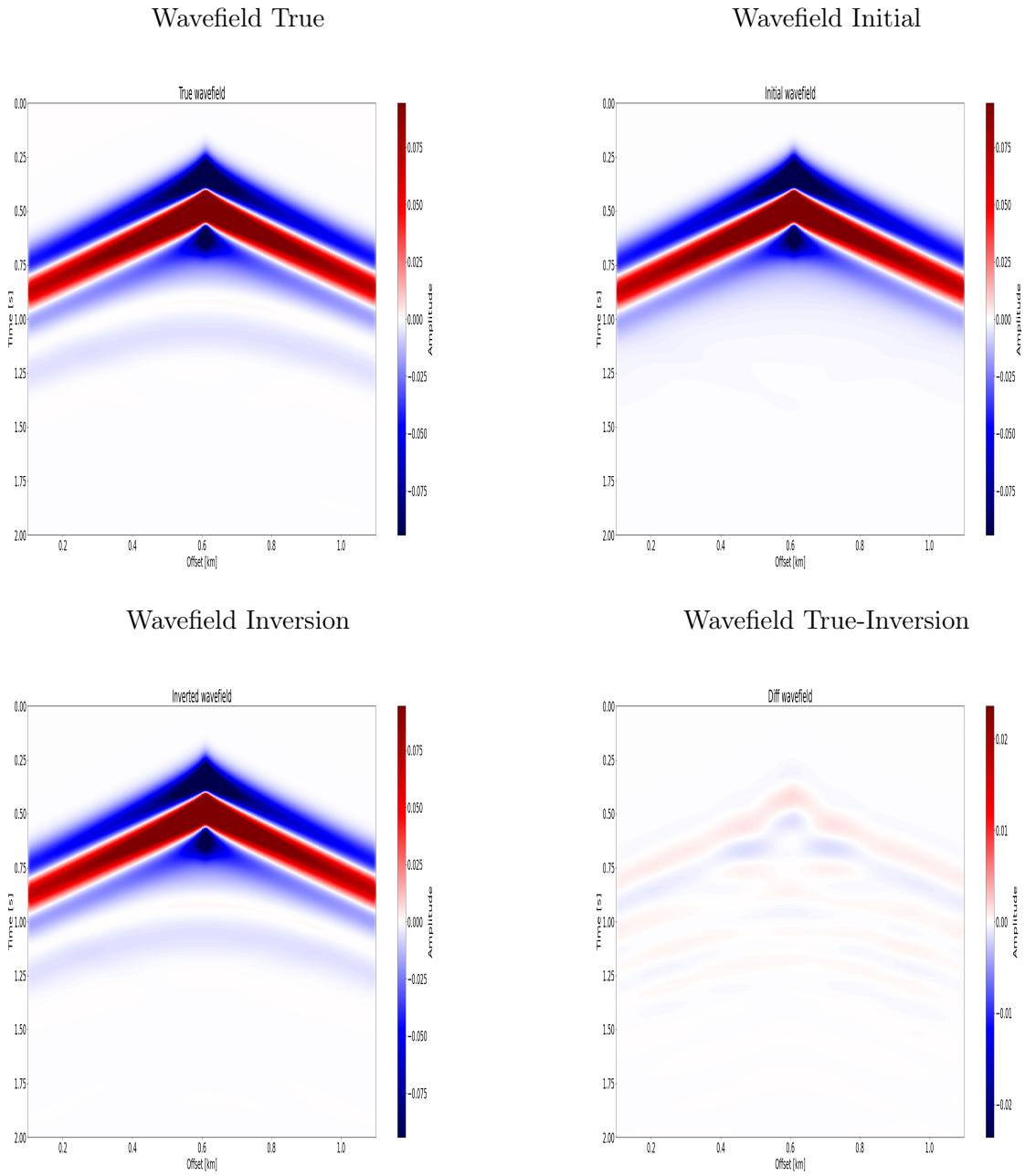
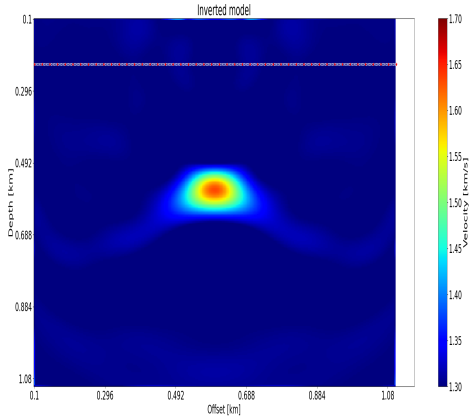


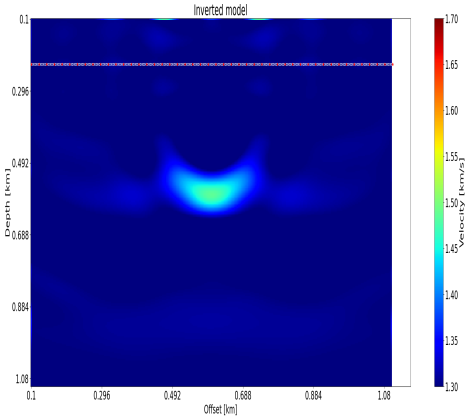
Figure 6.23: Wavefield in the time-receiver domain for the common shot gather associated to the centred source: "true" wavefield (upper-right figure); initial wavefield (upper-left figure); wavefield at the final FWI iteration (lower-right figure); difference between the "true" wavefield and the wavefield at the final FWI iteration (lower-left figure).

Different implementations of the $S_{\varepsilon,\lambda}$-misfit function are compared in Figures 6.24 and in 6.25. Both the trace-by-trace, 1-D unbalanced transport in time, and the offset-by-offset, 1-D unbalanced transport in space, implementations of the $S_{\varepsilon,\lambda}$-misfit function provide meaningful results after 25 iterations. The peak-amplitude of the velocity anomaly is better estimated with the trace-by-trace, 1D unbalanced transport in time, implementation than with the offset-by-offset, 1D unbalanced transport in space. As expected, the 2D $S_{\varepsilon,\lambda}$-misfit function based on the squared-transformed signals fails and does not recover a meaningful model.

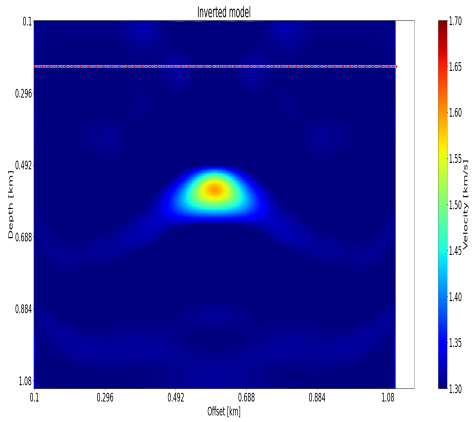
$S_{\varepsilon,\lambda}$ (2D - P_δ)



$S_{\varepsilon,\lambda}$ (2D-Squared Amp.)



$S_{\varepsilon,\lambda}$ (1D-Offset-P_δ)



$S_{\varepsilon,\lambda}$ (1D-Time-P_δ)

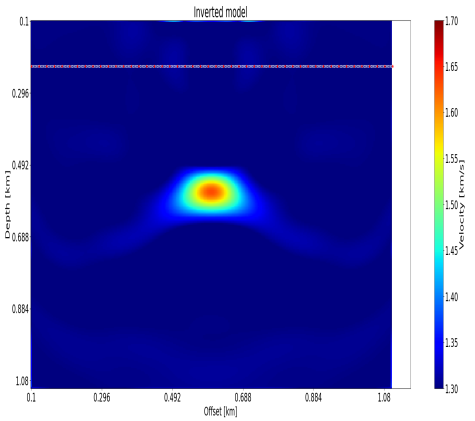


Figure 6.24: Reconstructed models for different implementation strategies of the $S_{\varepsilon,\lambda}$-misfit function: 2-D transport after decomposition of the signals into poistive and negative parts (upper-left figure); Squared-transform of the signals before 2-D transport (upper-left figure); 1-D transport in space (offset-by-offset) after decomposition of the signals into positive and negative parts (lower-left figure); 1-D transport in tile (trace-by-trace) (in time) after decomposition of the signals into positive and negative parts (lower-right figure). The later two implementations make use of the separability of the Gibbs matrix in the Sinkhorn algorithm and of the appropriate scaling of the transport domain.

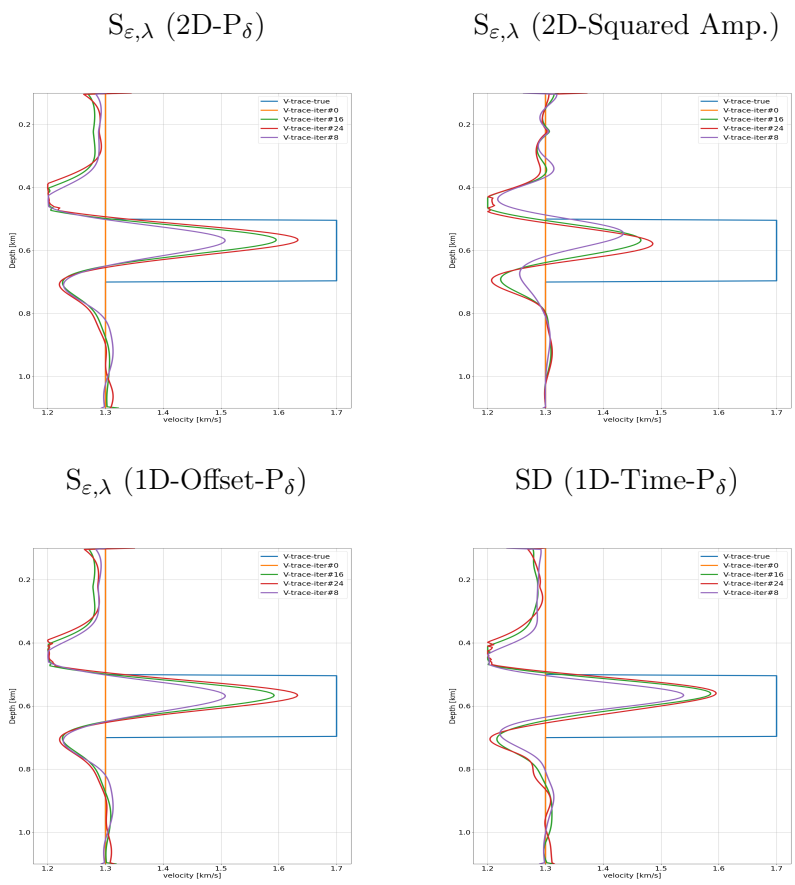


Figure 6.25: Velocity profile along a vertical section of the physical domain for different FWI iterations and different implementations of the $S_{\varepsilon,\lambda}$-misfit function.

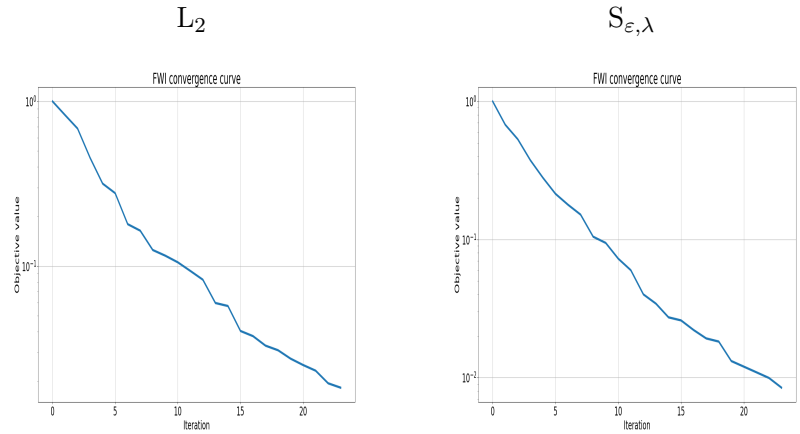


Figure 6.26: Convergence rate of the L_2 (left) and $S_{\varepsilon,\lambda}$ (right) based inversion

6.3 Layered reflection Models

6.3.1 One-layer reflection model

The one-layer model is inspired from (*Pladys et al.*, 2021).

The physical model is defined on a rectangular domain of 2000×1000 m dimension, and is parametrised by a homogeneous background acoustic wave velocity set to1500 m/s and by a 100 m thick layer at 300 m depth with a homogeneous wave velocity of 1600 m/s (postive anomaly) or 1400 m/s (negative anomaly).

The surface acquisition is the same for both models with 48 evenly-spaced seismic sources and 251 evenly-spaced receivers, both at 42 m depth below the upper boundary of the physical domain. The sources are point sources and the source-time function is a Ricker wavelet with central frequency f_c = 3Hz, and the recording time is set to 4 s.

The computational domain includes additional perfectly matched layers along each boundary of the physical domain. The constant-density acoustic wave equation is solved in space on a regular grid with 256×128 points using a sixth-order finite difference scheme and a second-order leap-frog integration scheme in time with constant time step satisfying the CFL condition, as implemented in PySIT.

The FWI problem is solved by a l-BFGS gradient-based method, as implemented in PySIT, with a fixed number of iterations set to 50. The initial model for both model configurations is a homogeneous model with wave velocity set to the correct background velocity 1500 m/s.

Both the predicted $\mathbf{d}_{\text{cal}}$ and the observed $\mathbf{d}_{\text{obs}}$ data, *i.e.* the latter corresponding to the "true" model, are physical realisations obtained with the same wave propagation solver, without considering additional noise to the data, *i.e.* the inversion crime setting.

Unless otherwise specified, the $S_{\varepsilon,\lambda}$-misfit function is formulated as discussed in Chapter 5. The observed and predicted signals are decomposed into positive and negative parts, with the P_δ - transform defined in (5.4), and transported separately using the Sinkhorn divergence formulation for unbalanced optimal transport.

One-layer reflection model: positive velocity anomaly

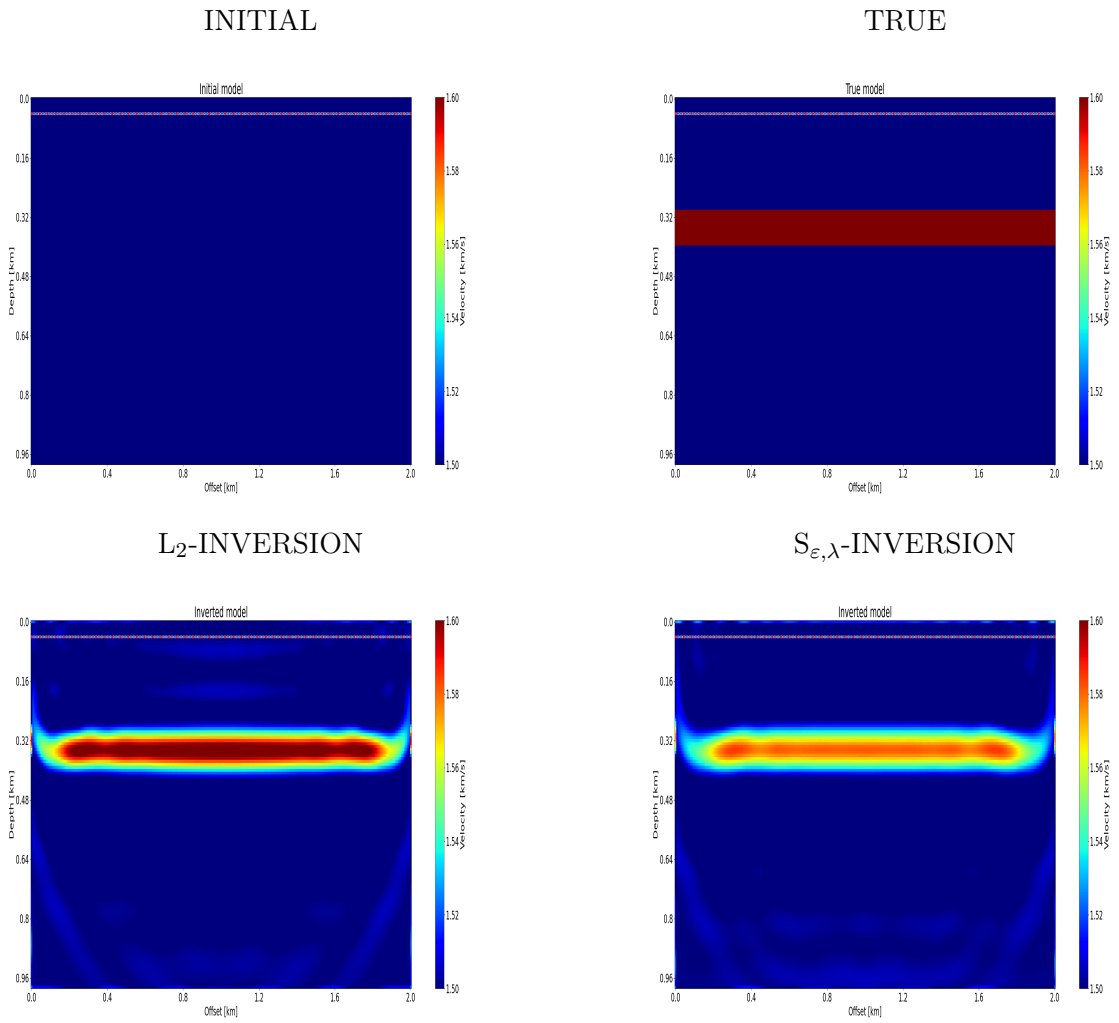


Figure 6.27: Final reconstructed models: initial model (upper-left figure) and targeted model (upper-right figure); FWI results with the L_2- misfit function (bottom-left figure) and the $S_{\varepsilon,\lambda}$-misfit function (bottom-right figure).

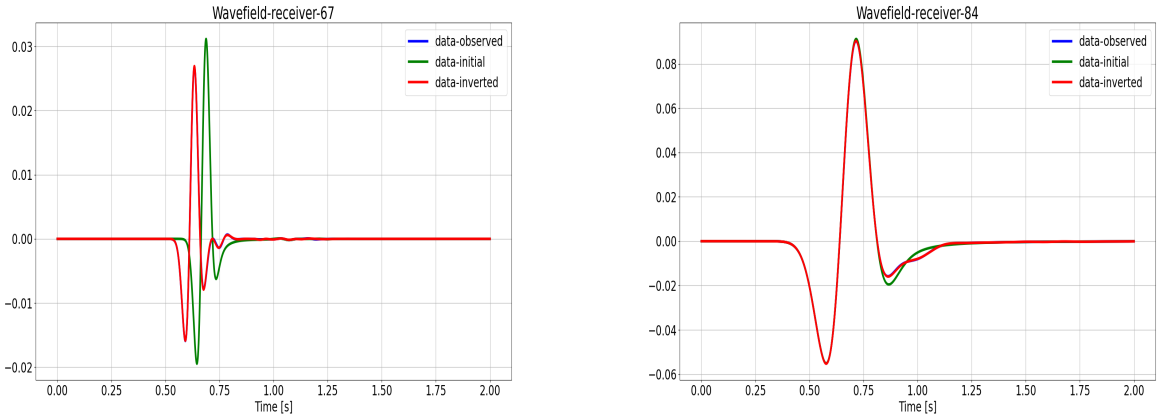


Figure 6.28: Velocity profiles reconstructed with the L_2 and the $S_{\varepsilon,\lambda}$ based inversion at receivers 67 and 84: initial (solid-green line) and predicted data (solid-red line) after 50 iterations, together with the velocity profile in the "True" model (solid-blue line).

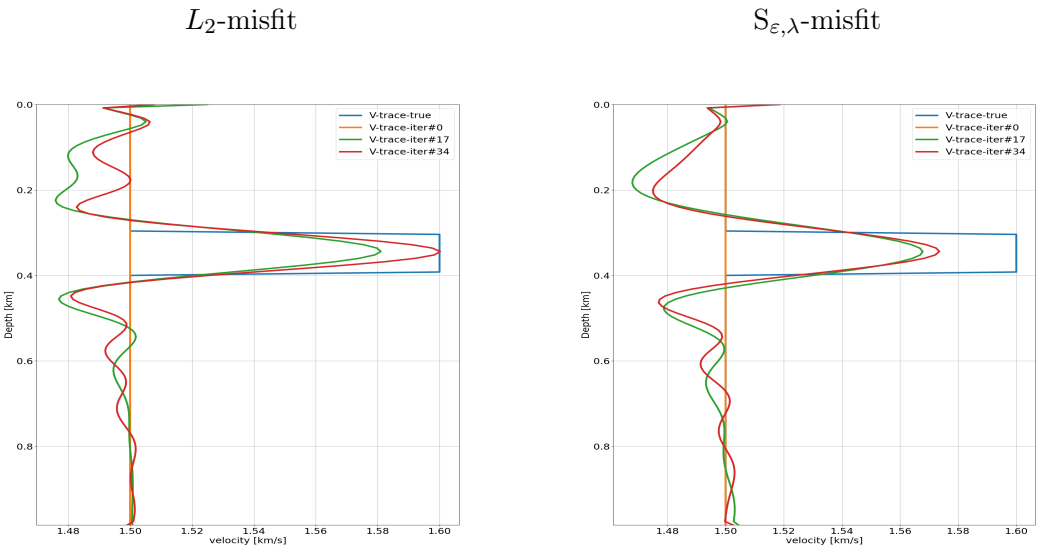


Figure 6.29: Velocity profiles in the physical domain at surface position $x = 1000$ m for different iterations of the L_2 and $S_{\varepsilon,\lambda}$ based inversion.

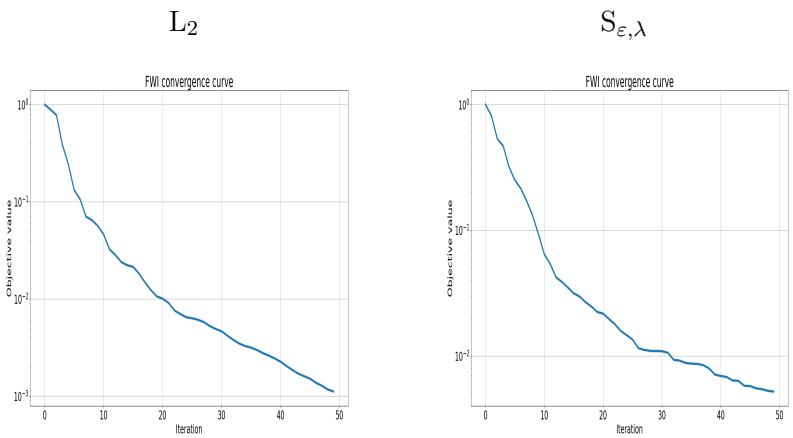


Figure 6.30: Convergence rate with the L_2 (left) and the $S_{\varepsilon,\lambda}$ (right) based inversion.

Reconstructed models, after 50 iterations, with the L_2 and $S_{\varepsilon,\lambda}$ based inversion are shown in Figure 6.27. The L_2 and $S_{\varepsilon,\lambda}$ based inversion provide similar meaningful reconstructions.

Results for the L_2 based inversion were expected since the L_2-misfit function is sensitive to amplitude variation and polarity. Results for the $S_{\varepsilon,\lambda}$-based inversion are less intuitive since the $S_{\varepsilon,\lambda}$-misfit function is based on signal decomposition into positive and negative parts and does not guarantees the preservation of phase and polarity information. The data are in phase for the $S_{\varepsilon,\lambda}$ based inversion, after 50 iterations, see Figure 6.28.

The initial and final predicted data for both L_2 and $S_{\varepsilon,\lambda}$ based inversion are shown in the time-receiver domain Figure 6.32.

The peak-amplitude of the velocity anomaly is correctly retrieved by the L_2 based inversion after 50 iterations, but only partially retrieved by the $S_{\varepsilon,\lambda}$ based inversion, see Figure 6.29. The $S_{\varepsilon,\lambda}$ based inversion recovers only half of the anomaly, with relatively more vertical smoothing than in the L_2 case, see also Figure 6.27.

This could possibly result from the entropic regularisation, which tends to spread mass and smooth discontinuity. The $S_{\varepsilon,\lambda}$ based inversion shows slower convergence rate compared to the L_2 based inversion, see Figure 6.30, but the misfit distance keep continuing to decrease after 50 iterations.

The L_2 and $S_{\varepsilon,\lambda}$ adjoint sources and optimisation gradients, at the first iteration, are shown in Figure 6.31. The $S_{\varepsilon,\lambda}$ adjoint source shows the signature of a mass-spreading over the whole domain as previously discussed, and the gradient is smoother than the L_2 one, even though both capture similar features.

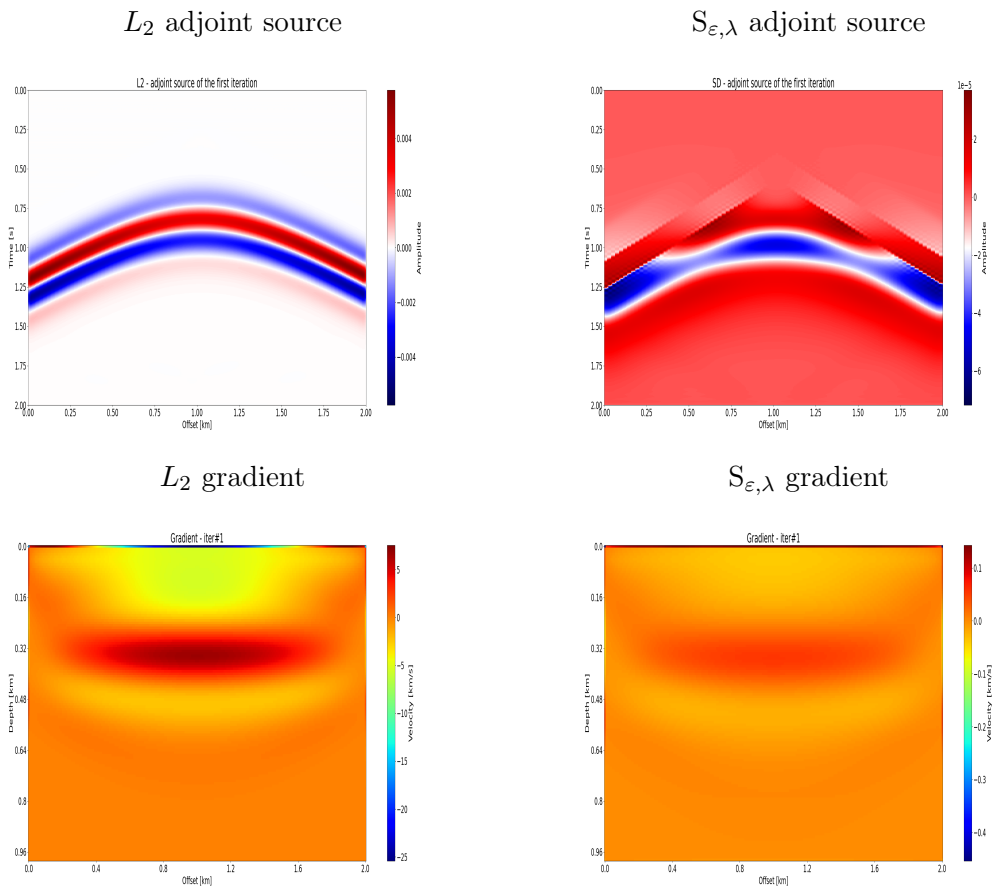


Figure 6.31: Adjoint source and gradient at first iteration for the L_2 and $S_{\varepsilon,\lambda}$ misfit functions.

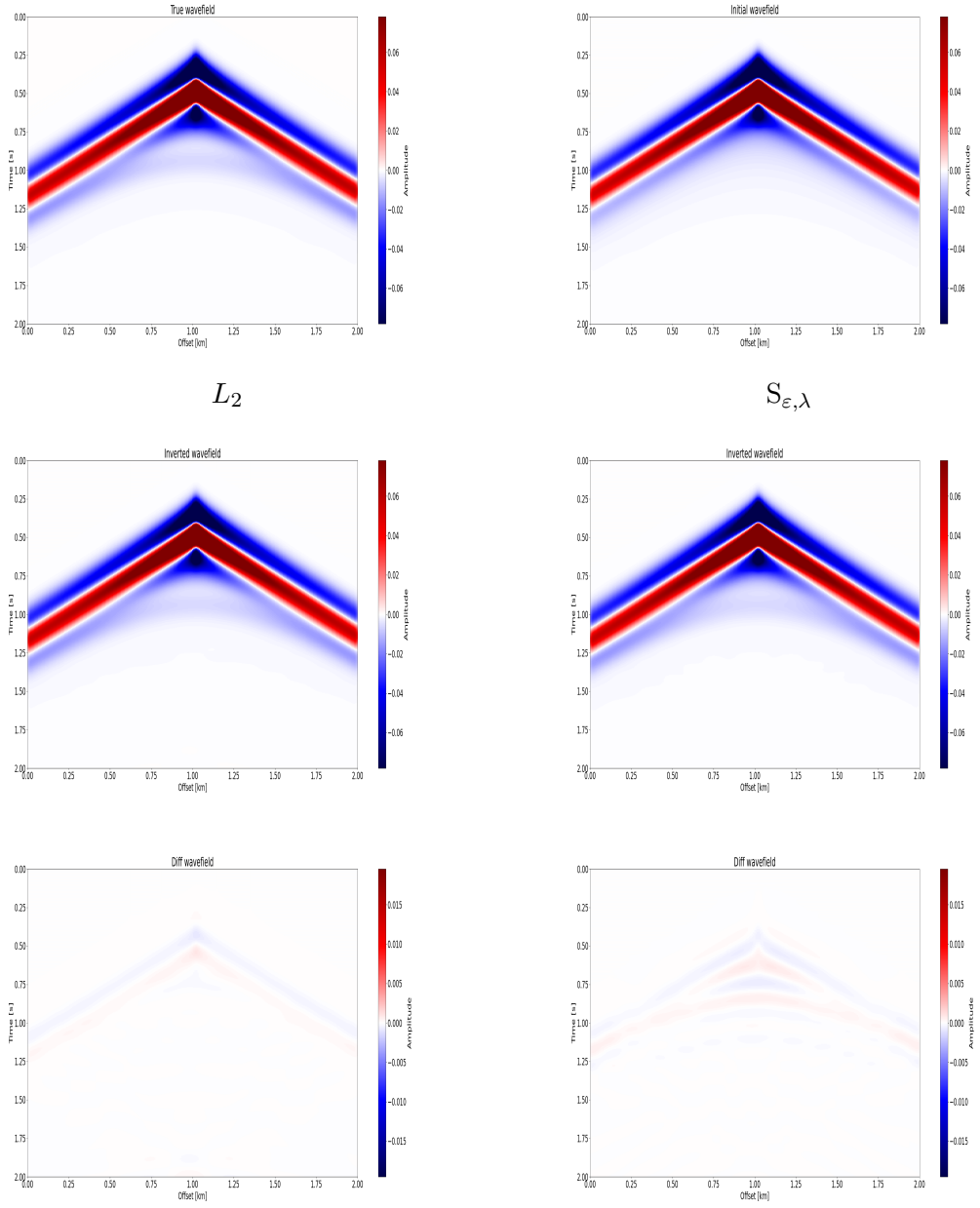


Figure 6.32: Wavefield in the time-receiver domain for a common shot gather associated to the centred source: "true" wavefield (upper-right figure); initial wavefield (upper-left figure); inverted wavefield and difference between the "true" wavefield and inverted wavefield for L_2 (lower-left figures) and Sε, λ (lower-right figures).

Inversion results for different implementations of the $S_{\epsilon,\lambda}$-misfit function are shown in Figure 6.33

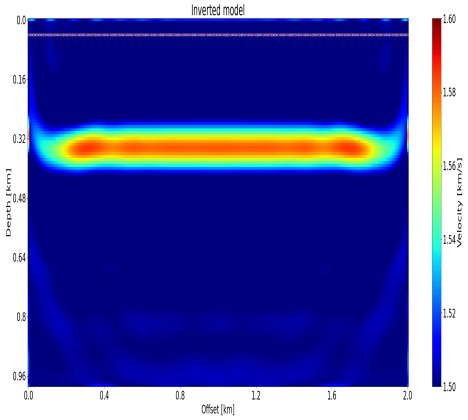
Somewhat surprisingly, meaningful model reconstructions, similar to results of the L_2 based inversion, are obtained with the 2D $S_{\varepsilon,\lambda}$-misfit function based on squared-amplitude transformation of the signals, for which reflected waves polarity is lost. The peak-amplitude of the velocity anomaly is well reconstructed after 50 iterations, and improved with respect to the 2-D $S_{\varepsilon,\lambda}$-misfit function implementation based on signal decomposition into positive and negative parts. This result might be related to the low kinematic errors associated with the initial model. For less control set-up and in the case of additional noise, this result might not be guaranteed.

Results obtained with trace-by-trace $S_{\varepsilon,\lambda}$-misfit function, *i.e.* 1-D unbalanced entropic transport in time, are similar to the L_2 results, Figure 6.27. The peak-amplitude of the velocity anomaly is well retrieved and

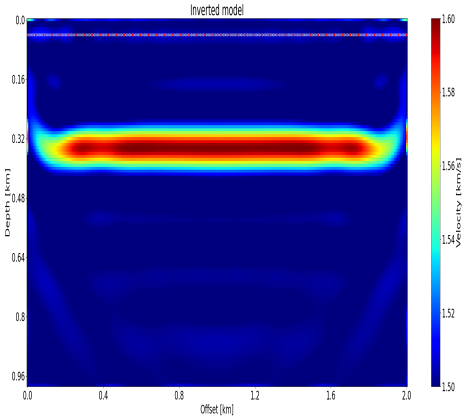
improved compared to the 2-D $S_{\varepsilon,\lambda}$ misfit function based the signal decomposition into positive and negative parts.

Results obtained with the offset-by-offset $S_{\varepsilon,\lambda}$, *i.e.* 1-D unbalanced transport in space, are less satisfactory. In particular the reconstruction of the positive anomaly layer is altered by strong horizontal smoothing artefacts.

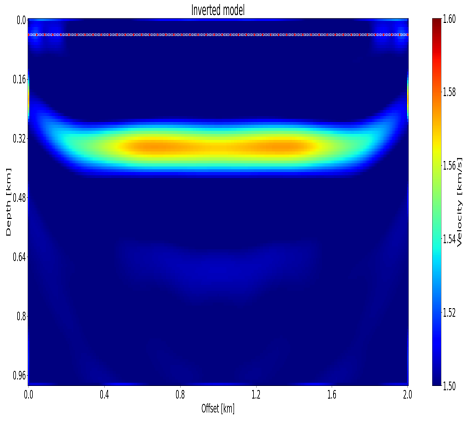
$S_{\varepsilon,\lambda}$ (2D - P_δ)



$S_{\varepsilon,\lambda}$ (2D-Squared Amp.)



$S_{\varepsilon,\lambda}$ (1D-Offset-P_δ)



$S_{\varepsilon,\lambda}$ (1D-Time-P_δ)

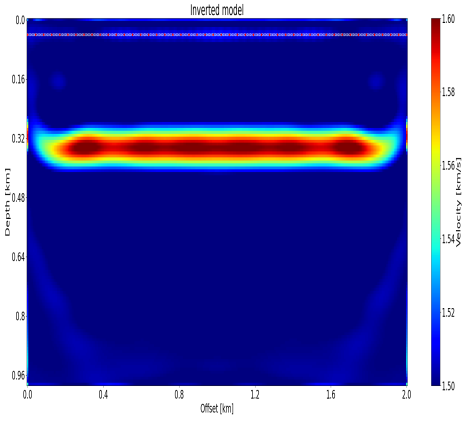
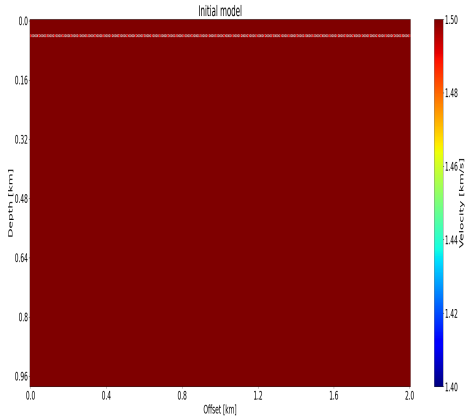


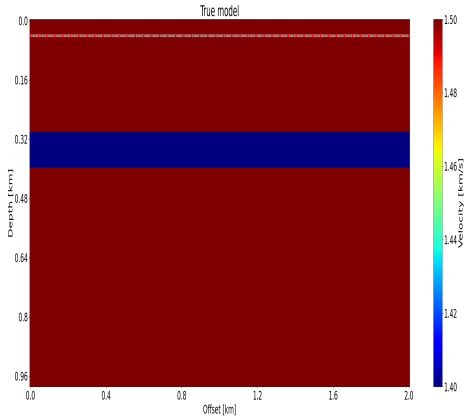
Figure 6.33: Reconstructed models for different implementation strategies of the $S_{\varepsilon,\lambda}$-misfit function: 2-D transport after decomposition of the signals into poistive and negative parts (upper-left figure); Squared-transform of the signals before 2-D transport (upper-left figure); 1-D transport in space (offset-by-offset) after decomposition of the signals into positive and negative parts (lower-left figure); 1-D transport in tile (trace-by-trace) (in time) after decomposition of the signals into positive and negative parts (lower-right figure). The later two implementations make use of the separability of the Gibbs matrix in the Sinkhorn algorithm and of the appropriate scaling of the transport domain.

6.3.2 One-layer reflection model: negative velocity anomaly

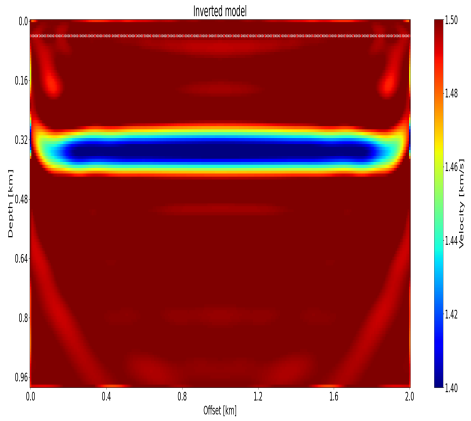
INITIAL



TRUE



L_2-INVERSION



$S_{\varepsilon,\lambda}$-INVERSION

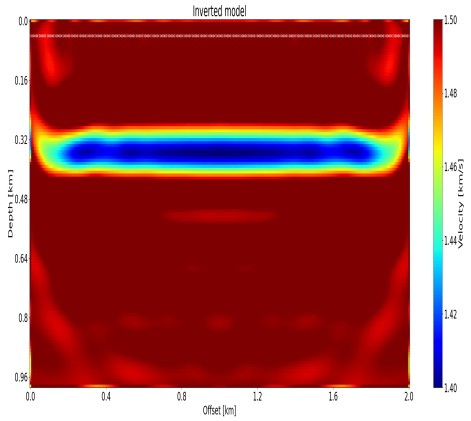
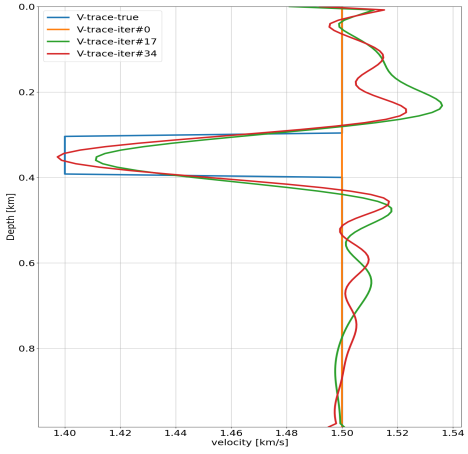


Figure 6.34: Final reconstructed models: initial model (upper-left figure) and targeted model (upper-right figure); FWI results with the L_2- misfit function (bottom-left figure) and the $S_{\varepsilon,\lambda}$-misfit function (bottom-right figure).

L_2-misfit



$S_{\varepsilon,\lambda}$-misfit

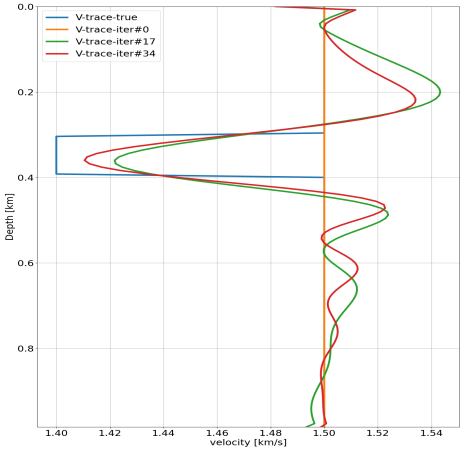


Figure 6.35: Velocity profiles in the physical domain at surface position $x = 1000$ m for different iterations with the L_2 and the $S_{\varepsilon,\lambda}$ based inversion.

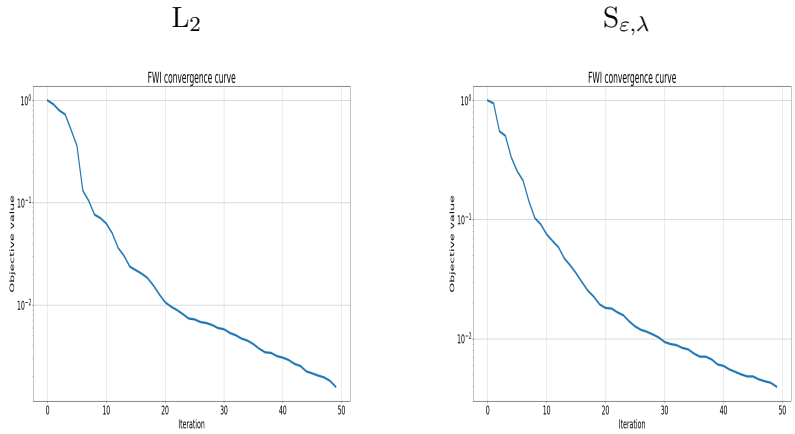


Figure 6.36: Convergence rate with the L_2 (left) and the $S_{\varepsilon,\lambda}$ (right) based inversion.

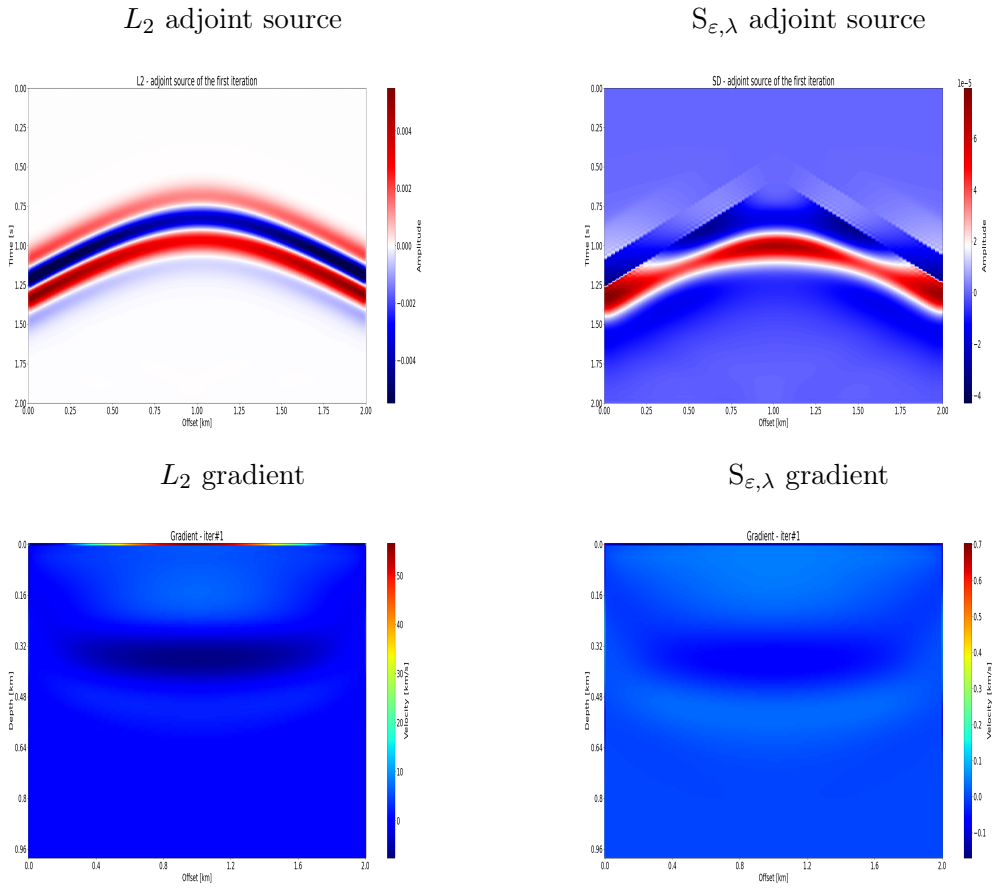


Figure 6.37: Adjoint source and gradient at first iteration for the L_2 and $S_{\varepsilon,\lambda}$ misfit functions

Reconstructed models, after 50 iterations, with the L_2 and $S_{\varepsilon,\lambda}$ based inversion are shown in Figure 6.27. Both inversion provide similar meaningful reconstructions.

For the L_2 based inversion, the recovered peak-amplitude of the velocity anomaly slightly exceeds the "true" one, see Figure 6.35, whereas for the $S_{\varepsilon,\lambda}$ based inversion, the peak-amplitude is only partially recovered after 50 iterations.

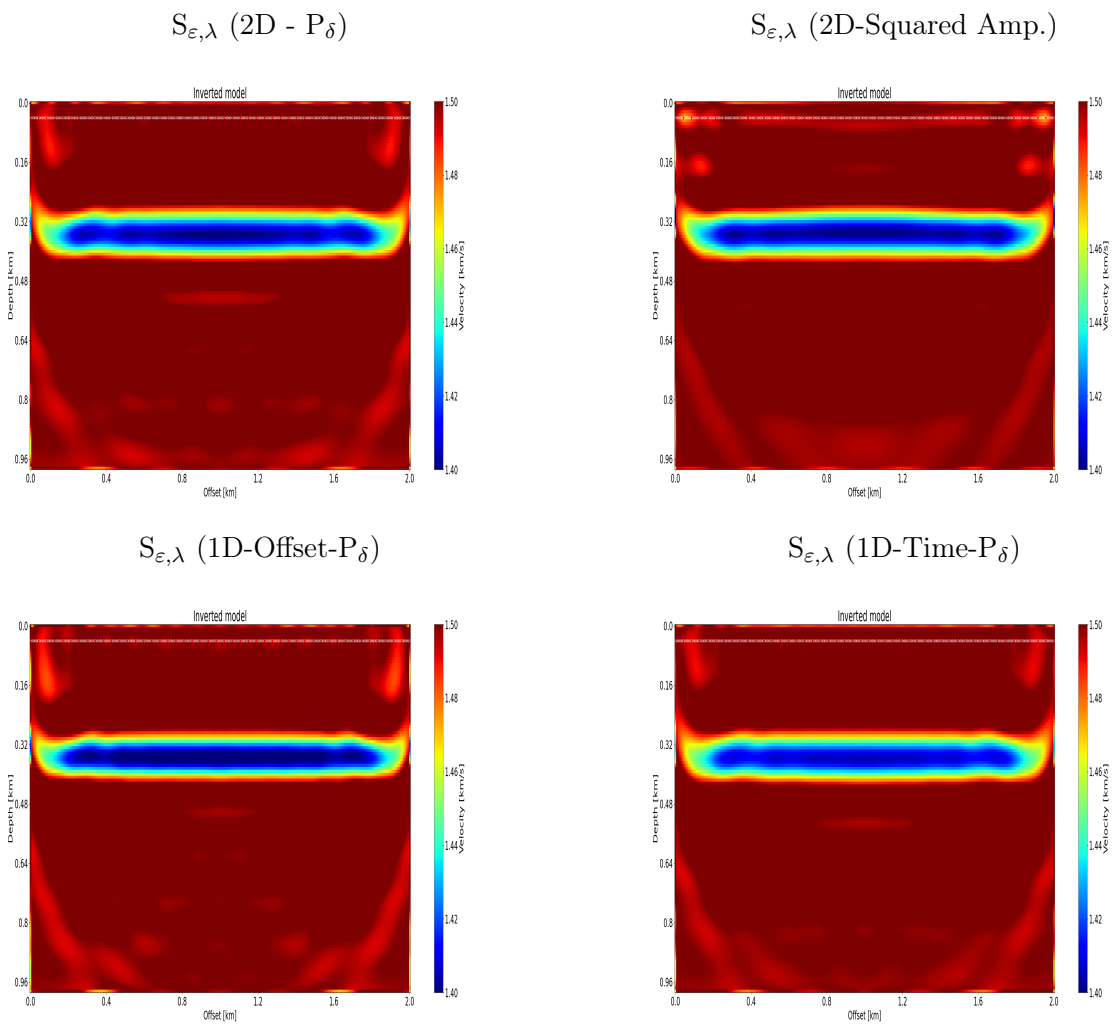


Figure 6.38: Reconstructed models for different implementation strategies of the $S_{\varepsilon,\lambda}$-misfit function: 2-D transport after decomposition of the signals into poistive and negative parts (upper-left figure); Squared-transform of the signals before 2-D transport (upper-left figure); 1-D transport in space (offset-by-offset) after decomposition of the signals into positive and negative parts (lower-left figure); 1-D transport in tile (trace-by-trace) (in time) after decomposition of the signals into positive and negative parts (lower-right figure). The later two implementations make use of the separability of the Gibbs matrix in the Sinkhorn algorithm and of the appropriate scaling of the transport domain.

The L_2 and $S_{\varepsilon,\lambda}$ adjoint sources and optimisation gradients, at the first iteration, are shown in Figure 6.37. The gradient associated to $S_{\varepsilon,\lambda}$ appears smoother compared to the L_2 one, even though both capture similar features.

The convergence rate of the $S_{\varepsilon,\lambda}$ based inversion is similar to the L_2 one, and the misfit distance keep continuing to decrease after 50 iterations see Figure 6.36.

Inversion results for different implementations of the $S_{\epsilon,\lambda}$-misfit function are shown in Figure 6.38 and Figure 6.39.

The 2-D $S_{\varepsilon,\lambda}$ misfit function based on squared-amplitude signal transformation, leads to a meaningful reconstruction similar to the L_2 results in Figure 6.34. The peak-amplitude of the velocity anomaly is only partially reconstructed after 50 iterations, and similar to the one obtained with the 2-D $S_{\varepsilon,\lambda}$-misfit function based on signal decomposition into positive and negative parts, see Fig 6.39.

The results obtained with the trace-by-trace $S_{\varepsilon,\lambda}$-misfit function, *i.e.* 1-D unbalanced transport in time, are very similar to the L_2 results, Figure 6.27. The peak-amplitude of the velocity anomaly is well retrieved

and improved compared to the 2-D $S_{\varepsilon,\lambda}$-misfit function, together with higher resolution of the lower interface of the interface.

The results obtained with the offset-by-offset $S_{\varepsilon,\lambda}$-misfit function, *i.e.* 1-D unbalanced transport in space, are more meaningfull than in the case of a positive velocity anomaly, see section 6.3.2.

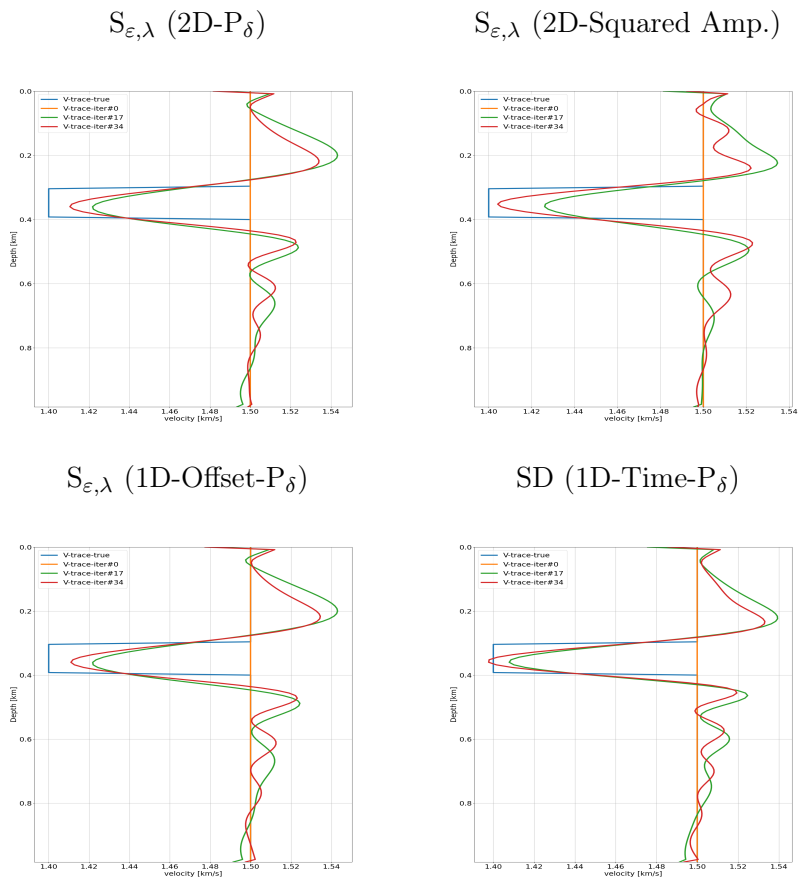


Figure 6.39: Velocity profile along a vertical section of the physical domain for different FWI iterations and different implementations of the $S_{\varepsilon,\lambda}$-misfit function.

6.3.3 Three-layer reflection model

Model configuration

The three-layer model is inspired from (*Yang*, 2018).

The physical model is defined on a rectangular domain of 15×6 km dimension, with three homogeneous layers: a first layer $[0, 1]$ km, with a homogeneous wave velocity set to 1 km/s; a second layer $[1, 2]$ km with a homogeneous wave velocity set to 2 km/s; and a third layer $[2, \infty]$ km with a homogeneous wave velocity set to 4 km/s.

The surface acquisition is composed of 48 evenly-spaced seismic sources and 300 evenly-spaced receivers, both at 100m depth below the upper interface. The sources are point sources and the source-time function is a Ricker wavelet with central frequency $f_c = 15$Hz, and the recording time is 4 s.

The computational domain includes additional perfectly matched layers along each boundary of the physical domain. The constant-density acoustic wave equation is solved in space on a regular grid with 300×120 points using a sixth-oder finite difference scheme and a second-order leap-frog integration scheme

in time with constant time step satisfying the CFL condition as implemented in PySIT.

The inversion problem is solved using the l-BFGS gradient-based method, as implemented in PySIT, with a fixed number of iterations set to 500 iterations. The initial model is a model where the deepest third layer is unknown, see Figure 6.40.

Both the predicted $\mathbf{d}_{\text{cal}}$ and the observed $\mathbf{d}_{\text{obs}}$ data, *i.e.* the latter corresponding to the "true" model, are physical realisations obtained with the same wave propagation solver, and without considering additional noise to the data, *i.e.* the inversion crime setting.

Unless otherwise specified, the $\mathrm{S}_{\varepsilon,\lambda}$-misfit function is formulated as in Chapter 5. The observed and predicted signals are decomposed into positive and negative parts, with the P_{δ} - transform defined in (5.4), and transported separately using the Sinkhorn divergence formulation for unbalanced optimal transport.

In this model configuration, there is no back-scattered information from the interior of the third layer returning to the receivers.

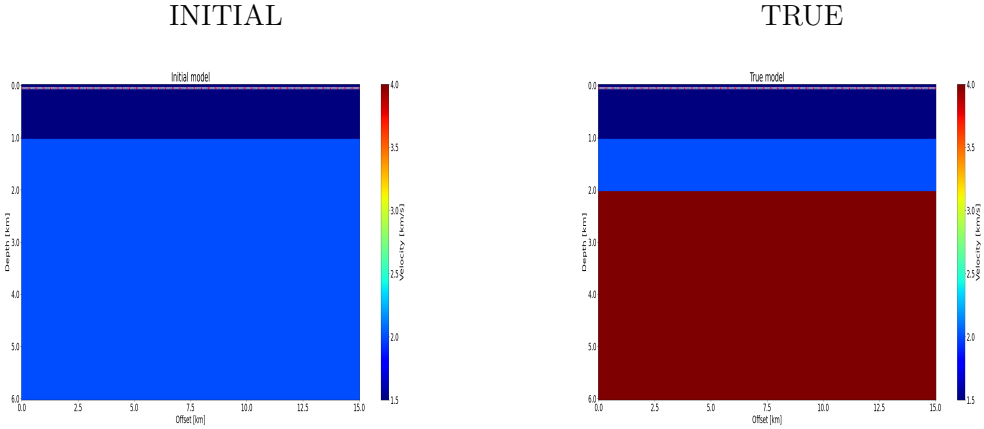


Figure 6.40: Three-layer model: initial (upper-left figure) and true (upper-right figure) models.

Results

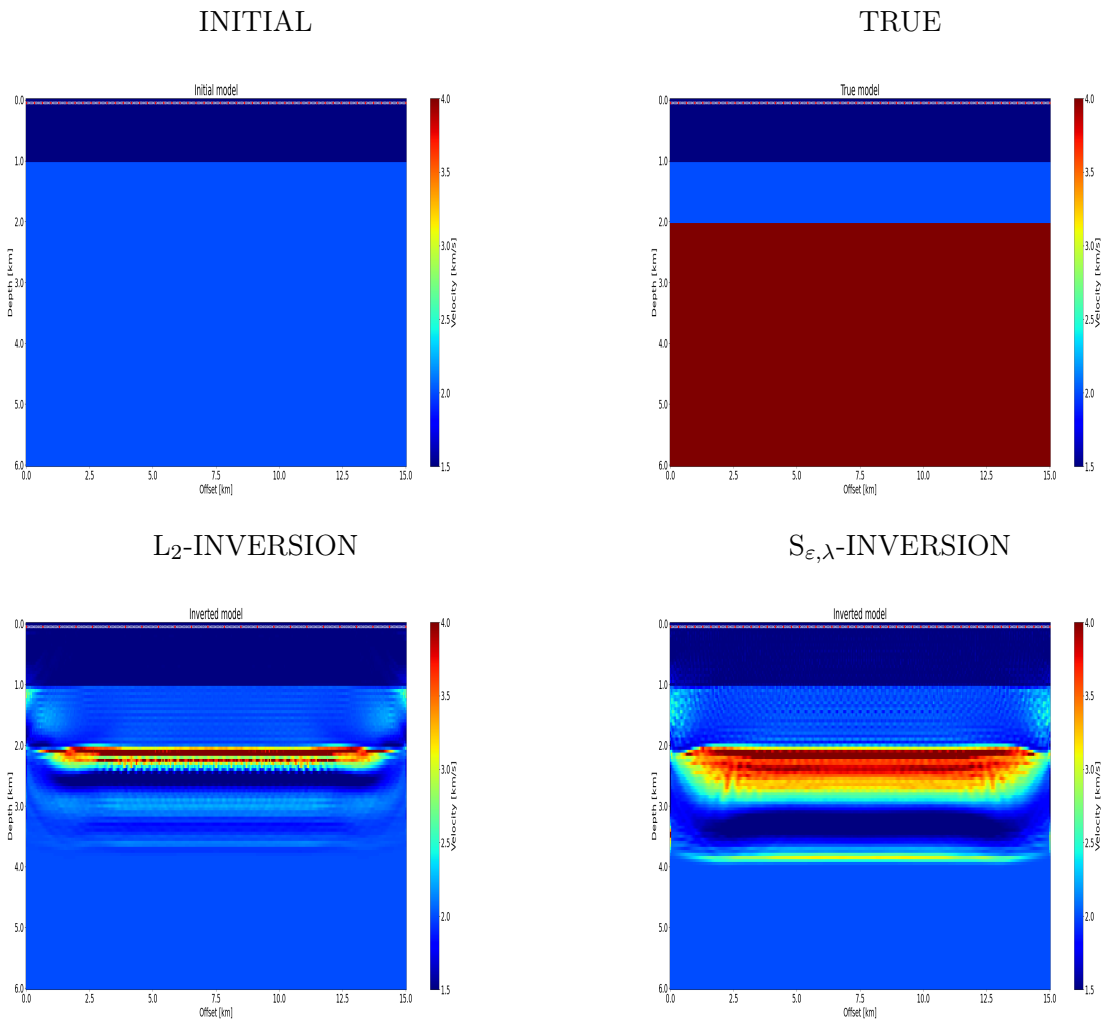


Figure 6.41: Final reconstructed models: initial model (upper-left figure) and targeted model (upper-right figure); FWI results with the L_2- misfit function (bottom-left figure) and the $S_{\varepsilon,\lambda}$-misfit function (bottom-right figure).

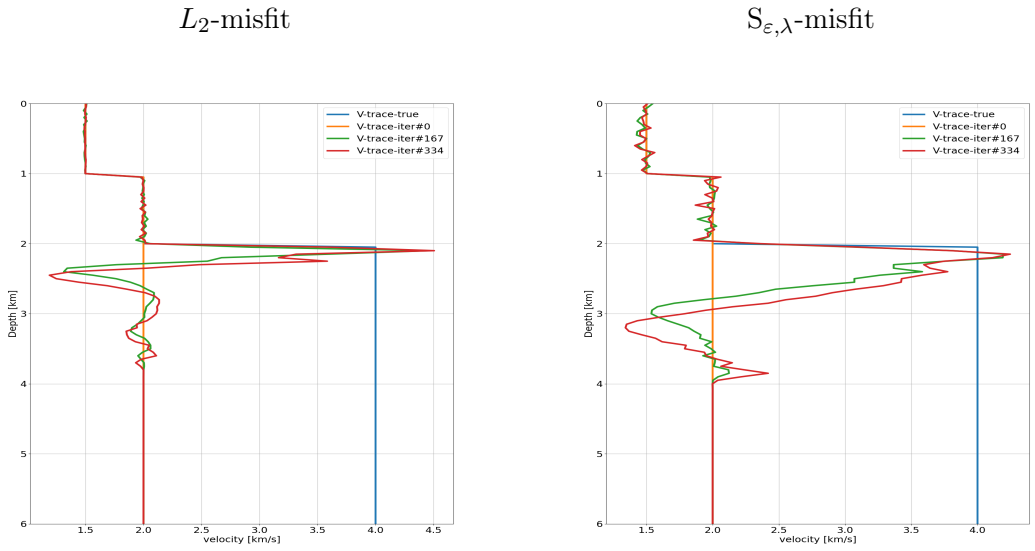


Figure 6.42: Velocity profiles in the physical domain at surface position $x = 7500$ m for different iterations of the L_2 and the $S_{\varepsilon,\lambda}$ based inversion

Reconstructed models with the L_2 and $S_{\varepsilon,\lambda}$ based inversion are shown in Figure 6.41. Both the L_2 and $S_{\varepsilon,\lambda}$ based inversion provide meaningful reconstructions. The L_2 based inversion does not provide however much information beside the local velocity change at the upper third-layer interface, whereas the $S_{\varepsilon,\lambda}$ based inversion recovers part of the velocity information below the upper third-layer interface.

Both inversions converge slowly, see Figure 6.43, with a relatively better convergence rate, after 500 iterations for $S_{\varepsilon,\lambda}$.

The first arrivals in the observed and predicted data are in phase but smaller-amplitude reflected phases behind are not properly recovered for larger offset, see Figure 6.44.

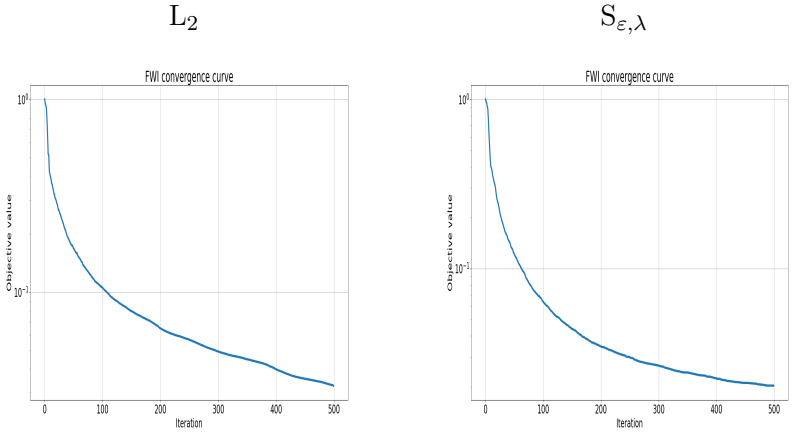


Figure 6.43: Convergence rate for the L_2 (left) and the $S_{\varepsilon,\lambda}$ (right) based inversion

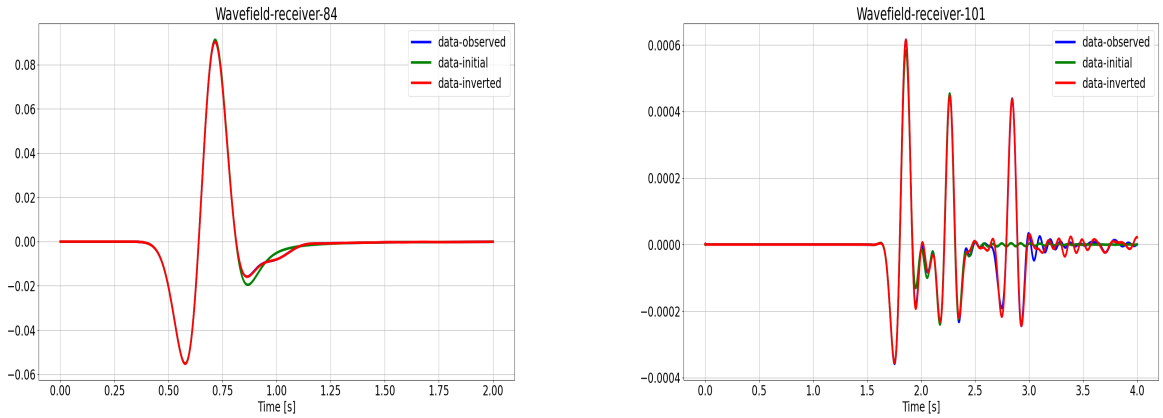


Figure 6.44: Seismic traces at receivers 84 and 101: initial (solid-green line) and predicted data (solid-red line) after 500 FWI iterations with the $S_{\varepsilon,\lambda}$-misfit function; observed data (solid-blue line)

A closer look at the reconstructed velocity models during the $L2$ and $S_{\varepsilon,\lambda}$ based inversion iterations, see Figure 6.42, shows that the $S_{\varepsilon,\lambda}$ based inversion recovers gradually part of the velocity information below the upper third-layer interface. Both the L_2 and $S_{\varepsilon,\lambda}$ based inversion seem to converge toward a local minimum with more artificial deep reflectors, see Figure 6.41, which could be the signature of an ill-posed inversion problem rather than a cycle-skipping problem.

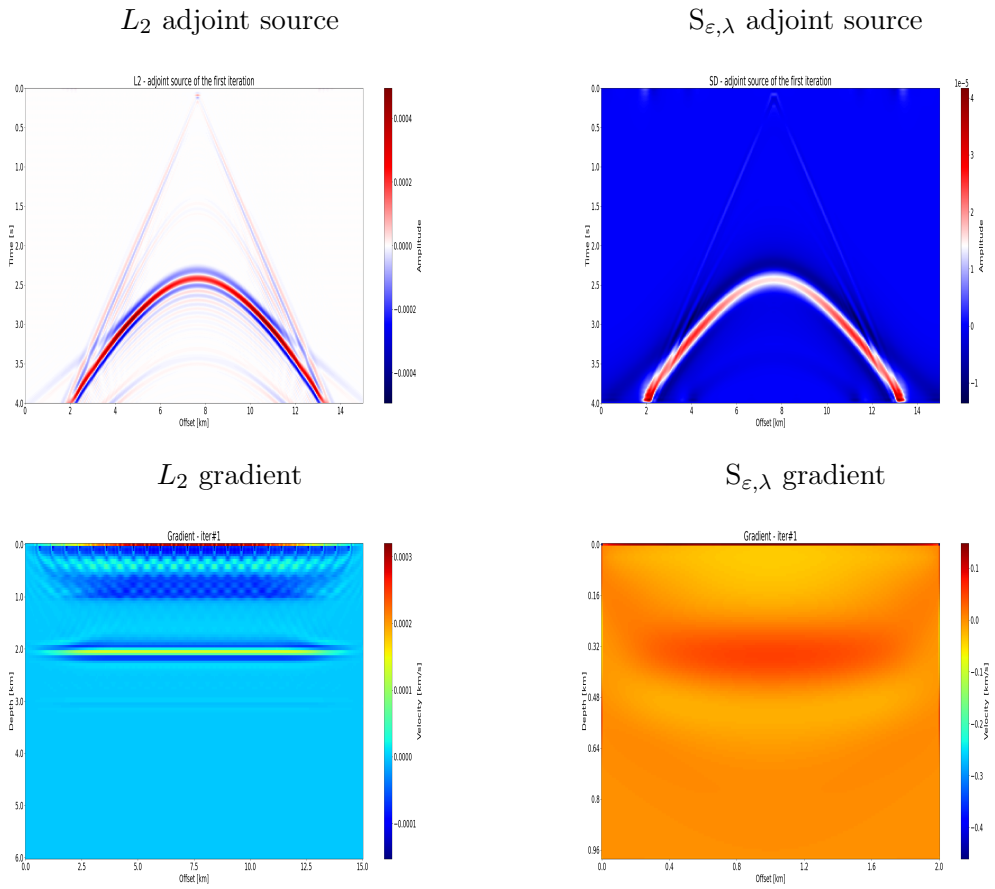


Figure 6.45: Adjoint source and gradient at first iteration for the L_2 and $S_{\varepsilon,\lambda}$ misfit functions.

It is not intuitive that the $S_{\varepsilon,\lambda}$ based inversion can recover velocity information in the third layer while no waves through this region return to the receivers. Velocity information of the third layer is printed in the amplitude of the reflected phases associated to the third-layer interface and in the head waves, the later being more easily retrieved with the 2D $S_{\varepsilon,\lambda}$-misfit function, even for limited offset.

As the $S_{\varepsilon,\lambda}$ based inversion gradually recovers deep velocity information, the reconstructed velocity model in the shallower region exhibits higher-amplitude, zero-mean oscillations around the "true" velocity model, compared to the L_2 based inversion. The amplitude of these oscillations slowly decrease as inversion iterate. Both the $L2$ and $S_{\varepsilon,\lambda}$ based inversion overshoot the peak-amplitude of the velocity contrast at the third-layer interface.

The initial and final predicted data are shown in the time-receiver domain Figure 6.46 for both the L_2 and $S_{\varepsilon,\lambda}$ based inversion. The L_2 and $S_{\varepsilon,\lambda}$ adjoint sources and optimisation gradients, at the first iteration, are shown in Figure 6.45.

The initial predicted data do not contain the reflected phases at the third-layer interface, which has a strong signature in the L_2 and the $S_{\varepsilon,\lambda}$ adjoint sources beside kinematic errors on reflected phases associated to the second-layer interface and the diving waves.

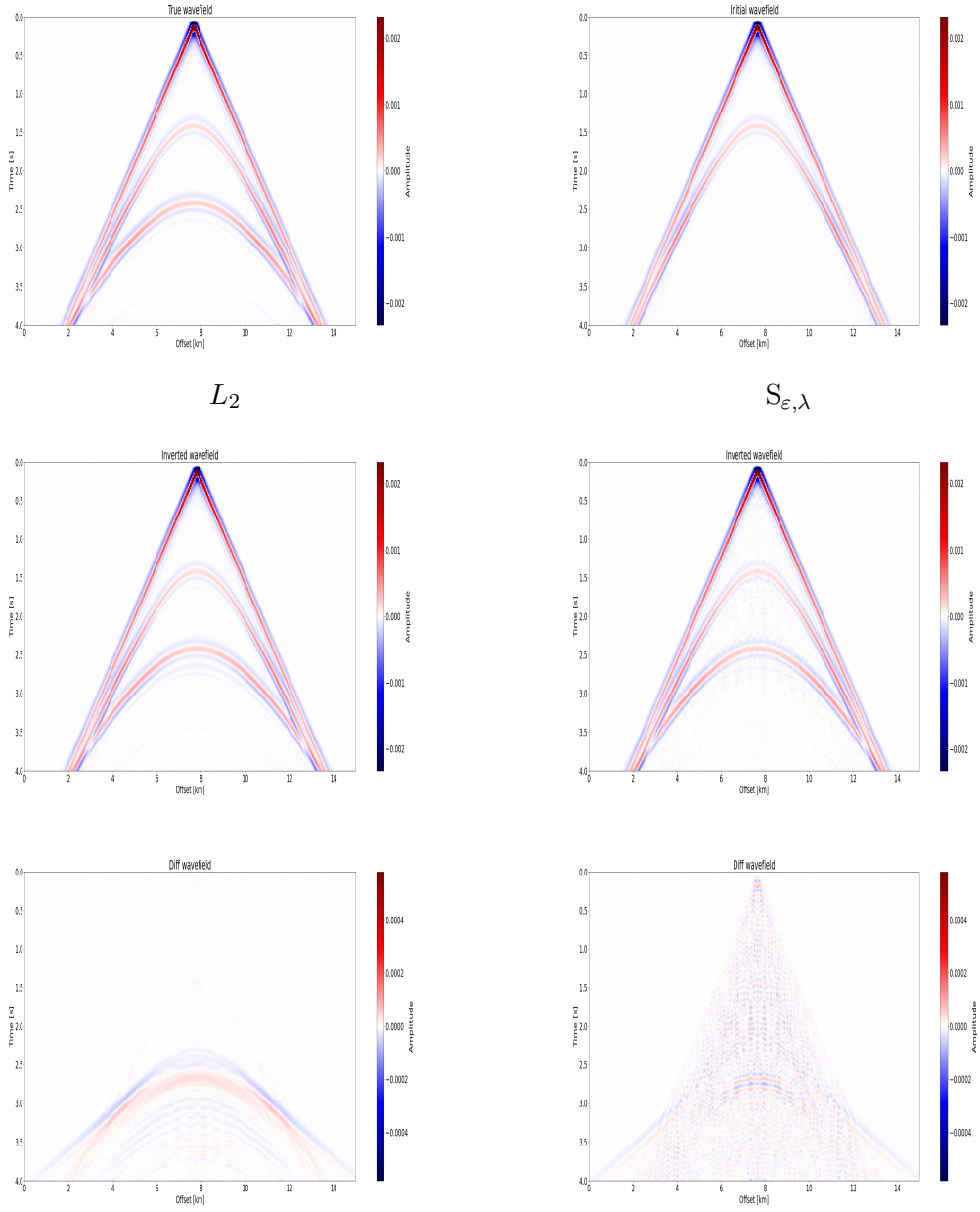


Figure 6.46: Wavefield in the time-receiver domain for a common shot gather associated to the centred source: "true" wavefield (upper-right figure); initial wavefield (upper-left figure); inverted wavefield and difference between the "true" wavefield and inverted wavefield for L_2 (lower-left figures) and $S_{\varepsilon,\lambda}$ (lower-right figures).

The $S_{\varepsilon,\lambda}$ contains additional signature of a mass-spreading over the whole domain. This can be associated to the penalisation of the marginals in the Sinkhorn divergence formulation of unbalanced entropic optimal transport, and more specifically of the local gradient of the Kullback-Leibler. The level of mass is linked to how mass-unbalanced are the predicted and observed data, and as such to the initial model. For strongly unbalanced observed and predicted data the transport cost relatively decreases. The gradient in the $S_{\varepsilon,\lambda}$ based inversion is smoother at the third-layer interface, and with with less subsurface artefacts compared to L_2 based inversion.

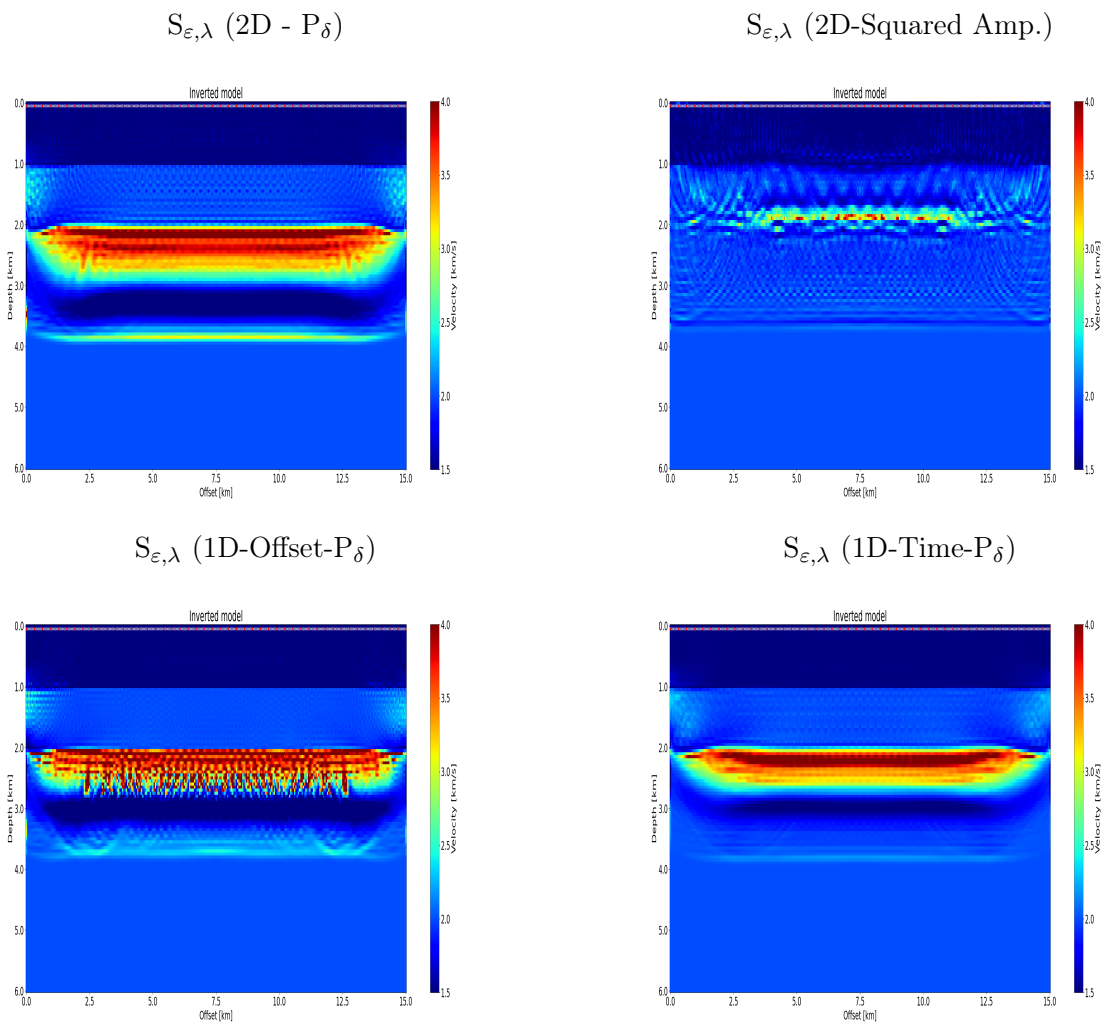


Figure 6.47: Reconstructed models for different implementation strategies of the $S_{\varepsilon,\lambda}$-misfit function: 2-D transport after decomposition of the signals into poistive and negative parts (upper-left figure); Squared-transform of the signals before 2-D transport (upper-left figure); 1-D transport in space (offset-by-offset) after decomposition of the signals into positive and negative parts (lower-left figure); 1-D transport in tile (trace-by-trace) (in time) after decomposition of the signals into positive and negative parts (lower-right figure). The later two implementations make use of the separability of the Gibbs matrix in the Sinkhorn algorithm and of the appropriate scaling of the transport domain.

Inversion results for different implementations of the $S_{\epsilon,\lambda}$-misfit function are shown in Figure 6.47 and Figure 6.48.

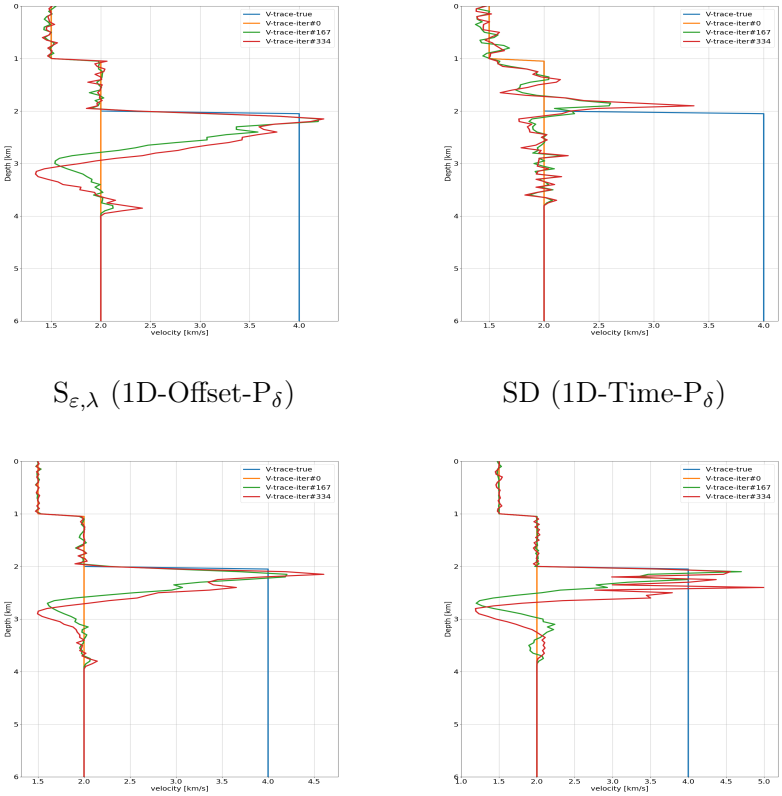
The 2-D $S_{\varepsilon,\lambda}$-misfit function based on squared-amplitude transformation of the signals, without normalisation, leads to unsatisfactory model reconstruction, note the different velocity-amplitude scale in figure 6.47. The velocity contrast at the third-layer interface is not correctly retrieved. This is not actually surprising as reflected waves polarity are lost in the squared-transformation.

Interestingly, results obtained with the offset-by-offset $S_{\varepsilon,\lambda}$-misfit function, *i.e.* 1-D unbalanced transport in space, are meaningful and similar to the 2-D $S_{\varepsilon,\lambda}$ results based on the signal decomposition into positive and negative parts. The inversion gradually recovers also the velocity below the third-layer interface. In this configuration, *i.e.* homogeneous, horizontally-layered medium, 1-D unbalanced transport in space retrieves velocity information associated to head waves and reflection waves at the third-layer interface. The peak-amplitude of the velocity anomaly exceeds slightly the one retrived by the 2D $S_{\varepsilon,\lambda}$ based inversion.

Results obtained with the trace-by-trace $S_{\varepsilon,\lambda}$ misfit function, *i.e.* 1-D unbalanced transport in time, are

also meaningful in this case, and the inversion gradually recovers the velocity below the third-layer interface. The amplitude of the zero-mean oscillations in the subsurface region seems better mitigated than with 2D transport.

$S_{\varepsilon,\lambda}$ (2D-P_δ) $S_{\varepsilon,\lambda}$ (2D-Squared Amp.)



$S_{\varepsilon,\lambda}$ (1D-Offset-P_δ) SD (1D-Time-P_δ)

Figure 6.48: Velocity profile along a vertical section of the physical domain for different FWI iterations and different implementations of the $S_{\varepsilon,\lambda}$-misfit function.

6.4 Marmousi Models

6.4.1 Mini-square Marmousi I model

Model configuration

This model is derived from the synthetic Marmousi I, originally generated at the Institut Français du Petrole (IFP) using a 2D acoustic wave solver (*Versteeg*, 1994).

The model is defined on a 9000 × 3000 m domain and contains many reflectors, steep dips, and strong velocity variations in bot lateral and vertical dimensions, with a minimum velocity of 1500 m/s and a maximum velocity of 5500 m/s.

The physical model here is defined on a mini-square domain *extracted* from the Marmousi I model, with origin $x = 4680$ m and 3000 × 3000 m dimension. The constant density, acoustic wave velocity model is shown in Figure 6.49

The surface acquisition is composed of 48 evenly-spaced seismic sources and 150 evenly spaced receivers, both at 500 m depth below the upper interface. The sources are point sources and the source-time function is a Ricker wavelet with central frequency $f_c = 10$Hz, and the recording time is set to 3 s.

The computational domain includes additional perfectly matched layers along each side of the physical domain. The constant-density acoustic wave equation is solved in space on a regular grid with 150 × 150 points, *i.e.* $dx = dz = 20$ m, using a sixth-order finite difference scheme and in time using a second-order leap-frog integration scheme with constant time step satisfying the CFL condition, as implemented in PySIT.

The inversion problem is solved using the l-BFGS gradient-based method, as implemented in PySIT, with a fixed number of FWI iterations set to 200 iterations.

The initial model, Figure 6.49, is derived from the true velocity model using a smooth low-pass filter with a frequency of 1/300, which preserves part of the long-wavelength components of the true model.

Both the predicted $\mathbf{d}_{\text{cal}}$ and the observed $\mathbf{d}_{\text{obs}}$, *i.e.* the latter corresponding to the "true" model, are physical realisations obtained with the same wave propagation solver, and without considering additional noise to the data, *i.e.* the inversion crime setting.

Unless otherwise specified, the $\text{S}_{\varepsilon,\lambda}$-misfit function is formulated as in Chapter 5. The observed and predicted signals are decomposed into positive and negative parts, using the P_δ-transform defined in (5.4), and transported separately using the Sinkhorn divergence formulation for unbalanced optimal transport.

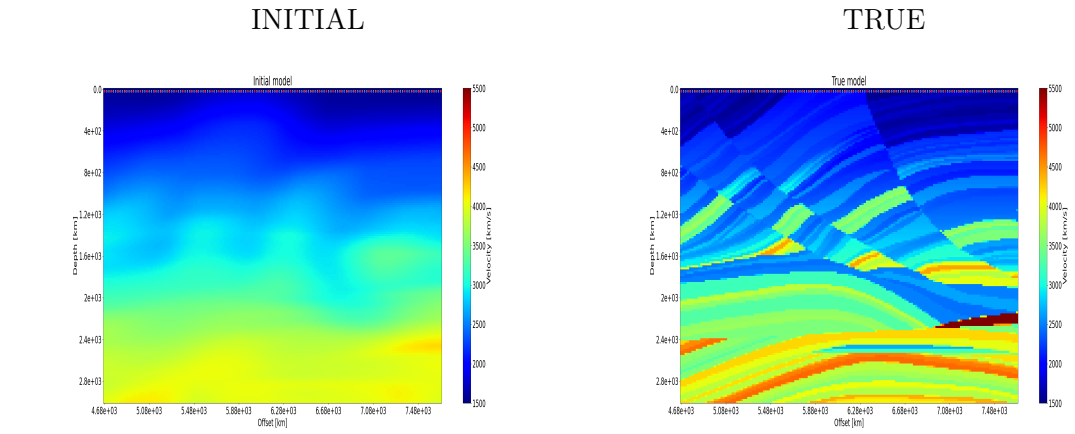


Figure 6.49: Mini-square Marmousi I: initial (upper-left figure) and true (upper-right figure) models.

Results

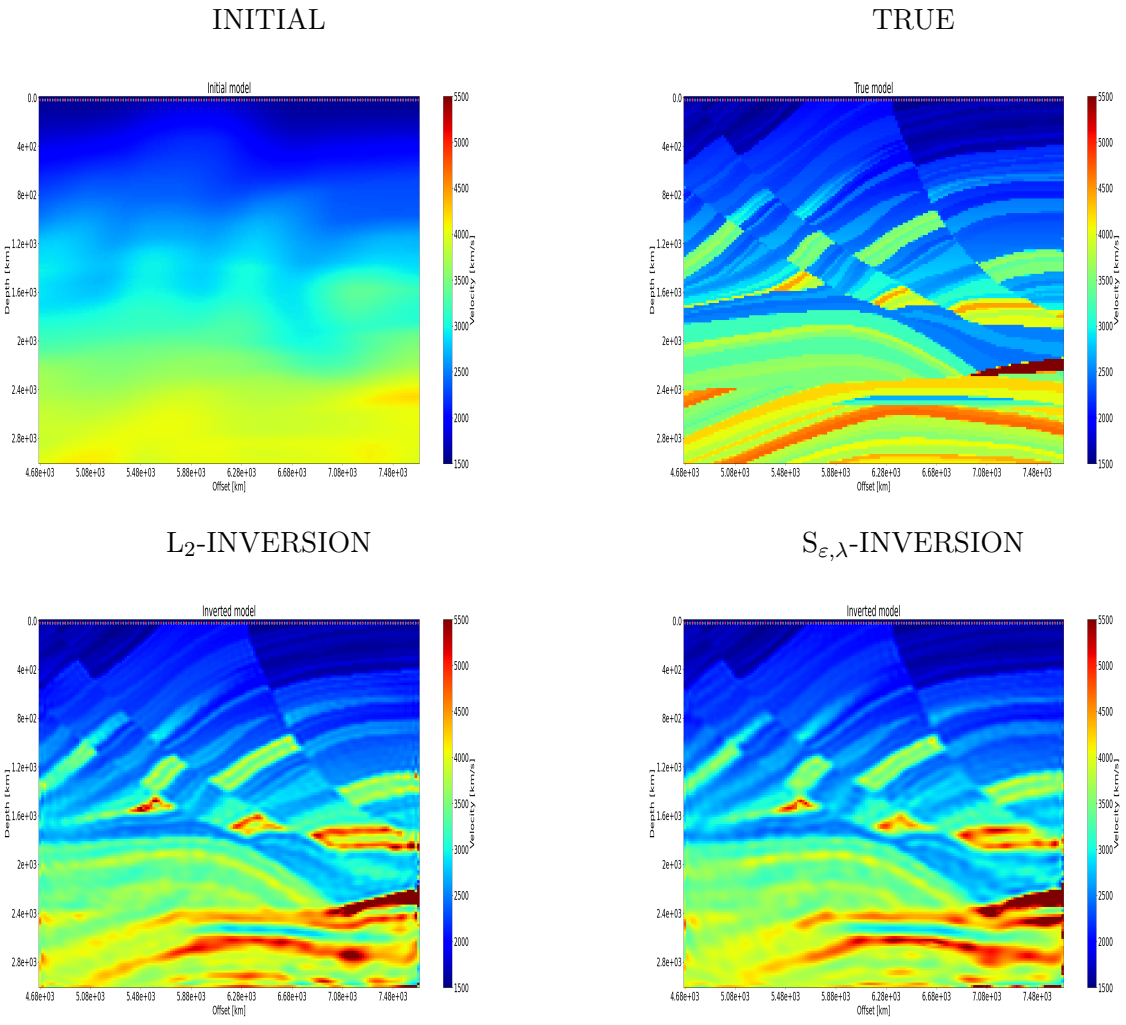


Figure 6.50: Final reconstructed models: initial model (upper-left figure) and targeted model (upper-right figure); FWI results with the L_2- misfit function (bottom-left figure) and the $S_{\varepsilon,\lambda}$-misfit function (bottom-right figure).

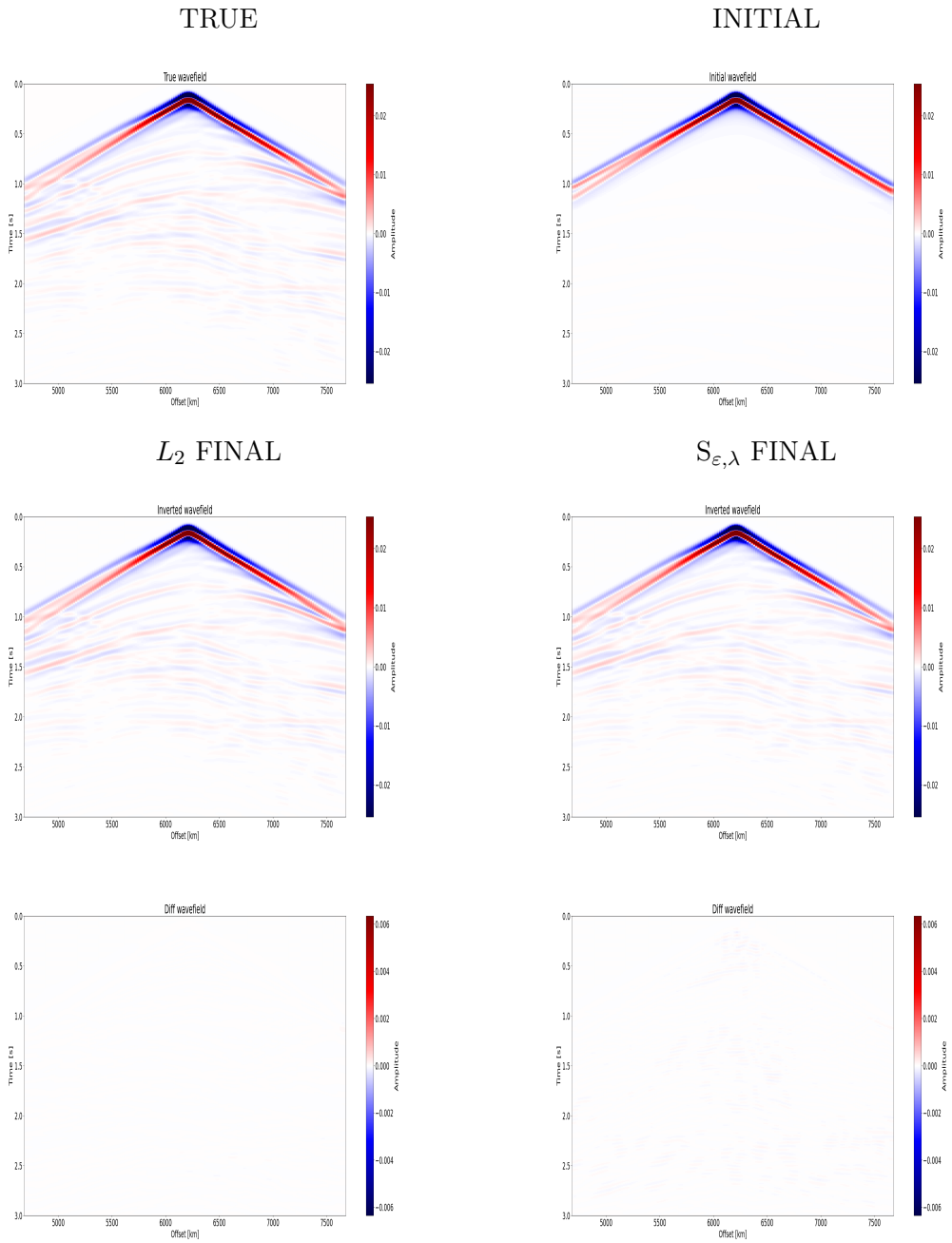


Figure 6.51: Wavefield in the time-receiver domain for a common shot gather associated to the centred source: "true" wavefield (upper-right figure); initial wavefield (upper-left figure); inverted wavefield and difference between the "true" wavefield and inverted wavefield for L_2 (lower-left figures) and $S_{\varepsilon,\lambda}$ (lower-right figures).

Reconstructed models with the L_2 and $S_{\varepsilon,\lambda}$ based inversion are presented in Figure 6.50.

Both the L_2 and $S_{\varepsilon,\lambda}$ based inversion recover similar velocity models in the shallow region (above 1.5 km), with accurately resolved fine layers and normal faults, but have more trouble resolving a high resolution and accurate velocity model in the deep region because reflections from the dipping reflectors and the anticline structure have not been sufficiently recorded due to the limited offset.

Pseudo velocity logs at two surface locations reconstructed with the L_2 and the $S_{\varepsilon,\lambda}$ based inversion are compared in Figure 6.54. Velocity models in both inversion recover reasonably well the true velocity model above 1500 m, whereas quality of the inverted models degrades with depth. Both inversion fail to update at depth (below 2000 m) the low wavenumber structure of the velocity model and place reflectors at wrong

positions.

The initial and final predicted data are shown in the time-receiver domain Figure 6.51 for both the L_2 and $S_{\varepsilon,\lambda}$ based inversion.The L_2 and $S_{\varepsilon,\lambda}$ adjoint sources and optimisation gradients, at the first iteration, are shown in Figure 6.52.

The predicted data at the first iteration miss, as expected, the reflected phases and are only reasonably in phase at short offset for the diving waves. This is reflected into the L_2 adjoint source.

The $S_{\varepsilon,\lambda}$ adjoint source is more complicated with signature of mass-spreading over the whole domain as a result of mass-unbalanced predicted and observed data induced by the smooth initial model.

The gradient in the L_2 and $S_{\varepsilon,\lambda}$ based inversion are different and smoother for $Sl_{\varepsilon,\lambda}$. Signature of the limited offset is visible in both cases.

The convergence rate of the $S_{\varepsilon,\lambda}$ based inversion, over 200 iterations, is relatively faster than the L_2 one, rapidly reducing the misfit distance by one order of magnitude after 20 iterations. The misfit distance keeps continuing slowly decreaseing after 200 iterations.

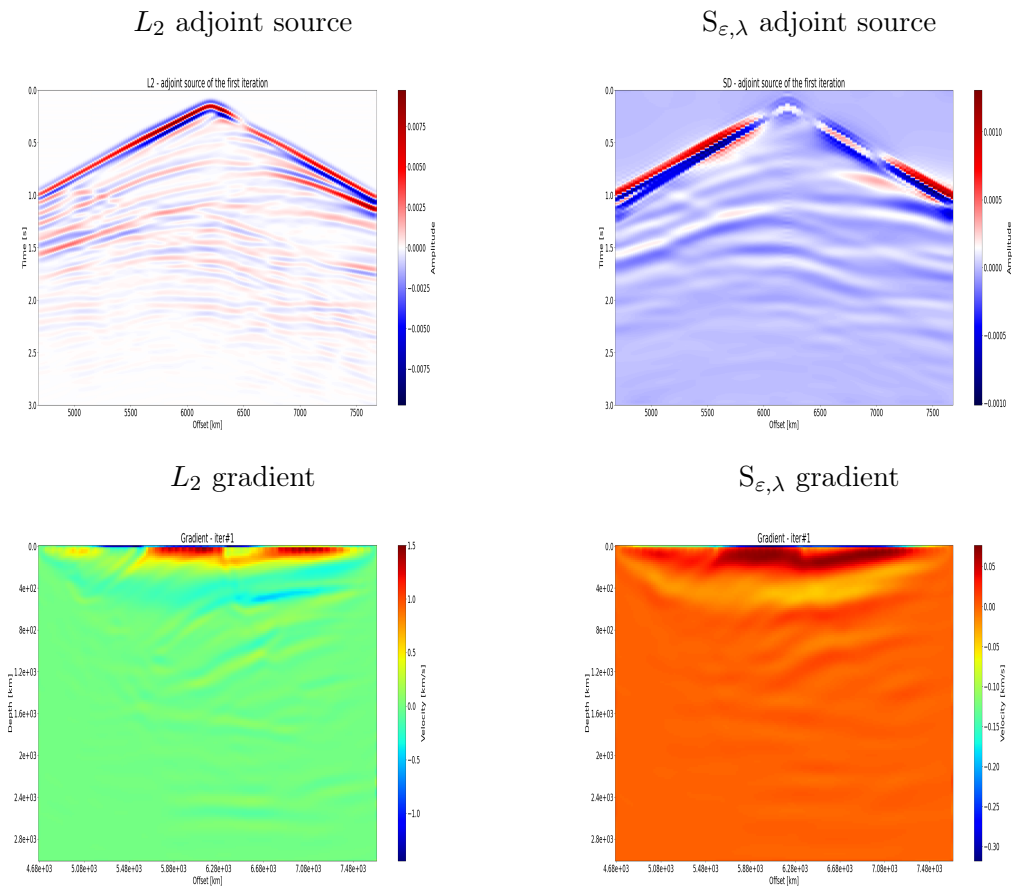


Figure 6.52: Adjoint source and gradient at first iteration for the L_2 and $S_{\varepsilon,\lambda}$ misfit functions

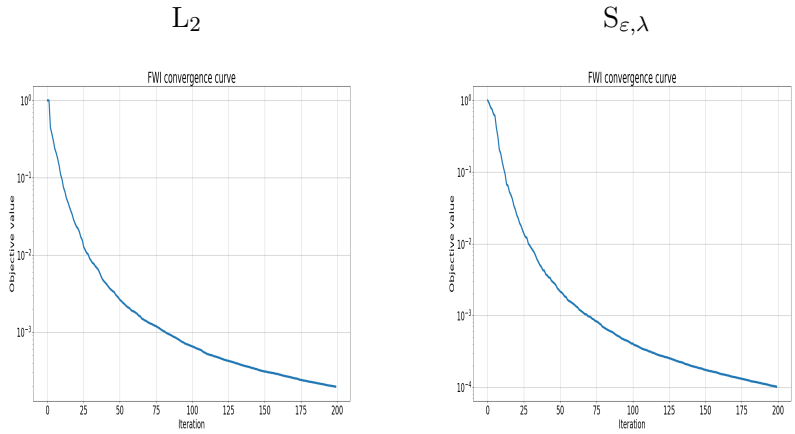


Figure 6.53: FWI convergence curve with the L_2 (left) and the $S_{\varepsilon,\lambda}$ (right) misfit functions.

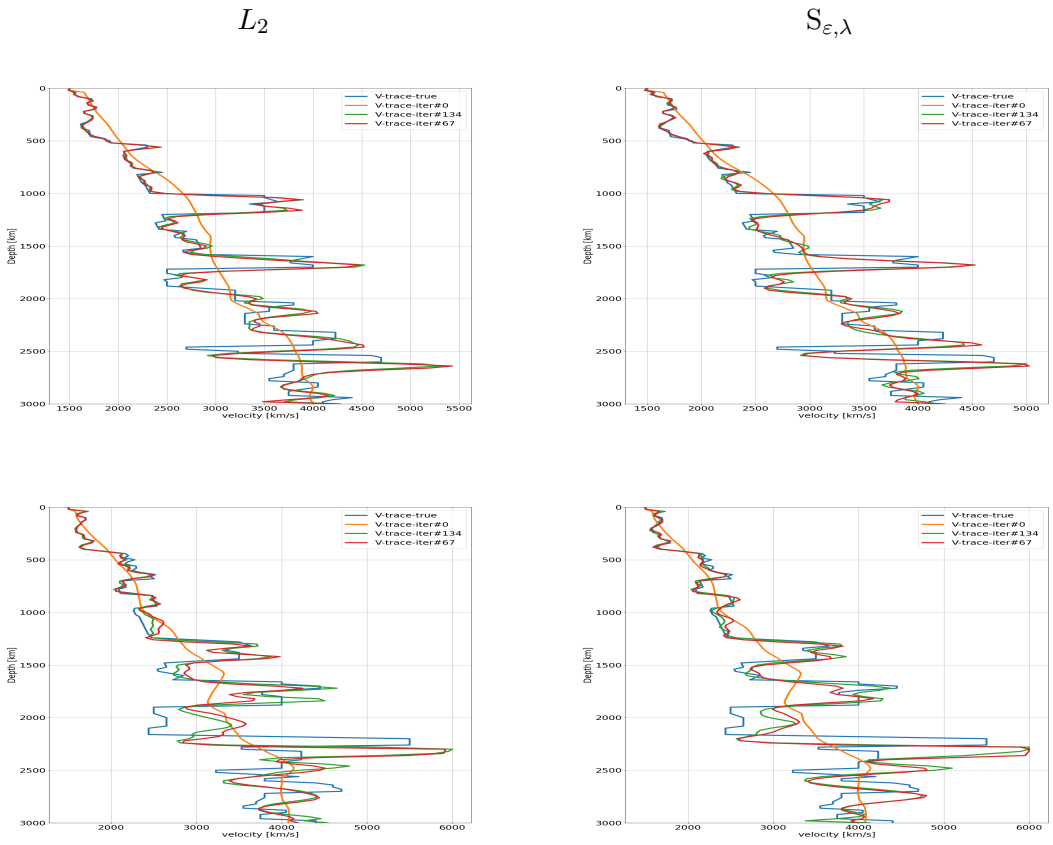


Figure 6.54: Pseudo velocity logs at two surface locations $x = 6480$ m (first row) and $x = 7380$ m (second row) for the L_2 (left) and $S_{\varepsilon,\lambda}$ based inversion

6.4.2 Fracture-zone Marmousi II model

Model configuration

This example is derived from the Marmousi II P-wave velocity model (*Martin et al.*, 2006). The original dimension of the Marmousi II model is 18×3.5 km, with 400 m deep water layer at the top.

The physical model, *i.e.* fracture zone model, is defined on a domain, *extracted* from the Marmousi II model, with origin $x = 8630$ m and 3520×3000 m dimension, with a constant density, acoustic wave velocity model as shown in Figure 6.55

The surface acquisition is composed of 48 evenly-spaced seismic sources and 176 evenly-spaced receivers, both at 500m depth below the upper boundary of the physical domain. The sources are point sources and the source-time function is a Ricker wavelet with central frequency $f_c = 4$Hz, and the recording time is set to 7s.

The computational domain includes additional perfectly matched layers along each side of the physical domain. The constant-density acoustic wave equation is solved in space on a regular grid of 176×150 points, with $dx = dz = 20$m, using a sixth-order finite difference scheme and in time using a second-order leap-frog integration scheme with constant time step satisfying the CFL condition, as implemented in PySIT.

The inversion problem is solved using the l-BFGS gradient-based method, as implemented in PySIT, with a fixed number of FWI iterations set to 250 iterations.

Two initial models are considered: a S-500 initial model, which is derived from the "true" velocity model applying a smooth low-pass filter with a frequency set to 1/500; a 1D-GRAD initial model, which is a homogeneous model with a linearly depth-increasing velocity from 1500 m/s at the bottom of the water layer to 3500 m/s at depth. While the S-500 initial model conserves some of the long-wavelength content of the true model, the 1D-GRAD initial does not contain anymore signature of the long wavelength components of the true model. With both initial models it is a challenging task for FWI to recover large low-wavenumber discrepancies between the initial and "true" model.

Both the predicted $\mathbf{d}_{\text{cal}}$ and the observed $\mathbf{d}_{\text{obs}}$ data, *i.e.* the latter corresponding to the "true" model, are physical realisations obtained with the same wave propagation solver, and without considering additional noise to the data, *i.e.* the inversion crime setting.

Unless otherwise specified, the $\text{S}_{\varepsilon,\lambda}$-misfit function is formulated as in Chapter 5. The observed and predicted signals are each decomposed into positive and negative parts, using the P_{δ}-transform defined in (5.4), and transported separately using the Sinkhorn divergence formulation for unbalanced optimal transport.

Results

Reconstructed physical models with the L_2 and $S_{\varepsilon,\lambda}$ based inversion are shown in Figure 6.55, for the two initial models.

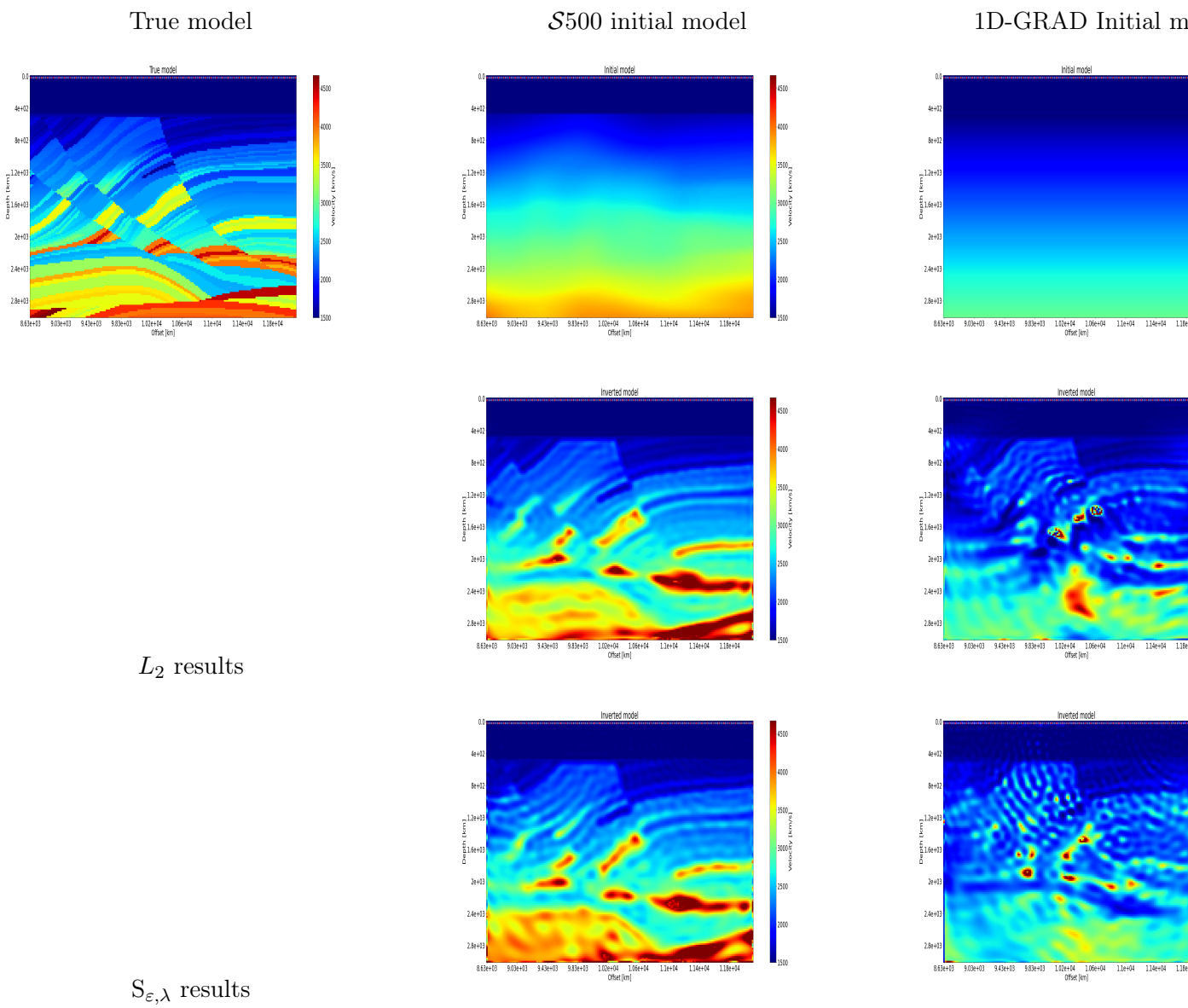


Figure 6.55: Fracture zone Marmousi II model: results for L_2 (second row) and $S_{\varepsilon,\lambda}$ (third row) misfit functions, with the S-500 (second column) and 1D-GRAD (third column) initial models.

S − 500 initial model

Both the L_2 and $S_{\varepsilon,\lambda}$ based inversion reconstruct partially meaningful models. Velocity models in both cases are similar with low-resolution images of the fine layers and normal faults above 1500 m. Both inversion have trouble resolving an accurate velocity model in the deep region (below 1500 m) because of the large low wave-number discrepancies between the S-500 initial model and the "true" model, and because reflections and the anticline structure in the deep region are not sufficiently recorded due to the limited offset.

Pseudo velocity logs, at two surface locations, reconstructed with the L_2 and the $S_{\varepsilon,\lambda}$ based inversion are compared in Figure 6.56.

Velocity models in both inversion recover reasonably well the true velocity model above 1500 m, whereas quality of the inverted models degrades rapidly with depth. Both inversion fail to update at depth (below 1500 m) the low wavenumber structure of the velocity model and place reflectors at wrong positions.

The initial and final predicted data are shown in the time-receiver domain Figure 6.57 for both the L_2 and $S_{\varepsilon,\lambda}$ based inversion. The adjoint sources and optimisation gradient, at the first iteration, are shown in

Figure 6.58.

The observed data and the predicted data at the last iteration for both the L_2 and $S_{\varepsilon,\lambda}$ based inversion are reasonably in phase in the shallow region (above 1500 m).

In the predicted data at the first iteration reflected phases, as expected, are missing whereas diving phases are reasonably in phase only at short offset. This reflects into the L_2 and the $S_{\varepsilon,\lambda}$ adjoint sources. The $S_{\varepsilon,\lambda}$ adjoint source is smooth and capture information of both diving and reflection phases. The source captures also signature of a mass-spreading over the whole domain as a result of the unbalanced-mass between predicted and observed data. At first iteration, the he L_2 and the $S_{\varepsilon,\lambda}$ based inversion gradients look different with different sensitivity regions at depth.

The $S_{\varepsilon,\lambda}$ based inversion convergence rate appears faster than the L_2-based inversion one and the misfit distance keeps continuously decreasing after 250 iterations.

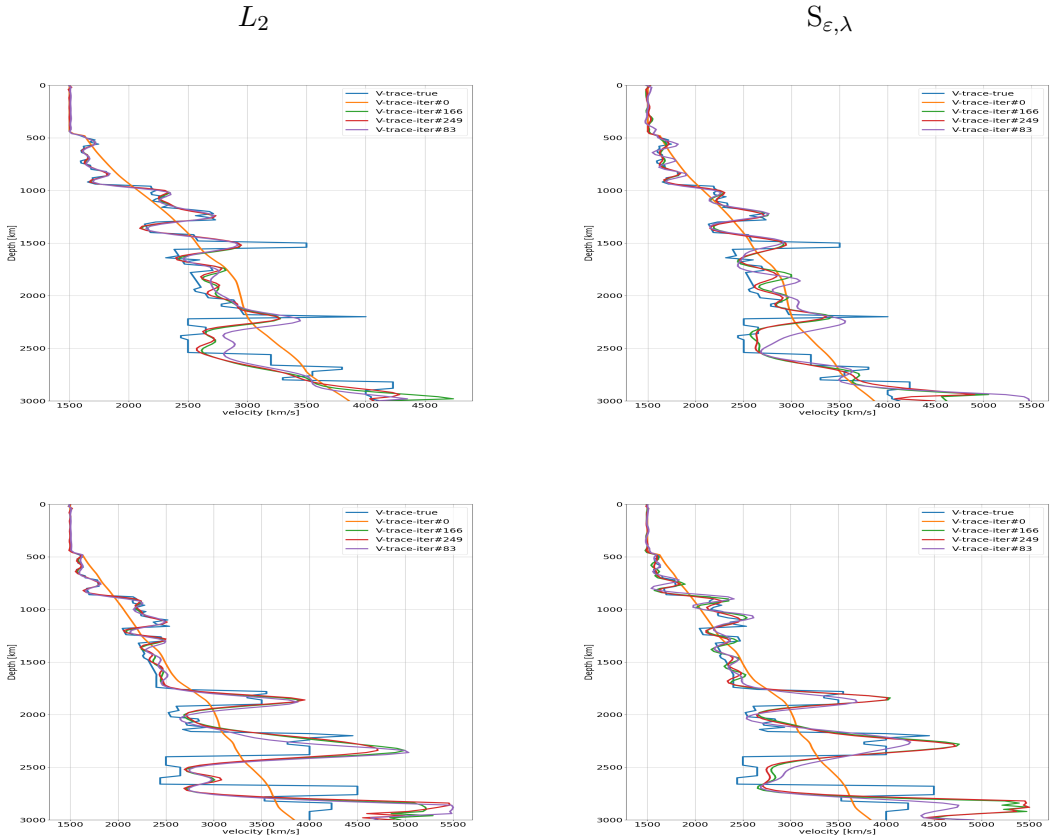


Figure 6.56: S-500 initial model: pseudo velocity logs at two surface locations $x = 10730$ m (first row) and $x = 11430$ m (second row) for the L_2 (left) and $S_{\varepsilon,\lambda}$ based inversion

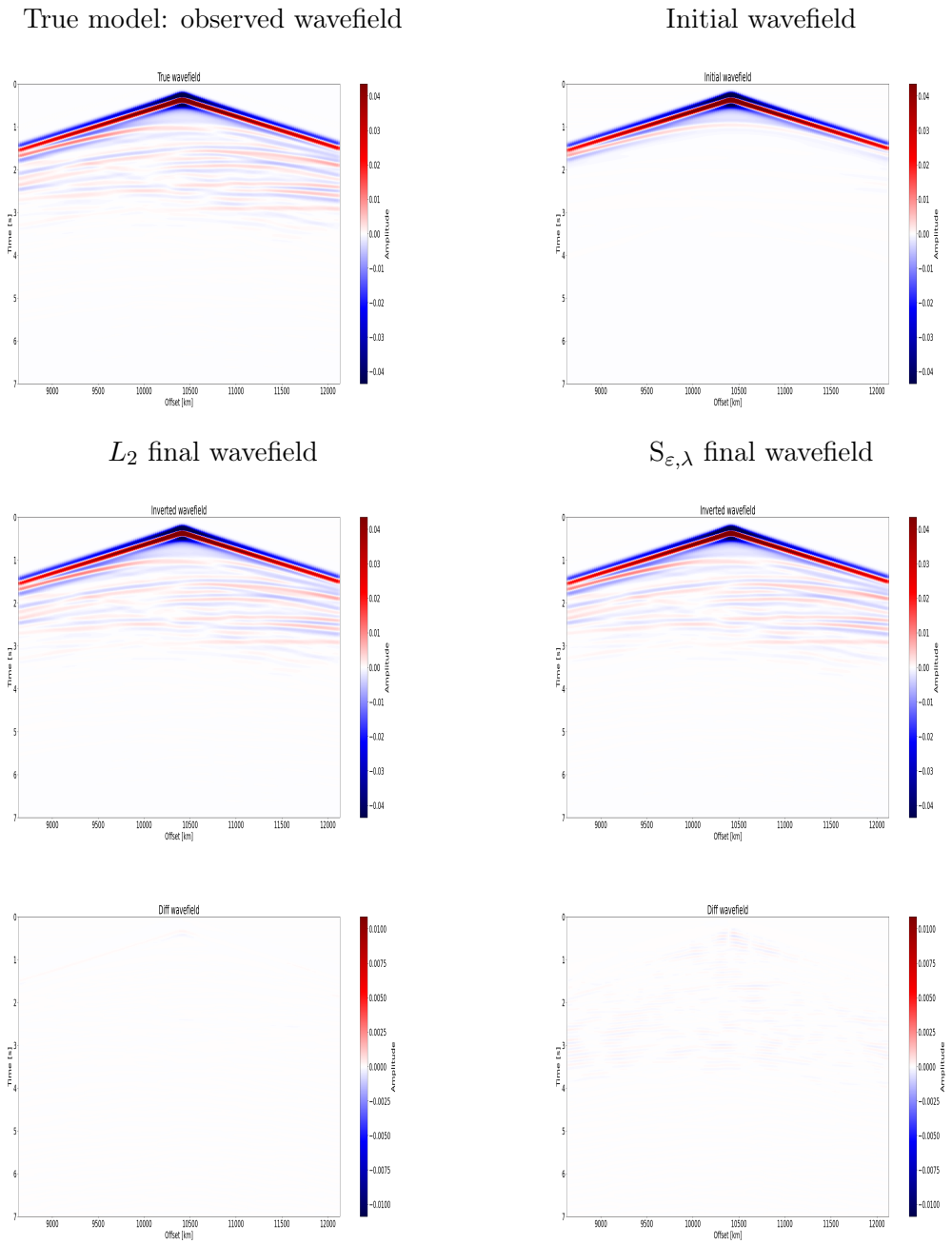


Figure 6.57: S-500 initial model: wavefield in the time-receiver domain for a common shot gather associated to a source et the middle of the domain: observed wavefield (upper-right figure); initial wavefield (upper-left figure); final wavefield for the inverted model and difference between the observed and final wavefield for L_2 (lower-left figures) and $S_{\varepsilon,\lambda}$ (lower-right figures).

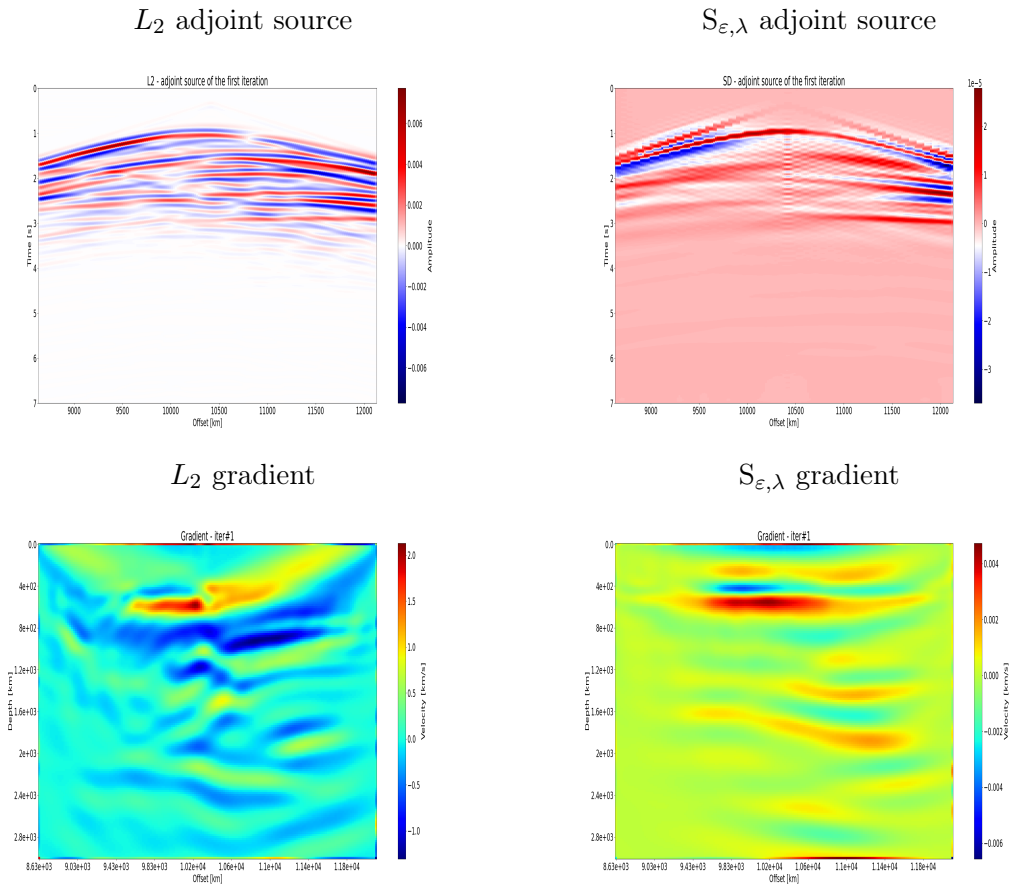


Figure 6.58: S-500 initial model: adjoint source and gradient at first iteration for the L_2 and $S_{\varepsilon,\lambda}$ misfit functions

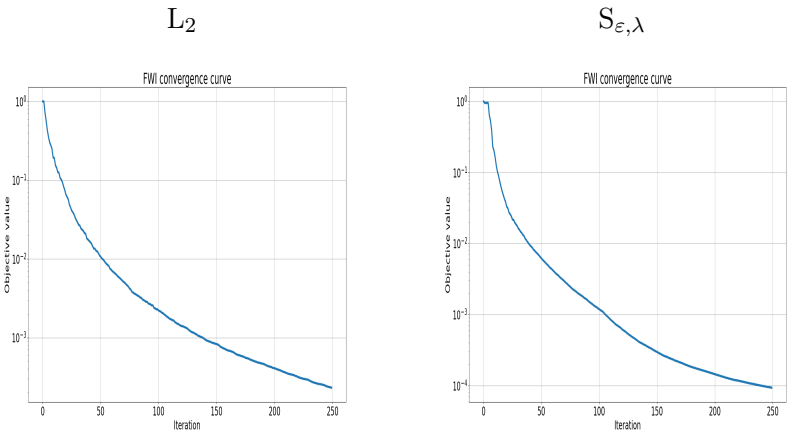


Figure 6.59: *S*-500 initial model: FWI convergence with the L_2 (left) and the $S_{\varepsilon,\lambda}$ (right) misfit functions

Results with the 2-D and the 1-D trace-by-trace $S_{\varepsilon,\lambda}$ misfit functions are shown in Figure 6.60 for comparison.

Results with the trace-by-trace $S_{\varepsilon,\lambda}$-misfit function, *i.e.* 1-D unbalanced transport in time, are not satisfactory compared to the 2-D $S_{\varepsilon,\lambda}$ and the 1-D L_2 misfit functions. The inversion does not recover long-wavelength components and exhibit more artefacts.

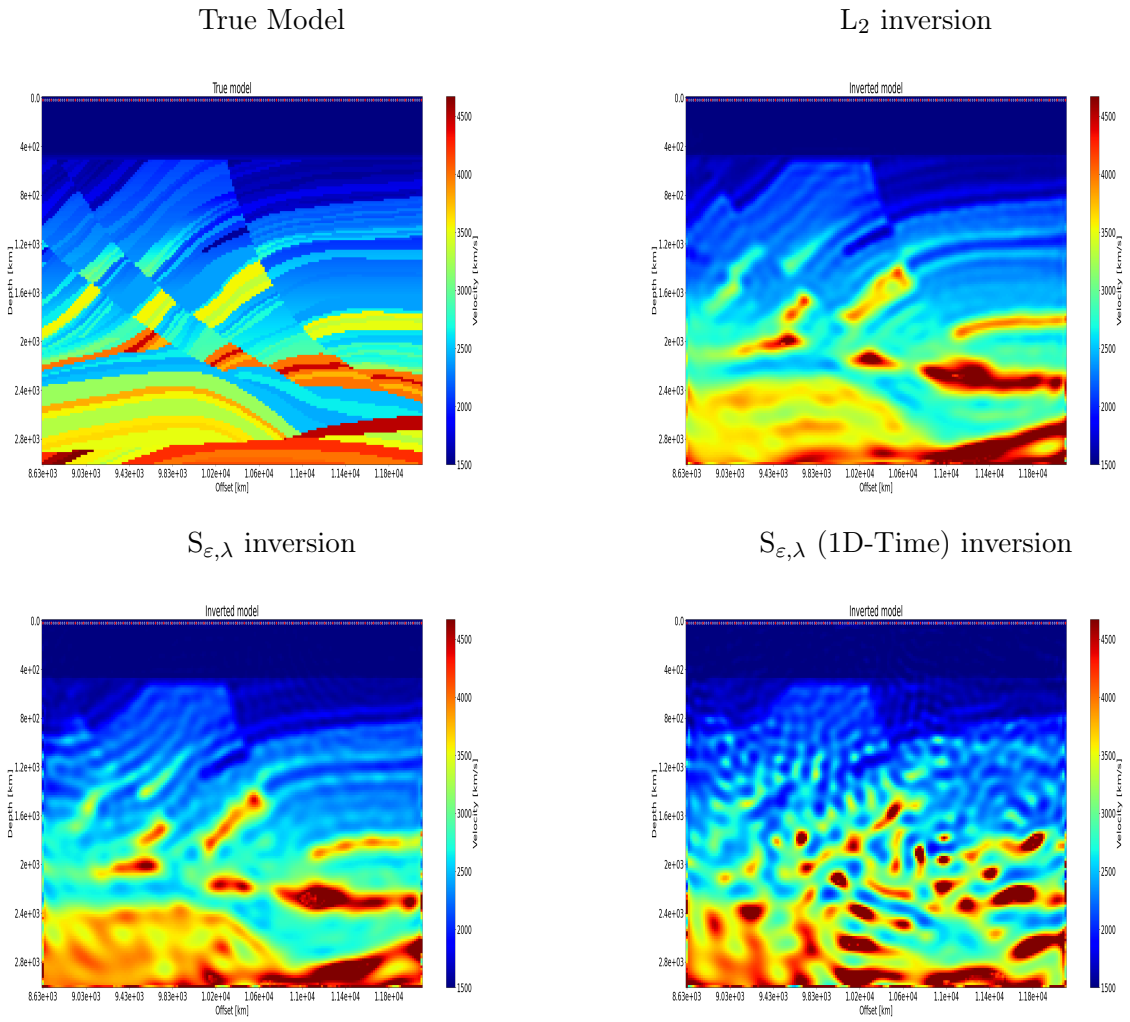


Figure 6.60: Reconstructed models for different implementation strategies of the $S_{\varepsilon,\lambda}$-misfit function: 2-D transport after decomposition of the signals into poistive and negative parts (upper-left figure); Squared-transform of the signals before 2-D transport (upper-left figure); 1-D transport in space (offset-by-offset) after decomposition of the signals into positive and negative parts (lower-left figure); 1-D transport in tile (trace-by-trace) (in time) after decomposition of the signals into positive and negative parts (lower-right figure). The later two implementations make use of the separability of the Gibbs matrix in the Sinkhorn algorithm and of the appropriate scaling of the transport domain.

1D-GRAD initial model

Reconstructed models with the L_2 and $S_{\varepsilon,\lambda}$ based inversion are not satisfactory after 250 iterations., even if some long-wavelength components start emerging in a more satisfactory way for $S_{\varepsilon,\lambda}$. It is an extremely challenging task for FWI to recover large low-wavenumber discrepancies between this initial model and the true model, especially in the deeper region (below 1500 km) where reflections from the dipping reflectors and the anticline structure are not sufficiently recorded due to the limited offset.

Emerging signatures of the fine layers and normal faults in the shallow region (above 1500 km) appear better retrieved with the L_2 based inversion, whereas the $S_{\varepsilon,\lambda}$ based inversion is relatively less prone to spurious artefacts in the deeper region (below 1500 m) than the L_2 based inversion, which may suffer from cycle-skipping associated to deep reflections in the central region.

Pseudo velocity logs, at two surface locations, reconstructed with the L_2 and the $S_{\varepsilon,\lambda}$ based inversion are compared in Figure 6.62. Velocity models in both inversion recover reasonably well the true velocity model above 1000 m, whereas quality of the inverted models degrades rapidly with depth. Both inversion fail to

update at depth below 1000 m the velocity model and place reflectors at wrong positions.

The initial and final predicted data are shown in the time-receiver domain Figure 6.64 for both L_2 and $S_{\varepsilon,\lambda}$, based inversion. The adjoint sources and optimisation gradients, at the first iteration, are shown in Figure 6.63.

The predicted data at last iteration does not fit well the observed data for both the L_2 and $S_{\varepsilon,\lambda}$ based inversion. Long wavelength error components appears for L_2 possibly as a result of cycle-skipping issue.This appears less an issue for $S_{\varepsilon,\lambda}$.

The L_2 and $S_{\varepsilon,\lambda}$ adjoint source are again different and does not capture the same space and time information in the early and late times of the source. The gradient of the $S_{\varepsilon,\lambda}$ exhibit large lobe-sided, long-wavelength updates in the shallow part (above 800 m) and relatively less sensitivity in the deeper region. The gradient of the L_2 based inversion exhibits more complicated and stronger updates in the region between 800-2000 m depth.

The convergence rate, over the first 250 iterations, appears faster with the $S_{\varepsilon,\lambda}$ based inversion compared to the L_2 based inversion, and the misfit keeps continuing decreasing.

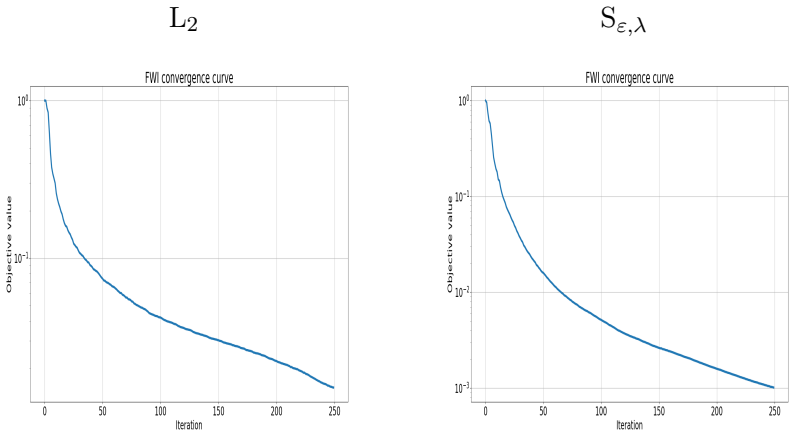


Figure 6.61: 1D-GRAD initial model: FWI convergence with the L_2 (left) and the $S_{\varepsilon,\lambda}$ (right) misfit functions.

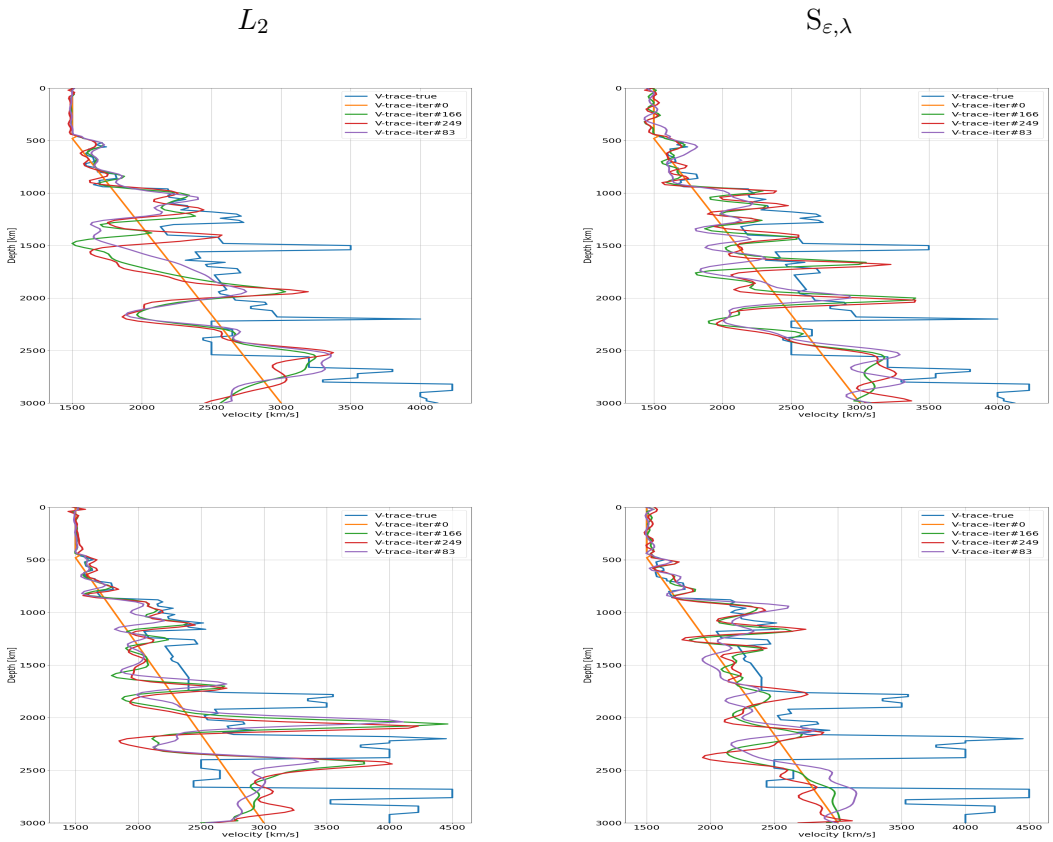


Figure 6.62: 1D-GRAD initial model: pseudo velocity logs at two surface locations $x = 10730$ m (first row) and $x = 11430$ m (second row) for the L_2 (left) and $S_{\varepsilon,\lambda}$ based inversion

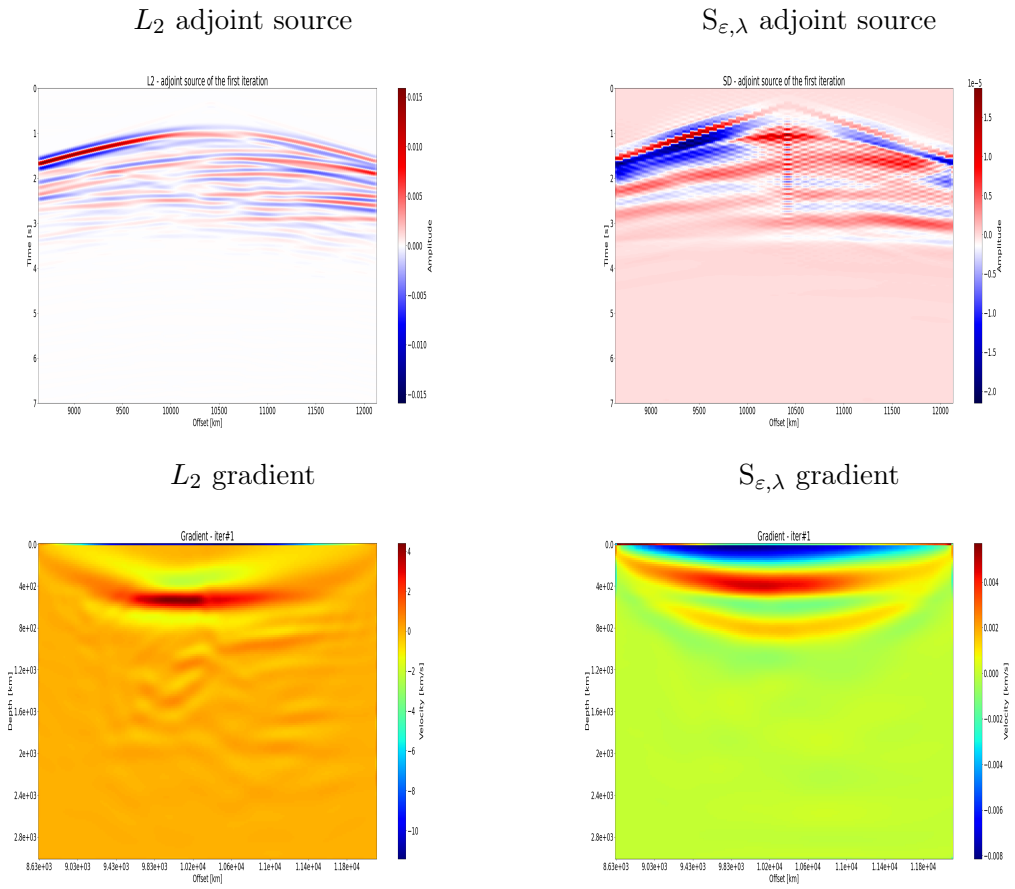


Figure 6.63: 1D-GRAD initial model: adjoint source and gradient at first iteration for the L_2 and $S_{\varepsilon,\lambda}$ misfit functions.

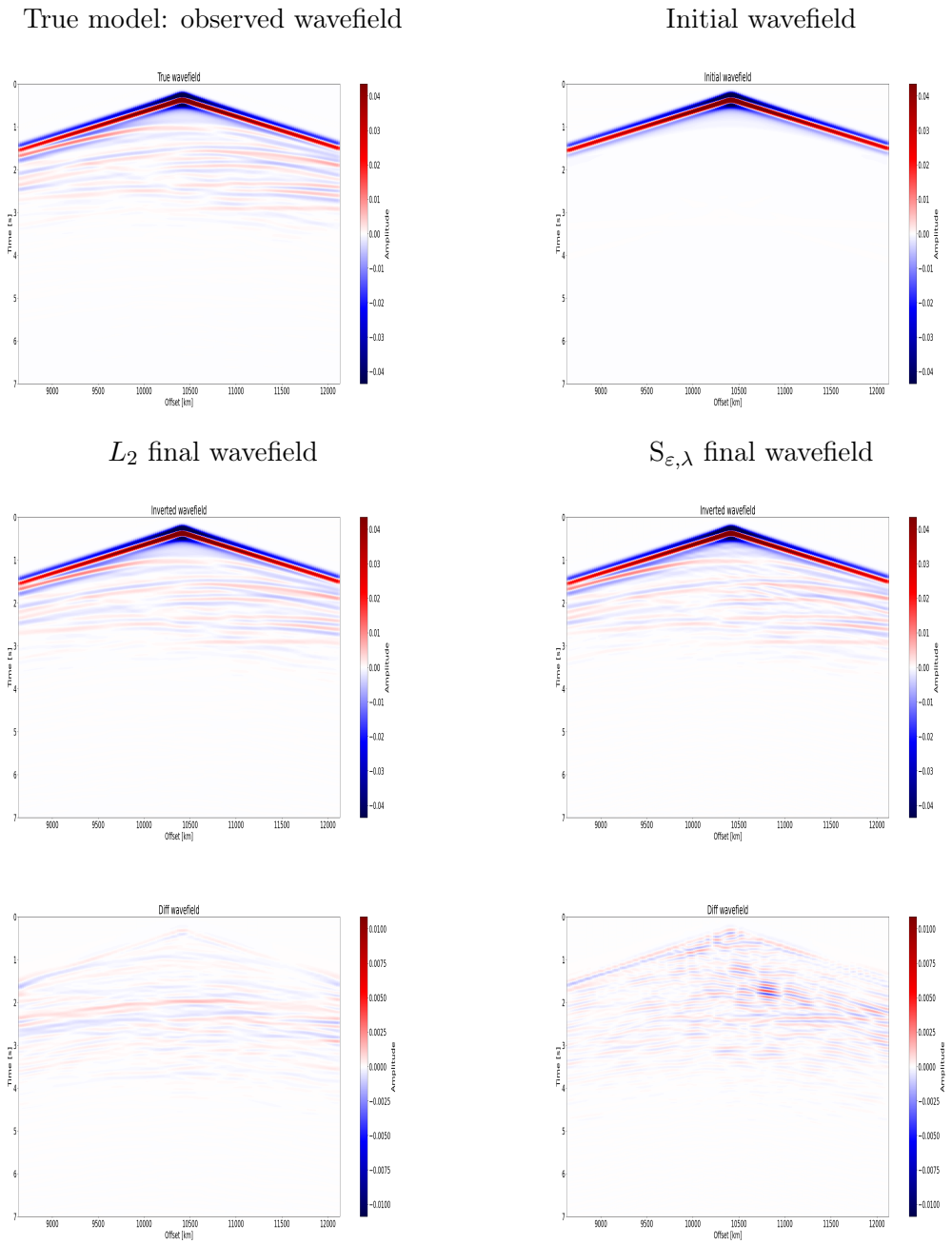


Figure 6.64: 1D-GRAD initial model: wavefield in the time-receiver domain for a common shot gather associated to a source et the middle of the domain: observed wavefield (upper-right figure); initial wavefield (upper-left figure); final wavefield for the inverted model and difference between the observed and final wavefield for L_2 (lower-left figures) and $S_{\varepsilon,\lambda}$ (lower-right figures).

Chapter 7

Conclusion

This thesis aims at investigating application of entropic unbalanced optimal transport in the context of full-waveform seismology. We provide a brief overview of full-waveform inversion in seismology and of the main issues encountered in practice and that are associated to the inherent ill-posedness and the non-convexity of the inversion problem, together with a brief review of different research directions that have been attempting to resolve these issues.

The premise of this study is that transport based distances can be a reasonably effective tool, partly because of their Lagrangian nature, for designing objective functions alternative to classical L^p based misfit functions that improve convexity and enlarge the basin of attraction of the full-waveform inversion problem. This has been already promoted by a number of studies investigating transport distances, such as 2-Wasserstein distance, earth mover's distance and Graph optimal distance. This Thesis can be seen as a contribution to this new trend.

More specifically in Chapter 4, we present the Sinkhorn divergence formulation of entropic unbalanced optimal transport, correcting the bias introduced by the entropic penalisation, that benefits from the easy to use and cost effective Sinkhorn algorithm associated to the entropic penalisation, and specifically discuss its use in the context of full-waveform inversion. The $\mathrm{S}_{\varepsilon,\lambda}$ distance still interprets signals as positive measures, requiring ad-hoc signal transformation into positive and negative parts (discussed in chapter 5) but unbalanced transport avoids ad-hoc normalisation methods that can often dampen features and reduce the intensity range of a signal.

This new misfit function is assessed in the most generic and reproducible way through several synthetic cases (presented in chapter 5 and 6) from 1D toy models (focussing on shifted patterns) to 2D FWI numerical illustrations on simple canonical configurations in transmission and reflection, and on more realistic heterogeneous configurations extracted from Marmousi models.

Comparison between misfit functions based on L^2 and different formulations of $\mathrm{S}_{\varepsilon,\varepsilon}$, associated to different signal transformations and ground cost metrics linked to transport domain scaling, is made simple by the adjoint state formalism of the FWI (presented in Chapter 2). For reproducibility, all misfit functions and numerical tests developed in this Thesis have been implemented within the Python Seismic Imaging Toolbox (PySIT), an open-source platform developed by Russel J. Hewett and Laurent Demanet in the Imaging and

Computing group of the Department of Mathematics at MIT.

The numerical illustrations provided in this Thesis make however a number of assumptions, for sake of simplicity and computing cost, that need to be kept in mind when reaching conclusions. First they all assume constant density, acoustic wave propagation (only one model parameter) and as such mode conversion and anisotropy are not considered. Second, all observables are point samples solution of the acoustic wave equation, using the same solver and without additional noise such as acquisition noise (full control set-up, so called inversion crime setting). Third seismic sources are assumed to be known and modelled as point sources with a Ricker wavelet source-time function.

As such, the numerical illustrations provided in this Thesis have to be considered as academic illustrations that need to be further investigated and evaluated in more realistic contexts before drawing any practical conclusion in term of field data applications.

At the end of this document, we cannot unfortunately provide a definitive conclusion on the relevance of using Entropic Unbalanced OT in FWI even just in the context of academic models. We here give where we stand and the various ideas arising from this work.

Based on chapter 6 we can say, as expected, that $S_{\epsilon,\lambda}$ provides similar results as the 2-Wasserstein distance promoted by the Engquist group under the use of signal sign splitting. It is yet unclear whether the introduction of unbalanced OT, avoiding signal normalisation, improves the inversion or not.

Introducing unbalanced (to avoid normalisation) and entropic (for its numerical performance and flexibility) OT is appealing but introduces two new parameters ϵ and λ which have so far simply been set by trial and visual assessment on the numerical tests. We do not have a clear understanding of the impact of these parameters on the convexity of the misfit and how they can be optimised. One of the conclusion is therefore that this should be first investigated in more details in the 2D parametric example (section 5.2). The potential use of these parameters in simulated annealing strategies could also be investigated.

On the positive side, our Entropic misfit function is easy to implement and the Sinkhorn divergence formulation really allows to set ϵ such that the number of Sinkhorn iterations to convergence remains acceptable. In all our experiments, it remained on the order of the discretisation of the offset line and in practice, one OT gradient step never exceeded 10 times one L^2 gradient step and was more like 5 times on average. This is encouraging and indicates that the extension of this method to $3D$ FWI is possible thanks to the tensorisation technique (section 4.1.2).

Another interesting research direction raised by this study is the choice of the ground metric, the metric used to measure the distance (in time $\times$ offset) traveled by the mass (here the positive or the negative part of the signal). A first discussion just on scaling the dimensions of the common shot gather is given in section 4.4 with two unrelated conclusions: first, scaling the domain will affect the modulus of convexity for simple translating signals, and second when using entropic OT the scale of the domain will constrain the transport range and can indeed be used to control it. In the OT theory and also numerical practice, it is legit to replace the quadratic euclidean metric by a Riemanian or even a Finsler metric. The open question is wether such a metric depending on the model could have better convexity properties and can this new non-linearity in

the inverse problem be tackled theoretically and numerically.

The strategy developed in this thesis remains based on the awkward interpretation of seismic wavefield as probability densities and we pay the heavy price of poorly understood non-linear signal transformations even with the unbalanced OT (see above) normalisation, even though it avoids signal normalisation that often dampen the features and reduces the intensity range. In contrast with the L^p distances that only considers differences in intensity, OT distances are transport distances and as such less sensitive to high frequency perturbations, the transport being on the order of the wavelength of the perturbation.

The representation of the data as a graph in the $(\mathbb{R}_{\text{amplitude}} \times \mathbb{R}_{\text{time}} \times \mathbb{R}^{d-1}_{\text{offset}})$ space proposed by the Metivier group is more natural and gives good results. There is a numerical price to pay though and it has remain until today limited to trace-by-trace comparisons. The application of Entropic OT maybe of interest for its numerical efficiency and also the range controlling property mentioned above.

Finally let us remark that the "graph" strategy proposed above can be pushed further, *e.g.* a discretised seismogram line (a collection of ordered points) can be interpreted as a Dirac measure in $\mathbb{R}^{d-1}_{\text{offset}} \times \mathbb{R}^{n_t}_{\text{time}} \times \mathbb{R}_{\text{amplitude}}$ and the OT distance can be defined on probability distributions over this space. This strongly reduces the number of samples of the data distributions but the computational burden now resides on the computation and storage of the ground cost. Interestingly, an efficient software exists (developed by J. Feydy `https://www.kernel-operations.io/geomloss/`), based on an online strategy to evaluate the ground cost and on a sophisticated GPU implementation, to compute Sinkhorn divergences in high dimensional spaces. It may be interesting to test if OT distances based on deformations in these higher dimensional spaces can be useful.

References

Aghamiry, H. S., A. Gholami, and S. Operto (2020), Accurate and efficient data-assimilated wavefield reconstruction in the time domain, *Geophysics*, *2*(A7-A12), doi:10.1190/geo2019-0535.1.

Aki, K., A. Christoffersson, and E. Husebye (1977), Determination of the three-dimensional seismic structure of the lithosphere, *Journal of Geophysical Research: Solid Earth*, *82*(2), 277–296, doi:10.1029/JB082i002p00277.

Alkhalifah, T. (2015), Scattering-angle based filtering of the waveform inversion gradients, *Geophysical Journal International*, *200*(1), 363–373, doi:10.1093/gji/ggu379.

Alkhalifah, T., and Y. Choi (2012), Taming waveform inversion non-linearity through phase unwrapping of the model and objective function, *Geophysical Journal International*, *191*(3), 1171–1176, doi:10.1111/j.1365-246X.2012.05699.x.

Ambrosio, L., E. Mainini, and S. Serfaty (2011), Gradient flow of the chapman-rubinstein-schatzman model for signed vortices, *Annales de l'Institut Henri Poincare (C), Analyse Non Lineaire*, *28*(2), 217–246, doi:10.1016/j.anihpc.2010.11.006.

Arbanel, S., and D. Gottlieb (1997), A mathematical analysis of the pml method, *Journal of Computational Physics*, *134*(2), 357–363, doi:10.1006/jcph.1997.5717.

Bamberger, A., G. Chavent, and P. Lailly (1977), Une application de la théorie du contrôle à un problème inverse sismique, *Annales de Géophysique*, *33*, 183–199.

Bamberger, A., G. Chavent, C. Hemon, and P. Lailly (1982), Inversion of normal incidence seismograms, *GEOPHYSICS*, *47*(5), 757–882, doi:10.1190/1.1441345.

Baysal, E., D. D. Kosloff, and J. W. C. Sherwood (1983), Reverse time migration, *GEOPHYSICS*, *48*(11), 1514–1524, doi:10.1190/1.1441434.

Benamou, J. D. (2003), Numerical resolution of an "unbalanced" mass transport problem, *Mathematical Modelling and Numerical Analysis*, doi:10.1051/m2an:2003058.

Benamou, J. D., and V. Duval (2018), Minimal convex extensions and finite difference discretisation of the quadratic Monge-Kantorovich problem, doi:10.1017/S0956792518000451.

Benamou, J. D., B. D. Froese, and A. M. Oberman (2014), Numerical solution of the Optimal Transportation problem using the Monge-Ampère equation, *Journal of Computational Physics*, *260*, 107–126, doi:10.1016/j.jcp.2013.12.015.

Benamou, J. D., G. Carlier, M. Cuturi, L. Nenna, and G. Peyŕ (2015), Iterative bregman projections for regularized transportation problems, *SIAM Journal on Scientific Computing*, *37*(2), A1111–A1138, doi:10.1137/141000439.

Benamou, J.-D., G. Carlier, and R. Hatchi (2018), A numerical solution to Monge's problem with a Finsler distance as cost, *ESAIM Math. Model. Numer. Anal.*, *52*, 2133–2148.

Berenger, J.-P. (1994), A perfectly matched layer for the absorption of electromagnetic waves, *Journal of computational physics*, *114*(2), 185–200, doi:10.1006/jcph.1994.1159.

Bermúdez, A., L. Hervella-Nieto, A. Prieto, and R. Rodríguez (2006), An optimal perfectly matched layer with unbounded absorbing function for time-harmonic acoustic scattering problems, *Journal of Computational Physics*, *223*(2), 469–488, doi:10.1016/j.jcp.2006.09.018.

Biondi, B., and A. Almondin (2014), Simultaneous inversion of full data bandwidth by tomographic full-waveform inversion, *GEOPHYSICS*, *79*(3), WA129–WA140, doi:10.1190/geo2013-0340.1.

Biondi, B., and W. W. Symes (2004), Angle-domain common-image gathers for migration velocity analysis by wavefield continuation imaging, *GEOPHYSICS*, *69*(5), 1283–1298, doi:10.1190/1.1801945.

Biswas, R., and M. Sen (2017), 2d full-waveform inversion and uncertainty estimation using the reversible jump hamiltonian monte carlo, in *SEG Technical Program Expanded Abstracts 2017*, pp. 120–1285, Society of Exploration Geophysicists, doi:segam2017-17680416.1.

Bogachev, V. (2007), *Measure thory*, 1 ed., Springer-Verlag Berlin Heidelberg, doi:10.1007/978-3-540-34514-5.

Borisov, D., F. Gao, P. Williamlson, and J. Tromp (2019), Application of 2d full waveform inversion on exploration land data, *GEOPHYSICS*, *85*(2), 1MA – Z8, doi:10.1190/geo2019-0082.1.

Boschi, L. (2006), Global multiresolution models of surface wave propagation: comparing equivalently regularized born and ray theoritical solutions, *Geophysical Journal International*, *167*(1), 238–252, doi:10.1111/j.1365-246X.2006.03084.x.

Boschi, L., and A. M. Dziewonski (2000), Whole earth tomography from delay times of p, pcp, and pkp phases: Latera heterogeneities in the outer core or radial anisotropy in the mantle?, *Journal of Geophysical Research: Solid Earth*, *105*(B6), doi:10.1029/2000JB900059.

Bozdağ, E., J. Tramper, and J. Tromp (2011), Misfit functions for waveform inversion baseds on instantaneous phase and envelope measurements, *Geophysical Journal International*, *185*(2), doi:10.1111/j.1365-246X.2011.04970.x.

Brenier, Y. (1991), Polar factorization and monotone rearrangement of vector-valued functions, *Communications on Pure and Applied Mathematics*, *44*(4), 375–417, doi:10.1002/cpa.3160440402.

Brossier, R., S. Operto, and J. Virieux (2009), Seismic imaging of complex onshore structures by 2d elastic frequency-domain full-waveform inversion, *GEOPHYSICS*, *74*(6), doi:10.1190/1.3215771.

Bunks, C., F. Salek, S. Zaleski, and G. Chavent (1995), Multiscale seismic inversion, *GEOPHYSICS*, *60*(5), 1457–1473, doi:10.1190/1.1443880.

Cance, P., and Y. Capdeville (2015), Validity of the acoustic approximation for elastic waves in heterogeneous media, *Geophysics*, *80*(4), T161–T173, doi:10.1190/geo2014-0397.1.

Capdeville, Y., and L. Métivier (2018), Elastic full-waveform inversion based on the homogeneization method: theoritical framework and 2-d numerical illustrations, *Geophysical Journal International*, *213*(2), 1093–1112, doi:10.1093/gji/ggy039.

Chauris, H. (2021), *Seismic imaging: a practical approach*, chap. 5. Full waveform inversion, pp. 123–146, EDP Sciences, doi:10.1051/978-2-7598-2351-2-007.

Chauris, H., and M. Noble (2001), Two-dimensional velocity macro model estimation from seismic reflection data by local differential semblance optimization: applications to synthetic and real data sets, *Geophysical Journal International*, *144*(1), 14–26, doi:10.1046/j.1365-246x.2001.00279.x.

Chavent, G. (1974), Identification of functional parameters in partial differential equations, in *Joint Automatic Control Conference*, 12, pp. 155–156.

Chavent, G., M. Dupuy, P. Lemmonier, et al. (1975), History matching by use of optimal theory, *Society of Petroleum Engineers Journal*, *15*(01), 74–86.

Chen, G., and Z. Wang (2017), Robust full-waveform inversion based on particle swarm optimization, in *SEG Technical Program Expanded Abstracts 2017*, pp. 1302–1306, doi:10.1190/segam2017-17634273.1.

Chizar, L., G. Peyré, B. Schmitzer, and F.-X. Vialard (2018), Unbalanced optimal transport: Dynamic and kantorovich formulations, *Journal of Functional Analysis*, *274*(11), 3090–3123, doi:10.1016/j.jfa.2018.03.008.

Chizat, L. (2017), Transport optimal de mesures positives: modèles, méthodes numériques, applications, Ph.D. thesis, Paris Sciences et Lettres.

Chizat, L., G. Peyré, B. Schmitzer, and F. X. Vialard (2018a), An Interpolating Distance Between Optimal Transport and Fisher–Rao Metrics, *Foundations of Computational Mathematics*, *18*(1), doi:10.1007/s10208-016-9331-y.

Chizat, L., G. Peyré, B. Schmitzer, and F.-X. Vialard (2018b), Scaling algorithms for unbalanced optimal transport problems, *Mathematics of Computation*, *87*(314), 2563–2609, doi:10.1090/mcom/3303.

Choi, Y., and T. Alkhalifah (2012), Application of multi-source waveform inversion to marine streamer data using the global correlation norm, *Geophysical Prospecting*, *60*(4), 748–758, doi:10.1111/j.1365-2478.2012.01079.x.

Claerbout, J. F. (1971), Toward a unified theory of reflector mapping, *GEOPHYSICS*, *36*(3), 467–618, doi:10.1190/1.1440185.

Claerbout, J. F. (1976), *Fundamentals of geophysical data processing with applications to petroleum prospecting*, McGraw-Hill, New York (N. H.), doi:10.1016/0148-9062(78)91918-6.

Claerbout, J. F., and S. M. Doherty (1972), Downward continuation of moveout-corrected seismograms, *GEOPHYSICS*, *37*(5), 741–909, doi:10.1190/1.1440298.

Clayton, R., and B. Engquist (1977), Absorbing boundary consitions for acoustic and elastic wave equations, *Bulletin of the Seismological Society of America*, *67*(6), 1259–1540, doi:10.1785/BSSA0670061529.

Cocher, E., H. Chauris, and R. Plessix (2017), Seismic iterative migration velocity analysis: two strategies to update the velocity model, *Computational Geosciences*, *21*(4), 759–780, doi:10.1007/s10596-017-9656-x.

Combettes, P. L., and J.-C. Pesquet (2011), Proximal splitting methods in signal processing, in *Fixed-point algorithms for inverse problems in science and engineering*, *Springer Optimization and Its Applications*, vol. 49, edited by H. Bauschke, R. Burachik, P. Combettes, V. Elser, D. Luke, and H. Wolkowicz, pp. 185–212, Springer, NY., doi:10.1007/978-1-4419-9569-8_10.

Conforti, G., and L. Tamanini (2021), A formula for the time derivative of the entropic cost and applications, *J. Funct. Anal.*, *280*(11), 108,964, doi:10.1016/j.jfa.2021.108964.

Cupillard, P., and Y. Capdeville (2018), Non-periodic homogeneization of 3-d elastic media for the seismic wave equation, *Geophysical Journal International*, *213*(2), 983–1001, doi:10.1093/gji/ggy032.

Cuturi, M. (2013), Sinkhorn distances: Lightspeed computation of optimal transport, *Advances in Neural Information Processing Systems*, pp. 1–9.

Dahlen, F. A., S.-H. Hung, and G. Nolet (2000), Fréchet kernels of finite-frequency traveltimes – i. theory, *Geophysical Journal International*, *141*(1), 157–174, doi:10.1046/j.1365-246X.2000.00070.x.

Datta, D., and M. Sen (2016), Estmating a starting model for full-waveform inversion using global optimization method, *GEOPHYSICS*, *81*(4), R211–R223, doi:10.1190/geo2015-0339.1.

Diaz, J., and P. Joly (2006), A time domain analysis of pml models in acoustics, *Computer Methods in Applied Mechanics and Engineering*, *195*(29-32), 3820–3853, doi:10.1016/j.cma.2005.02.031.

Donno, D., H. Chauris, and H. Calandra (2013), Estimating the background velocity model with the normalised integration method, in *75th EAGE Conference abstracts*, p. TU704, European Association of Geoscientists & Engineers, doi:10.3997/2214-4609.20130411.

Ekström, G. M., J. Tromp, and E. Larson (1997), Measurements and global models of surface wave propagation, *Journal of Geophysical Research: Solid Earth*, *102*(B4), 8137–8157, doi:10.1029/96JB03729.

Engquist, B., and B. D. Froese (2014), Application of the wasserstein metric to seismic signals, *Communications in Mathematical Sciences*, *12*(5), 979–988, doi:10.4310/CMS.2014.v12.n5.a7.

Engquist, B., and A. Majda (1977), Absorbing boundary conditions for numerical simulation of waves, *Proceedings of the National Academy of Sciences*, *74*(5), 1765–1766, doi:10.1073/pnas.74.5.1765.

Engquist, B., and Y. Yang (2019), Seismic inversion and the data normalisation for optimal transport, *Methods and Applications Analysis*, *26*(2), 133–148, doi:10.4310/MAA.2019.v26.n2.a3.

Engquist, B., and Y. Yang (2021), Optimal transport based seismic inversion: Beyond cycle skipping, *Communications on Pure and Applied Mathematics*, doi:10.1002/cpa.21990.

Engquist, B., B. D. Froese, and Y. Yang (2016), Optimal transport for seismic full waveform inversion, *Communications in Mathematical Sciences*, *14*(8), 2309–2330, doi:10.4310/CMS.2016.v14.n8.a9.

Engquist, B., K. Ren, and Y. Yang (2020), The quadratic wasserstein metric for inverse data matching, *Inverse Problems*, *36*(5), 055,001, doi:10.1088/1361-6420/ab7e04.

Festa, G., and J.-P. Vilotte (2005), The newmak scheme as velocity-stress time-staggering: an efficient pml implementation for spectral element simulations of elastodynamics, *Geophysical Journal International*, *161*(3), 789–812, doi:10.1111/j.1365-246X.2005.02601.x.

Feydy, J., T. Séjourné, F.-X. Vialard, S.-i. Amari, A. Trouvé, and G. Peyré (2019), Interpolating between optimal transport and mmd using sinkhorn divergences, in *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, *Proceedings of Machine Learning Research*, vol. 89, edited by K. Chaudluri and M. Sugiyama, pp. 2681–2690, PMLR.

Fichtner, A. (2011), *Full seismic waveform modelling and inversion*, Advances in Geophysical and Environmental Mechanics and Mathematics, Elsevier New York, doi:10.1007/978-3-642-15807-0.

Fichtner, A., and A. Zunino (2019), Hamiltonian nullspace shuttles, *Geophysical Research Letters*, *46*(2), 644–651, doi:10.1029/2018GL080931.

Fichtner, A., H.-P. Bunge, and H. Igel (2006a), The adjoint method in seismology - i. theory, *Physics of the Earth and Planetary Interiors*, *157*, 86–104, doi:10.106/j.pepi.2006.03.016.

Fichtner, A., H.-P. Bunge, and H. Igel (2006b), The adjoint method in seismology - ii. applications: travel times and sensitivity functionals, *Physics of the Earth and Planetary Interiors*, *157*(105-123), doi:10.1016/j.pepi.2006.03.018.

Fichtner, A., B. L. Kennett, H. Igel, and H.-P. Bunge (2008), Theoretical background for continental - and global - scale full waveform inversion in the time-frequency domain, *Geophysical Journal International*, *175*(2), doi:10.1111/j.1365-246X.2008.03923.x.

Fichtner, A., J. Trampert, P. Cupillard, E. Saygin, T. Taymaz, Y. Capdeville, and A. V. nor (2013), Multi-scale full waveform inversion, *Geophysical Journal International*, *194*(1), doi:10.1093/gji/ggt118.

Fichtner, A., A. Zunino, and L. Gebraad (2019), Hamiltonian monte carlo solution of tomographic problems, *Geophysical Journal International*, *216*(2), 1344–1363, doi:doi:10.1093/gji/ggy496.

Friederich, W. (1999), Propagation of seismic shear and surface waves in laterally heterogeneous mantle by multiple forward-scattering, *Geophysical Journal International*, *136*(1), 180–204, doi:10.1046/j.1365-246X.1999.00720.x.

Friederich, W. (2003), The s-velocity structure of the east asian mantle for inversion of shear and surface waveforms, *Geophysical Journal International*, *153*(1), 88–102, doi:10.1046/j.1365-246X.2003.01869.x.

Froese, B. D. (2012), A numerical method for the elliptic monge-amperè equation with transport boundary conditions, *SIAM Journal on Scientific Computing*, doi:10.1137/110822372.

Gangbo, W., and R. J. McCann (1996), The geometry of optimal transportation, *Acta Math.*, *177*, 113–161, doi:10.1007/BF02392620.

Gangbo, W., W. Li, S. Osher, and M. Puthawala (2019), Unormalised optimal transport, *Journal of Computational Physics*, *399*, 108,940, doi:https://doi.org/10.1016/j.jcp.2019.108940.

Gao, Y., H. Song, J. Zhang, and Z. Yao (2017), Comparison of artificial absorbing boundaries for acoustic wave equation modelling, *Exploration Geophysics*, *48*(1), 76–93, doi:10.1071/EG15068.

Gauthier, O., J. Virieux, and A. Tarantola (1986), Two-dimensional nonlinear inversion of seismic waveforms: Numerical results, *GEOPHYSICS*, *51*(7), 1341–1519, doi:10.1190/1.1442188.

Gazdag, J. (1978), Wave equation migration with the phase shift methode, *GEOPHYSICS*, *43*(7), 1337–1556, doi:10.1190/1.1440899.

Gebraad, L., C. Boehm, and A. Fichtner (2020), Bayesian elastic full-waveform inversion using hamiltonian monte carlo, *Journal of Geophysical Research: Solid Earth*, *125*, e2019JB018,428, doi:10.1029/2019JB018428.

Genevay, A., G. Peyré, and M. Cuturi (2018), Learning generative models with sinkhorn divergences, in *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics, AISTATS 2018*, *Proceedings of Machine Learning Research*, vol. 84, edited by A. Storkey and F. Perez-Cruz, pp. 1608–1617, PMLR.

Górszczyk, A., L. Métivier, and R. Brossier (2020), Relaxing the initial model constraint for crustal-scale full waveform inversion with graph-space optimal transport, in *Conference Proceedings, EAGE 2020 Conference and Exhibition*, vol. 2020, pp. 1–5, European Association of Geoscientists & Engineers, doi:10.3997/2214-4609.202010943.

Górszczyk, A., R. Brossier, and L. Metivier (2021), Graph-space optimal transport concept for time-domain full-waveform inversion of ocean-bottom seismometer data: Nanka through velocity structure reconstructed from a 1d model, *Journal of Geophysical Research*, *126*(5), e2020JB021,504, doi:10.1029/2020JB021504.

Grote, M. J., and I. Sim (2011), Local nonreflecting boundary condition for time-dependent multiple scattering, *Journal of Computational Physics*, *230*(8), 3135–3154, doi:10.1016/j.jcp.2011.01.017.

Gu, Y. J., A. M. Dziewonski, W. Su, and G. Ekström (2001), Models of the mantle shear velocity and discontinuities in the pattern of lateral heterogeneities, *Journal of Geophysical Research: Solid Earth*, *106*(B6), 1169–1199, doi:10.1029/2001JB000340.

Gung, Y., and B. Romanowicz (2004), Q tomography of the upper mantle using three-component long-period waveforms, *Geophysical Journal International*, *157*(2), 813–830, doi:10.1111/j.1365-246X.2004.02265.x.

Hedjazian, N., Y. Capdeville, and T. Bodin (2021), Multiscale seismic imaging with inverse homogenization, *Geophysical Journal International*, *226*(1), 676–691, doi:10.1093/gji/ggab121.

Hewett, J., and L. Demanet (2020), *PySIT: Python Seismic Imaging Toolbox v1.0*, doi:10.5281/zenodo.3603367.

Hewett, R., L. Demanet, and the PySIT Team (2013), PySIT: Python seismic imaging toolbox v0.5, release 0.5.

Huang, G., R. Nammour, and W. Symes (2017), Full-waveform inversion via soure-receiver extension, *GEOPHYSICS*, *82*(3), R153–R171, doi:10.1190/GEO2016-0301.1.

Huang, G., R. Nammour, and W. W. Symes (2018), Source-independent extended waveform inversion based on space-time source extension: Frequency-domain extension, *GEOPHYSICS*, *83*(5), R449–R461, doi:10.1190/geo2017-0333.1.

Jannane, M., et al. (1989), Wavelengths of earth structures that can be resolved from seismic reflection data, *GEOPHYSICS*, *54*(7), 806–932, doi:10.1190/1.1442719.

Kantorovich, L. V. (1942), On the translocation of masses, in *Journal of Mathematical Sciences*, doi:10.1007/s10958-006-0049-2.

Kantorovich, L. V., and G. S. Rubinshtein (1957), On a functional space and certain extremum problems., *Dokl. Akad. Nauk SSSR*, *115*, 1058–1061.

Komatitsch, D., and R. Martin (2007), An unsplit convolutional perfectly matched layer improved at grazing incidence for seismic wave equation in poroelastic media, *GEOPHYSICS*, *72*(SM155-SM167), doi:10.1190/1.2757586.

Komatitsch, D., J. Ritsema, and J. Trom (2002), The spectral-element method, beowulf computing, and global seismology, *Science*, *298*(5599), 1737–1742, doi:10.1126/science.1076024.

Komatitsch, D., Z. Xie, E. Bozdağ, E. S. de Andrade, D. Peter, Q. Liu, and J. Tromp (2016), Anelastic sensitivity kernels with parsimonius storage for adjoint tomography and full waveform inversion, *Geophysical Journal International*, *206*(3), 1467–1478, doi:10.1093/gji/ggw224.

Kondratyev, S., L. Monsaingeon, and D. Vorotnikov (2016a), A new optimal transport distance on the space of finite Radon measures, *Adv. Diff. Equations*, *21*, 1117–1164.

Kondratyev, S., L. Monsaingeon, and D. Vorotnikov (2016b), A new optimal transport distance on the space of finite radon measures, *Advances in Differential Equations*.

Lailly, P. (1983), The seismic inverse problem as a sequence before stack migration, in *Conference on inverse scattering: theory and applications*, pp. 206–220, Society for Industrial and Applied Mathematics, Philadelphia, PA.

Lambaré, G. (2008), Stereotomography, *Geophysics*, *73*(5), VE25–VE34, doi:10.1190/1.2952039.

Lellmann, J., D. A. Lorenz, C. Schönlieb, and T. Vlakonen (2014), Imaging with kantorovich-rubinstein discrepency, *SIAM Journal on Imaging Sciences*, *7*(4), 2833–2859, doi:10.1137/140975528.

Léonard, C. (2014), A survey of the Schrödinger problem and some of its connections with optimal transport, *Discrete Contin. Dyn. Syst.*, *34*, 1533–1574.

Lévy, B., and E. L. Schwindt (2018), Notions of optimal transport theory and how to implement them on a computer, *Compututers & Graphics*, *72*, 135–148, doi:10.1016/j.cag.2018.01.009.

Li, D., M. P. Lamoureux, and W. Liao (2021), Application of an unbalanced optimal transport distance and a mixed l1/wasserstein distance to full waveform inversion.

Li, X.-D., and B. Romanowicz (1996), Global mantle shear velocity model developed using nonlinear asymptotic coupling theory, *Journal of Geophysical Research: Solid Earth*, *101*(B10), 22,245–22,272, doi:10.1029/96JB01306.

Li, Y. E., and L. Demanet (2015), Phase and amplitude tracking for seismic-event separation, *GEOPHYSICS*, *80*(6), WD59–WD72, doi:10.1190/geo2015-0075.1.

Liero, M., A. Mielke, and G. Savaré (2018), Optimal Entropy-Transport problems and a new Hellinger–Kantorovich distance between positive measures, *Inventiones Mathematicae*, *211*(3), 969–1117, doi:10.1007/s00222-017-0759-8.

Lions, J. L. (1971), *Optimal control of systems governed by partial differential equations*, *Grundelhren der mathematischen Wisseschaften*, vol. 170, Springer-Verlag, Berlin Heidelberg.

Liu, D. C., and J. Nocedal (1989), On the limited memory bfgs method for large scale optimisation, *Mathematical Programming*, *45*, 503–528, doi:10.1007/BF01589116.

Liu, J., H. Chauris, and H. Calandra (2012), The normalised integration method - and alternative to full waveform inversion?, in *Symposium on the Application of Geophysics to Engineering and Environmental Problems Proceedings*, pp. 5–8, Environmental & Engineering Geophysical Society, doi:10.3997/2214-4609.20144373.

Liu, Q., and J. Tromp (2008), Finite-frequency sensitivity kernels for global seismic wave propagation based upon adjoint methods, *Geophysical Journal International*, *174*(1), doi:10.1111/j.1365-246X.2008.03798.x.

Lombardi, D., and E. Maitre (2015), Eulerian models and algorithms for unbalanced optimal transport, *ESAIM: Mathematical Modelling and Numerical Analysis*, doi:10.1051/m2an/2015025.

Luo, J., and R.-S. Wu (2015), Seismic envelope inversion: reduction of local minima and noise resistance, *Geophysical Prospecting*, *63*(3), 597–614, doi:10.1111/1365-2478.12208.

Luo, J., R. shan Xu, and F. Gao (2018), Time-domain inversion using instantaneous phase information with damping, *Journal of Geophysics and Engineering*, *15*(1032-1041), doi:10.1088/1742-2140/aaa984.

Luo, S., and P. Sava (2011), A decovolution-based objective function for wave-equation inversion, in *SEG Technical Program Expanded Abstracts 2011*, pp. 2788–2792, Society of Exploration Geophysicists, doi:10.1190/1.3627773.

Luo, Y., and G. Schuster (1991), Wave-equation travel time inversion, *GEOPHYSICS*, *56*(5), 598–730, doi:10.1190/1.1443081.

Mainini, E. (2012), A description of transport cost for signed measures, *Journal of Mathematical Sciences*, doi:10.1007/s10958-012-0718-2.

Martin, G. S., R. Wiley, and K. Marfurt (2006), Marmousi2: an elastic upgrade for marmousi, *The Leading Edge*, *25*(2), 156–166, doi:10.1190/1.2172306.

Masters, G., S. Johnson, G. Laske, H. Bolton, and J. Davies (1996), A shear-velocity model of the [and discussion], *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, *354*(1711), 1385–1411, doi:10.1098/rsta.1996.0054.

Mégnin, C., and B. Romanowicz (2000), The three-dimensional shear velocity structure from the inversion of body, surface and higher-mode waveforms, *Geophysical Journal International*, *143*(3), 709–728, doi:10.1046/j.1365-246X.2000.00298.x.

Mérigot, Q. (2011), A multiscale approach to optimal transport, *Eurographics Symposium on Geometry Processing*, doi:10.1111/j.1467-8659.2011.02032.x.

Mérigot, Q., and B. Thibert (2021), Xhapter 2 - optimal transport: discretization and algorithms, in *Geometric Partial Differential Equations - Part II*, *Handbook of Numerical Analysis*, vol. 22, edited by A. Bonito and R. H. Nochetto, pp. 133–212, Elsevier, doi:10.1016/bs.hna.2020.10.001.

Messud, J., and A. Sedova (2019), Multidomensional optimal transport for 3d fwi: Demonstration of field data, in *Conference Proceedings, 81st EAGE Conference and Exhibition 2019*, vol. 2019, pp. 1–5, doi: 10.3997/2214-4609.201900869.

Métivier, L., R. Brossier, Q. Mérigot, E. Oudet, and J. Virieux (2016a), Measuring the misfit between seismograms using an optimal transport distance: Application to full waveform inversion, *Geophysical Journal International*, *205*(1), 345–377, doi:10.1093/gji/ggw014.

Métivier, L., R. Brossier, Q. Mérigot, E. Oudet, and J. Virieux (2016b), Increasing the robustness and applicability of the full-waveform inversion: An optimal transport distance strategy, *The Leading Edge*, *35*(2), 1060–1067, doi:10.1190/tle35121060.1.

Métivier, L., R. Brossier, Q. Mérigot, E. Oudet, and J. Virieux (2016), An optimal transport approach for seismic tomography: Application to 3D full waveform inversion, *Inverse Problems*, *32*(11), 115,008, doi:10.1088/0266-5611/32/11/115008.

Métivier, L., A. Allain, R. Brossier, Q. Mérigot, E. Oudet, and J. Virieux (2018), Optimal transport for mitigating cycle skipping in full-waveform inversion: A graph-space transform approach, *Geophysics*, *83*(5), R515–R540, doi:10.1190/geo2017-0807.1.

Metivier, L., R. Brossier, Q. Merigot, and E. Oudet (2019), A graph space optimal transport distance as a generalisation of the l^p distances: application to a seismic imaging inverse problem, *Inverse problems*, *35*(8), doi:10.1088/1361-6420/ab206f.

Monge, G. (1781), Mémoire sur la théorie des déblais et de remblais, *Histoire de l'Académie Royale des Sciences de Paris, avec les Mémoires de Mathématique et de Physique pour la même année*, pp. 666–705.

Montelli, R., G. Nollet, F. Dahlen, G. Masters, E. R. Engdahl, and S.-H. Hung (2004), Finite-frequency tomography reveals a variety of plumes in the mantle, *Science*, *303*(5656), 338–343, doi:10.1126/science.1092485.

Mora, P. (1987), Nonlinear two-dimensional elastic inversion of multioffset seismic data, *GEOPHYSICS*, *52*(9), 1211–1228, doi:10.1190/1.1442384.

Nataf, F. (2013), Absorbing boundary conditions and perfectly matched layers in wave propagation problems, in *Direct and Inverse Problems in Wave Propagation and Applications*, *Radon series on computational and applied mathematics*, vol. 14, edited by I. Graham, U. Langer, J. Melenk, and M. Sini, pp. 219–232, De Gruyter, doi:10.1515/9783110282283.219.

Nemeth, T., E. Normark, and F. Qin (1997), Dynamic smoothing in crosswell traveltime tomography, *GEOPHYSICS*, *62*(1), doi:10.1190/1.1444115.

Nemeth, T., C. Wu, and G. T. Schuster (1999), Least-square migration of incomplete reflection data, *GEOPHYSICS*, *64*(1), 13–310, doi:10.1190/1.1444517.

Nemirovski, A., and U. Rothblum (1999), On complexity of matrix scaling, *Linear Algebra and its Applications*, *302-303*, 435–460, doi:10.1016/S0024-3795(99)00212-8.

Nocedal, J. (1980), Updating quasi-newton matrices with limited storage, *Mathematics of Computation*, *35*(151), 772–782, doi:10.2307/2006193.

Nocedal, J., and S. J. Wright (2006), *Numerical Optimization: Theoretical and Practical Aspects*, Springer series in operation research and financial engineering, second edition ed., Springer, New York, NY, USA.

Operto, S., Y. Gholami, V. Prieux, A. Robodetti, R. Brossier, L. Métivier, and J. Virieux (2013), A guided tour of multiparameter full-waveform inversion with multi-component data: From theoty to practice, *The Leading Edge*, *32*(9), 1040–1054, doi:10.1190/tle32091040.1.

Operto, S., A. Miniussi, R. Brossier, L. Combe, L. Métivier, V. Monteiller, A. Ribodetti, and J. Virieux (2015), Efficient 3-d frequency-domain mono-parameter full-waveform inversion of ocean-bottom cable data: application to valhall in the visco-acoustic vertical transverse isotropic approximationb, *Geophysical Journal International*, *202*(2), 1362–1391, doi:10.1093/gji/ggv226.

Pal, S. (2019), On the difference between entropic cost and the optimal transport cost, https://arxiv.org/pdf/1905.12206.pdf.

Pasalic, D., and R. McGarry (2010), Convolutional perfectly matched layer for isotropic and anisotropic acoustic wave equations, in *SEG Technical Program Expanded Abstracts 2010*, pp. 2925–2929, Society of Exploration Geophysicists, doi:10.1190/1.3513453.

Peter, D., et al. (2011), Forward and adjoint simulations of seismiv wave propagation on fully unstructured hexahedral meshes, *Geophysical Journal International*, *186*(721-739), doi:10.1111/j.1365-246X.2011.05044.x.

Peyré, G., and M. Cuturi (2019), Computational optimal transport, *Foundations and Trends in Machine Learning*, *11*(5-6), 1–257, doi:10.1561/2200000073.

Piccoli, B., and F. Rossi (2014), Generalized Wasserstein Distance and its Application to Transport Equations with Source, *Archive for Rational Mechanics and Analysis*, doi:10.1007/s00205-013-0669-x.

Piccoli, B., and F. Rossi (2016), On Properties of the Generalized Wasserstein Distance, *Archive for Rational Mechanics and Analysis*, doi:10.1007/s00205-016-1026-7.

Pladys, A., R. Brossier, Y. Li, and L. Metivier (2021), On cycle-skipping function modification for full-waveform inversion: comparison of five recent approaches, *GEOPHYSICS*, *86*(4), R563–R587, doi:10.1190/geo2020-0851.1.

Plessix, R.-E. (2006), A review of the adjoint-state method for computing the gradient of a functional with geophysical applications, *Geophysical Journal International*, *167*(2), 495–503, doi:10.1111/j.1365-246X.2006.02978.x.

Plessix, R.-E., and C. Perkins (2010), Full waveform inversion of a deep water ocean bottom seismometer dataset, *First Break*, *28*(4), 71–78, doi:10.3997/1365-2397.2010013.

Plessix, R.-E., W. Mulder, and A. ten Kroode (2000), Automatic cross-well tomography by semblance and differential semblance optimization: Theory and gradient computation, *Geophysical Prospecting*, *48*(5), 919–935, doi:10.1046/j.1365-2478.2000.00217.x.

Plessix, R.-E., G. Baeten, J. W. de Maag, F. ten Kroode, and Z. Rujie (2012), Full waveform inversion and distance separated simultaneous sweeping: a study with a land seismic data set, *Geophysical Prospecting*, *60*(4), 733–747, doi:10.1111/j.1365-2478.2011.01036.x.

Poncet, R., J. Messud, M. Bader, G. Lambaré, G. Viguier, and C. Hidalgo (2018), Fwi with optimal transport: a 3d implementation and an application on a field dataset, in *Conference Proceedings, 80th EAGE Conference and Exhibition 2018*, pp. 1–5, European Association of Geoscientists & Engineers, doi:10.3997/2214-4609.201801029.

Pratt, R., Z.-M. Song, P. Williamson, and M. Warner (1996), Two-dimensional velocity models from wide-angle seismic data wevefield inversion, *Geophysical Journal International*, *124*(2), 323–340, doi:10.1111/j.1365-246X.1996.tb07023.x.

Pratt, R. G. (1990), Inverse theory applied to multi-source cross-hole tomography. part 2: Elastic wave-equation method, *Geophysical Prospecting*, *38*(3), 311–329, doi:10.1111/j.1365-2478.1990.tb01847.x.

Pratt, R. G. (1999), Seismic waveform inversion in the frequency domain, part i: Theory and verification in a physical scale model, *GEOPHYSICS*, *64*(3), 888–901, doi:10.1190/1.1444597.

Pratt, R. G., and N. R. Goulty (1991), Combining wave-equation imaging with travel time tomography to form high-resolution images from crosshole data, *GEOPHYSICS*, *56*(2), 168–323, doi:10.1190/1.1443033.

Pratt, R. G., and R. M. Schipp (1999), Seismic waveform inversion in the frequency domain – part 2: Faukt delineation in sediments using crosshole data, *GEOPHYSICS*, *64*(3), 902–914, doi:10.1190/1.1444598.

Pratt, R. G., and M. H. Worthington (1990), Inverse inverse theory applied to multi-source cross-hole tomography. part 1: acoustic wave-equation method, *Geophysical Prospecting*, *38*(3), 287–310, doi:10.1111/j.1365-2478.1990.tb01846.x.

Qiu, L., J. Ramos-Martinez, A. Valenciano, Y. Yang, and B. Engquist (2017), Full-waveform inversion with an exponentially encoded optimal-transport, in *SEG Technical Program Expanded Abstracts 2011*, pp. 1286–1290, Society of Exploration Geophysicists, doi:10.1190/segam2017-17681930.1.

Rawlinson, N., and M. Sambridge (2003), Seismic traveltime tomography of the crust and lithosphere, *Advances in Geophysics*, *46*, 81–198, doi:10.1016/S0065-2687(03)46002-0.

Ray, A., A. Sekar, G. M. Hoversten, and U. Albertin (2016), Frequency domain full waveform elastic inversion of marine seismic data from the alba field using bayesian trans-dimensional algorithm, *Geophysical Journal International*, *205*(2), 915–937, doi:10.1093/gji/ggw061.

Ray, A., S. Kaplan, J. Washbourne, and U. Albertin (2018), Low frequency full waveform seismic inversion within a tree based bayesian framework, *Geophysical Journal International*, *212*(1), 522–542, doi:10.1093/gji/ggx428.

Ritsema, J., H. J. van Heijst, and J. Woodhouse (1999), Complex shear wave velocity structure imaged beneath africa and iceland, *Science*, *286*(5446), 1925–1928, doi:10.1126/science.286.5446.1925.

Roden, J. A., and S. D. Gedney (2000), Convolution pml (cpml): An efficient fdtd implementation of the cfs-pml for arbitrary media, *Microwave and Optical Technology Letters*, *27*(5), 334–339, doi:10.1002/1098-2760(20001205)27:5<334::AID-MOP14>3.0.CO;2-A.

Romanowicz, B. (2000), Anomalous splitting of free oscillations: A reevaluation of possible interpretations, *Geophysical Research Letters*, *105*(B9), 21,559–21,578, doi:10.1029/GL016i005p00401.

Romanowicz, B. (2003), Global mantle tomography: progress status in the past 10 years, *Annual Review of Earth and Planetary Sciences*, *31*(1), 303–328, doi:10.1146/annurev.earth.31.091602.113555.

Sambridge, M., and K. Gallagher (2011), Inverse theory, monte carlo method, in *Encyclopedia of Solid Earth Geophysics*, edited by G. H.K., Encyclopia of Earth Sciences, Springer Netherlands, Dordrecht, doi:10.1007/978-90-481-8702-7_192.

Santambrogio, F. (2015), *Optimal Transport for Applied Mathematicians*, *Progress in nonlinear differential equations and their applications*, vol. 87, Birkhäuser Basel, doi:10.1007/978-3-319-20828-2.

Sava, P., and S. Fomel (2006), Time-shift imaging condition in seismic migration, *GEOPHYSICS*, *71*(6), S209–S217, doi:10.1190/1.2338824.

Schmitzer, B. (2019), Stabilized Sparse Scaling Algorithms for Entropy Regularized Transport Problems, *SIAM Journal on Scientific Computing*, *41*(3), A1443–A1481, doi:10.1137/16m1106018.

Schuster, G. T., and A. Quintus-Bosz (1993), Wavepath eikonal travel time inversion: Theory, *GEOPHYSICS*, *58*(9), 1248–1387, doi:10.1190/1.1443514.

Sedova, A., J. Messud, H. Prigent, S. Masclet, G. Royle, and G. Lambaré (2019), Acoustic land full waveform inversion on a broadband land dataset: The impact of optimal transport, in *Conference Proceedings, 81st EAGE Conference and Exhibition 2019*, vol. 2019, pp. 1–5, doi:10.3997/2214-4609.201901568.

Séjourné, T., J. Feydy, F.-X. Vialard, A. Trouvé, and G. Peyré (2019), Sinkhorn divergences for unbalanced optimal transport, doi:arXiv:1910.12958, arXiv preprint.

Shen, P. (2004), Wave equation migration velocity analysis by differential semblance optimization, Ph.D. thesis, Rice University.

Shen, P., and W. W. Symes (2008), Automatic velocity analysis vi shot profile migration, *GEOPHYSICS*, *73*(5), VE49–VE59, doi:10.1190/1.2972021.

Shen, P., W. Symes, and C. Stolk (2003), Differential semblance velocity analysis by wave-equation migration, in *SEG Technical Program Expanded Abstracts 2003*, pp. 2132–2135, Society of Exploration Geophysicists, doi:10.1190/1.1817759.

Shen, X., I. Ahmed, A. Brenders, J. Dellinger, J. Etgen, and S. Michell (2018), Full waveform inversion: The next leap forward in subsalt imagin, *The Leading Edge*, *37*, 67b1–67b6, doi:10.1190/tle37010067b1.1.

Shipp, R. M., and S. C. Singh (2002), Two-dimensional full wavefield inversion of wide-aperture marine seismic streamer data, *Geophysical Journal International*, *151*(2), 325–344, doi:10.1046/j.1365-246X.2002.01645.x.

Sigloch, K., N. McQuarrie, and G. Nolet (2008), Two-stage subduction history under north america inferred from multiple-frequency tomography, *Nature Geoscience*, *1*(7), 458–462, doi:10.1038/ngeo231.

Sim, I. (2010), Nonreflecting boundary conditions for the time dependent wave equation, Ph.D. thesis, University of Basel, Lausane, doi:10.5451/unibas-005269594.

Sinkhorn, R., and P. Knopp (1967), Concerning nonnegative matrices and doubly stochastic matrices, *Pacific Journal of Mathematics*, *21*(2), 343–348, doi:http://dx.doi.org/10.2140/pjm.1967.21.343.

Sirgue, L., and R. G. Pratt (2004), Efficient waveform inversion and imaging: A strategy for selecting temporal frequencies, *GEOPHYSICS*, *69*(1), 16–297, doi:10.1190/1.1649391.

Sirgue, L., O. Barkved, J. Dellinger, J. Etgen, U. Albertin, and J. Kommedal (2010), Full waveform inversion: the next leap forward in imaging at valhall, *First Break*, *28*(4), 65–70, doi:10.3997/1365-2397.2010012.

Stolt, R. H. (1978), Migration by fourier transform, *GEOPHYSICS*, *143*(1), 23–48, doi:10.1190/1.1440826.

Stopin, A., R.-E. Plessix, and S. A. Abri (2014), Multiparameter waveform inversion of a large wide-azimuth low-frequency land data set in oman, *GEOPHYSICS*, *79*(3), 1MJ – Z84, doi:10.1190/geo2013-0323.1.

Sun, B., and T. Alkhalifah (2019), The application of an optimal transport to a preconditionned data matching function for robust waveform inversion, *GEOPHYSICS*, *84*(6), R923–R945, doi:10.1190/geo2018-0413.1.

Sun, D., and W. W. Symes (2012), Waveform inversion via nonlinear differential semblance optimization, in *SEG Technical Program Expanded Abstracts 2012*, pp. 1–7, Society of Exploration Geophysicists, doi:https://doi.org/10.1190/segam2012-1190.1.

Symes, W. W. (2008), Migration velocity analysis and waveform inversion, *Geophysical Prospecting*, *56*(6), 765–790, doi:10.1111/j.1365-2478.2008.00698.x.

Symes, W. W., and J. J. Carrazone (1991), Velocity inversion by differential semblance optimization, *GEOPHYSICS*, *56*(5), 598–730, doi:10.1190/1.1443082.

Tang, Y., S. Lee, A. Baumstein, and D. Hinkley (2013), Tomographically enhanced full wavefield inversion, in *SEG Technical Program Expanded Abstracts 2013*, pp. 1037–1041, Society of Exploration Geophysicists, doi:10.1190/segam2013-1145.1.

Tape, C., Q. Liu, and J. Tromp (2007), Finite-frequency tomography using adjoint methods - methodology and examples using membrane surface waves, *Geophysical Journal International*, *168*(3), 1105–1129, doi:10.1111/j.1365-246X.2006.03191.x.

Tarantola, A. (1984a), Inversion of seismic reflection data in the acoustic approximation, *GEOPHYSICS*, *49*(3), 1140–1395, doi:10.1190/1.1441754.

Tarantola, A. (1984b), Linearized inversion of seismic reflection data, *Geophysical Prospecting*, *32*(6), 998–1015, doi:10.1111/j.1365-2478.1984.tb00751.x.

Tarantola, A. (1987), *Inverse problem theory: methods for data fitting and model parameter estimation*, Elsevier New York.

Tarantola, A. (1988), A theoretical background for the inversion of seismic waveforms including elasticity and attenuation, *Pure and Applied Geophysics*, *128*, 365–399, doi:10.1007/BF01772605.

Tarantola, A. (2005), *Inverse problem theory and methods for model parameter estimation*, Society for Industrial and Applied Mathematics (SIAM), doi:10.1137/1.9780898717921.

Thorpe, M., S. Park, S. Kolouri, G. K. Rohde, and D. Slepčev (2017), A transportation l^p distance for signal analysis, *Journal of Mathematical Imaging and Vision*, *59*(2), 187–210, doi:10.1007/s10851-017-0726-4.

Trampert, J., and J. H. Woodhouse (1995), Global phase velocity maps of love and raileigh waves between 40 and 150 seconds, *Geophysical Journal International*, *122*(2), 675–690, doi:10.1111/j.1365-246X.1995.tb07019.x.

Tromp, J. (2020), Seismic wavefield imaging of the earth'sinterior across scales, *Nature Reviews Earth & Environment*, *1*(1), 40–53, doi:10.1038/s43017-019-0003-8.

Tromp, J., C. Tape, and Q. Liu (2005), Seismic tomography, adjoint methods, time reversal and banana-doughnut kernels, *Geophysical Journal International*, *160*(1), 195–216, doi:10.1111/j.1365-246X.2004.02453.x.

van Leeuwen, T., and F. J. Herrmann (2013), Mitigating local minima in full-waveform inversion by expanding the seach space, *Geophysical Journal International*, *195*(1), 661–667, doi:10.1093/gji/ggt258.

van Leeuwen, T., and W. Mulder (2008), Velocity analysis based on data correlation, *Geophysical Prospecting*, *56*(6), 791–803, doi:10.1111/j.1365-2478.2008.00704.x.

van Leeuwen, T., and W. Mulder (2010), A correlation-based misfit criterion for wave-equation traveltime tomography, *Geophysical Journal International*, *182*(3), 1383–1384, doi:10.1111/j.1365-246X.2010.04681.x.

Čeverný, V., and I. Pšenčík (2011), Seismic, ray theory, in *Encyclopedia of Exploration Geophysics*, edited by H. K. Gupta, pp. 1244–1258, Springer Netherlands, Dordrecht, doi:10.1007/978-90-481-8702-7_53.

Čeverný, V., and J. E. Soares (1992), Fresnel volume ray tracing, *GEOPHYSICS*, *57*(7), 902–915, doi:10.1190/1.1443303.

Vershik, A. (2013), Long history of the monge-kantorovich transportation problem, *Math Intelligencer*, *35*(4), 1–9, doi:10.1007/s00283-013-9380-x.

Versteeg, R. (1994), The marmousi experience: velocity model determination on a synthetic complex data set, *The Leading Edge*, *13*(9), 927–936, doi:10.1190/1.1437051.

Vialard, F.-X. (2019), An elementary introduction to entropic regularization and proximal methods for numerical optimal transport, notes presented in the CEA-EDF-INRIA summer school.

Vigh, D., K. Jiao, D. Watts, and D. Sun (2014), Elastic full-waveform inversion application using multicomponent measurements of seismic data collection, *GEOPHYSICS*, *79*(2), R63–R77, doi:10.1190/geo2013-0055.1.

Villani, C. (2009), *Optimal Transport, Grundlehren der mathematischen Wissenschaffen*, vol. 338, Springer-Verlag Berlin Heidelberg, doi:10.1007/978-3-540-71050-9.

Virieux, J., and S. Operto (2009), An overview of full-waveform inversion in exploration geophysics, *GEOPHYSICS*, *74*(6), doi:10.1190/1.3238367.

Virieux, J., A. Asnaashari, R. Brossier, L. Métivier, A. Ribodetti, and W. Zhou (2014), An introduction to full waveform inversion, in *Encyclopedia of Exploration Geophysics*, Geophysical References Series: R1-1-R1-40, chap. 6, Society of Exploration Geophysicists, doi:10.1190/1.9781560803027.entry6.

Wang, C., D. Yingst, P. Farmer, and J. Levei (2016), Full waveform inversion with the reconstructed wavefield method, in *SEG Technical Program Expanded Abstracts 2016*, pp. 1237–1241, Society of Exploration Geophysicists, doi:segam2016-13870317.1.

Wang, Y., and Y. Rao (2009), Reflection seismic waveform tomography, *Journal of Geophysical Research: Solid Earth (1978–2012)*, *114*(B3), 1978–2012, doi:10.1029/2008JB005916.

Warner, M., and L. Guasch (2016), Adaptive waveform inversion: Theory, *GEOPHYSICS*, *81*(6), doi:10.1190/geo2015-0387.1.

Warner, M., A. Ratcliffe, T. Nangoo, J. Morgan, A. Umpleby, and N. Shah (2013), Anisotropic 3d full-waveform inversion, *GEOPHYSICS*, *78*(2), 1MA–Z51, doi:10.1190/geo2012-0338.1.

Wolfe, P. (1969), Convergence conditions for ascent methods, *SIAM review*, *11*(2), 226–235, doi:10.1137/1011036.

Wolfe, P. (1971), Convergence conditions for ascent methods. ii: some cirrections, *SIAM review*, *13*(2), 185–188, doi:10.1137/1013035.

Woodward, M. J., D. Nichols, O. Zdraweva, P. Whitfield, and T. Johns (1973), A decade of tomography, *GEOPHYSICS*, *73*(5), VE5–VE11, doi:10.1190/1.2969907.

Wu, R.-S., J. Luo, and B. Wu (2014), Seismic envelope inversion and modulation signal model, *GEOPHYSICS*, *79*(3), WA13–WA24, doi:10.1190/geo2013-0294.1.

Yang, Y. (2018), Optimal transport for seismic inverse problems, Ph.D. thesis, University of Texas at Austin, doi:10.15781/T2RR1Q49P.

Yang, Y., and B. Engquist (2017), Analysis of optimal transport and related misfit functions in full-waveform inversion, in *SEG Technical Program Expanded Abstracts 2017*, pp. 1291–1296, Society of Exploration Geophysicists, doi:10.1190/segam2017-17754089.1.

Yang, Y., and B. Engquist (2018), Analysis of optimal transport related misfit functions in full-waveform inversion, *GEOPHYSICS*, *83*(1), A7–A12, doi:10.1190/geo2017-0264.1.

Yang, Y., B. Engquist, J. Sun, and B. F. Hamfeldt (2018), Application of optimal transport and the quadratic wasserstein metric to full-waveform inversion, *GEOPHYSICS*, *83*(1), R43–R62, doi:10.1190/geo2016-0663.1.

Yilmaz, Ö. (2001), *Seismic data analysis*, Society of Exploration Geophysicists.

Yoshizawa, K., and B. L. N. Kennett (2004), Multimode surface wave tomography for the australian region using three-stage approach incorporating finite frequency effects, *Journal of Geophysical Research: Solid Earth*, *109*(B2), doi:10.1029/2002JB002254.

Yoshizawa, K., and B. L. N. Kennett (2005), Seisitivity kernels for finte-frequency surface waves, *Geophysical Journal International*, *162*(3), 910–926, doi:10.1111/j.1365-246X.2005.02707.x.

Zelt, C. A. (1999), Modelling strategies and model assessment for wide-angle seismic travel time data, *Geophysical Journal International*, *139*(1), 183–204, doi:10.1046/j.1365-246X.1999.00934.x.

Zhao, L., T. H. Jordan, and C. H. Chapman (2000), Three-dimensional fréchet differential kernels for seismicdelay time, *Geophysical Journal International*, *141*(3), 558–576, doi:10.1046/j.1365-246x.2000.00085.x.

Zhu, H., and S. Fomel (2016), Building good starting model for full-waveform inversion using adaptive matching filter misfit, *GEOPHYSICS*, *81*(5), U61–U72, doi:10.1190/geo2015-0596.1.

Zhu, H., E. Bozdağ, and J. Tromp (2015), Seismic structure of the european upper mantle based on adjoint tomography, *Geophysical Journal International*, *201*(1), 18–52, doi:10.1093/gji/ggu492.