UNIVERSITE DE NICE-SOPHIA ANTIPOLIS - UFR SCIENCES
Ecole Doctorale Sciences Fondamentales et Appliquées

THESE

pour obtenir le titre de

Docteur en Sciences
de l'Université de Nice-Sophia Antipolis

Spécialité : Sciences de l'Univers - Géophysique

présentée et soutenue par
Clara CASTELLANOS LOPEZ

Speed-up and regularization techniques for seismic full waveform inversion

Accélération et régularisation de la méthode d'inversion des formes d'ondes complètes en exploration sismique

Thèse dirigée par **Stéphane GAFFET** et **Stéphane OPERTO**
préparée au laboratoire Géoazur, Sophia-Antipolis

soutenue le 18 Avril 2014 devant le Jury

Hervé CHAURIS	Professeur MINES ParisTech	Rapporteur
Stéphane GAFFET	Directeur de Recherche CNRS	Directeur de Thèse
Josselin GARNIER	Professeur Paris VII	Examinateur
Ludovic MÉTIVIER	Chargé de Recherche CNRS	Invité
Guust NOLET	Professeur Université Nice-Sophia Antipolis	Examinateur
Stéphane OPERTO	Chargé de Recherche CNRS	Co-directeur de Thèse
René-Edouard PLESSIX	Chercheur Sénior SHELL	Rapporteur
Jean VIRIEUX	Professeur Université Joseph Fourier	Examinateur

Résumé

Les méthodes d'imagerie sismique ont pour objet d'estimer des propriétés du sous-sol à partir des ondes élastiques enregistrées au voisinnage de la surface de la Terre. Parmi les méthodes sismiques, l'inversion des formes d'ondes complètes (FWI) est un problème d'optimisation numérique local visant à exploiter la totalité de l'information contenue dans les formes d'onde. Bien que la FWI fasse aujoud'hui partie intégrante des séquences de traitement en exploration sismique pétrolière, cette technologie pose encore de nombreux défis, en particulier en termes de résolution, de coût calcul et d'optimisation non-linéaire. Cette thèse explore plusieurs méthodes visant à réduire le coût de calcul et à améliorer les techniques de régularisation interfacées avec le problème d'optimisation.

Actuellement, le principal obstacle à la mise en oeuvre de la FWI élastique en trois dimensions sur des cas d'étude réalistes réside dans le coût de calcul associé aux taches de modélisation sismique. Pour surmonter cette difficulté, je propose deux contributions. Tout d'abord, je propose de calculer le gradient de la fonctionnelle avec la méthode de l'état adjoint à partir d'une forme symétrisée des équations de l'elastodynamique formulées sous forme d'un système du premier ordre en vitesse-contrainte. Cette formulation auto-adjointe des équations de l'élastodynamique permet de calculer les champs incidents et adjoints intervenant dans l'expression du gradient avec un seul opérateur de modélisation numérique. Le gradient ainsi calculé facilite également l'interfaçage de plusieurs outils de modélisation avec l'algorithme d'inversion.

Le coût de calcul de la FWI est proportionnel au nombre de sources de l'expérience sismique ($> 10^3$ en 3D). Une approche possible pour réduire ce coût consiste à effectuer la modélisation sismique pour une combinaison linéaire des sources plutot que de traiter chaque source indépendamment. Les facteurs aléatoires utilisés dans cette combinaison linéaire ont pour fonction de réduire au cours des itérations les artefacts créés par cet assemblage artificiel. L'encodage de sources a été abondamment utilisé avec des méthodes d'optimisation du premier ordre tels que des algorithmes de plus grande pente ou de gradients conjugués. J'explore dans cette thèse dans quelle mesure son utilisation dans des algorithmes d'optimisation du second-odre de quasi-Newton et de Newton tronqué permettait de réduire encore le coût de la FWI. J'ai évalué cette méthodologie à l'aide de cas d'études synthétiques et réels 2D en utilisant une méthode de modélisation sismique en domaine fréquentiel fondée sur l'utilisation de solveurs directs. Néanmoins, les résultats de mon étude peuvent s'extrapoler à des modélisations sismiques 3D en domaine fréquentiel effectuées avec des solveurs itératifs.

Le problème d'optimisation associé à la FWI est mal posé au sens où les données ne permettent pas de déterminer un modèle du sous-sol unique, nécéssitant ainsi d'ajouter des contraintes de régularisation à la fonctionnelle à minimiser. Le type de régularisation généralement utilisée en imagerie sismique consiste à chercher le modèle le plus lisse possible permettant d'expliquer les données. Je montre ici comment une régularisation fondée sur la variation totale du modèle fournissait une représentation adéquate des modèles du sous-sol en préservant le caractère discontinu des interfaces lithologiques. Pour améliorer la qualité des images du sous-sol, je propose également un algorithme de débruitage fondé sur une variation totale locale au sein duquel j'incorpore l'information structurale fournie par une image migrée pour préserver les structures de faible dimension dans l'image.

Mots clés : inversion de forme d'onde, FWI, encodages des sources, optimisation numérique non linéaire, gradient stochastique, régularisation, variation totale, débruitage, méthodes de Newton.

Abstract

Seismic imaging methods allow to reconstruct the earth's subsurface parameters based on partial measurements of elastic waves at, or near, the surface. Full waveform inversion (FWI) is a numerical optimization problem that uses the whole waveform information of all arrivals to update the subsurface parameters that govern seismic wave propagation. FWI has shown to provide high quality images and now constitutes a production imaging tool in reservoir exploration. However, FWI still faces many challenges, specially in the topics of the image resolution, computational cost and non-linear optimization. This thesis concerns some aspects to reduce the computational cost and the use of regularization techniques in the optimization problem.

Currently, the main limitation to perform 3D elastic full waveform inversion on a production level is the computational cost it represents. With this in mind, we provide two contributions. First, we develop a self adjoint formulation of the isotropic first order velocity-stress elastic equations that allow to implement only one forward modeling operator in the gradient computation. Second, solving the forward problem is the most computationally expensive part of FWI, and its cost is proportional to the number of sources ($> 10^3$ in 3D). To gain efficiency, instead of solving the wave equation for each source, it is possible to solve the wave equation for a linear combination of the sources. This is known as source encoding, and has been widely used when combined with steepest descent or conjugate gradient algorithms in the optimization process. With the purpose of reducing even more the computational cost, we combine Newton and quasi-Newton optimization methods with source encoding techniques. We implement this in a 2D frequency domain acoustic modeling engine based on frequency, but the results are easily extendible to the 3D frequency scenario. Our synthetic numerical tests were carried out in the BP-2004 salt model which is a realistic configuration inspired by the Gulf of Mexico, and the real data we used is an ocean bottom cable dataset recorded from the Valhall field in the North Sea. We find that the lowest computational cost needed to attain a predefined misfit value is provided by l-BFGS, with periodic restarts. However, with noisy data, the most accurate and robust direction of descent is provided by Newton methods, with and without source encoding.

The optimization process requires regularization constraints because the model is not entirely constrained by the data, and more than one model may fit the observed data equally well. We see that the total variation of the model as a regularization term provides and adequate description of earth models. To improve the quality of the images of the earth parameters, we propose a local total variation denoising algorithm based on the incorporation of the information provided by a migrated image.

Key words : full waveform inversion, FWI, source encoding, optimization, stochastic optimization, non-linear optimization, regularization, total variation, TV denoising, Newton methods

Remerciements

Je ne pourrais pas résumer toute la gratitude que j'ai pour toutes les personnes qui m'ont accueilli, appris, guidé, et partagé des moments avec moi pendant ces années de thèse.

Je tiens tout d'abord à remercier profondément Stéphane Operto, qui m'as guidé dans mes travaux de recherche. J'apprécie énormément les temps qu'il a dédié pour me transmettre toutes ses connaissances, sa disponibilité pour discuter, et son implication dans mon travail. Je le remercie aussi d'avoir rendu ce manuscrit beaucoup plus lisible et de m'avoir aidé pour la rédaction en français. Mes remerciements vont également à Stéphane Gaffet avec qui j'ai commencé à travailler en géophysique pour la première fois lors du master, et qui m'a encouragé à faire une thèse. Je voudrais lui exprimer tout ma reconnaissance pour sa sympathie et ses conseils.

Je voudrais remercier Hervé Chauris, professeur de MINES ParisTech, et René-Eduard Plessix, chercheur pincipal à Shell, pour avoir accepté d'être les rapporteurs de ce manuscrit. Mes remerciements vont également à Josselin Garneier, Professeur à Paris VII, Ludovic Metivier, chargé de recherche CNRS, Guust Nolet, Professeur à Géoazur, et Jean Virieux, professeur à l'Université Joseph Fourier, pour m'avoir accordé l'honneur de faire partie du jury de ma thèse. Je voudrais exprimer toute ma gratitude à Guust et Jean, avec qui cela a toujours était un vrai bonheur pour moi de partager une discussion.

J'ai eu la chance de faire cette thèse dans un cadre très agréable grâce à l'ensemble du groupe SEISCOPE. Je remercie Ludovic Metivier avec qui j'ai eu le plaisir de travailler pendant ma thèse, Romain Brossier qui m'as toujours aidé à sortir des moments de confusion et Alessandra Ribodetti pour ses explications et sa sympathie en tout moment. A tous les membres de Seiscope actuellement à Geoazur ou déjà partis et avec qui j'ai partagé beaucoup des moments dans les congrès, les conférences et réunions : Yaser Gholami, Vincent Etienne, Vincent Prieux, Stephen Beller, Laure Combe et Vadim Montellier. Un remerciement aussi aux thèsards de l'équipe Seiscope à Grenoble avec qui j'ai pu partager des moments : Amir Asnaashari, Isabella Massoni, François Lavoue.

J'ai eu aussi l'occasion de partager des discussion et des réunions avec l'équipe Globalseis. Un grand merci à Laurent Stehly, Dylan Mikesell et Jean Charlety qui ont aussi contribué à mes travaux de recherche.

Je tiens à remercier aussi les professeurs du master Mathmods pour leur sympathie et investissement dans notre formation. Je pense surtout à Bruno Rubino, Pierre- Emmanuel Jabin, Victorita Dolean, Chiara Simene et Etienne Tanré.

Chaque jour de ces années de thèse ont été spécialement très agréables grâce à tous les thésards avec qui on partage les choses qui marchent et qui ne marchent pas, les débats à midi, les hauts et les bas. Merci à Clément, Yaser, Alain, Dlung, Nestor, Quentin, Juan Carlos, Maëlle, Victorien, Stephen, Benoit, Flore, Marianne, Imane et Alice et Maurin (même si vous n'êtes pas des thésards). Je tiens à remercier spécialement mes collègues avec qui j'ai partagé le bureau : Clément (mon prof de culture française), Alain (c'est trop cool de partager les astuces sur Matlab, linux, bash, etc) et Dlung (ta patience est contagieuse). J'ai appris beaucoup des choses grâce à vous : du stress drop, des failles, des satellites et orbites, jusqu'à la vie française. Keep calm and ...

Un grand merci à toutes les personnes qui font de la vie au laboratoire très agréable et qui sont toujours prêt a donner un coup de main. Je pense surtout à Jenny Trévisan, Jelena Giannetti, Caroline Ramel, Julien Olivier, Lionel Maurino, Valerie Mercier, Arielle Willm (OCA), Laure Miniussi (OCA) et Alain Miniussi (OCA).

Il y a une personne avec qui j'ai commencé cette aventure de la thèse et finalement on est arrivé à la fin. Camilo, merci d'être ma constante.

Mommy, Daddy, Juli, Pao : Gracias infinitas por mantenerme cerca a ustedes a pesar que estamos tan lejos.

Merci du fond du cœur à toutes les personnes à Géoazur avec qui j'ai partagé d'inoubliables moments ces dernières années. C'était un vrai bonheur d'apprendre une langue, une culture et un métier parmi vous.

Clara

"On se dit que le bon temps passe finalement comme une étoile filante..."

Resumé Etendu

Connaître la composition de l'intérieur de la Terre est d'un intérêt particulier pour de nombreuses applications académiques et industrielles. Cette thèse porte sur l'étude de la méthode d'inversion des formes d'ondes complètes (**FWI**)

$$\min_m \phi_0 = \min_m \|Pu(m) - d\|_2^2, \tag{1}$$

où u est le champ d'onde modélisé, d représente les données enregistrées et m sont les paramètres décrivant les propriétés du sous-sol que l'on cherche à imager. L'opérateur de projection P qui extrait la valeur du champ d'onde modélisé aux positions de l'espace où les données ont été enregistrées. Le champ d'onde modélisé est solution de l'équation d'onde. Par exemple, dans le cas de l'approximation acoustique, u est solution de

$$\left(\nabla^2 - \frac{1}{v^2}\frac{\partial^2}{\partial t^2}\right) u(x,t) = s(x,t) \qquad (x,t) \in \Omega \times [0,\infty), \tag{2}$$

où $v(x) = 1/\sqrt{\rho(x)\kappa(x)}$, κ est le module d'incompressibilité, ρ est al densité, v est la vitesse de propagation des ondes P. Ici, les paramètres m représentent aux différentes positions de l'espace une seule propriété physique, la lenteur au carré: $m = \{1/v^2(x)\}$. La résolution de l'équation (18) est dénommée problème direct, et la résolution de l'équation (17) est le problème inverse. La FWI nécessite la résolution de ces deux problèmes de manière alternée. Le problème direct fournit u pour un modèle initial m donné, et ce champ d'onde u est utilisé dans un deuxième temps par le problème inverse pour mettre à jour le modèle du sous-sol. En d'autres termes, le problème direct consiste à résoudre une équation aux dérivés partielles (EDP), et le problème inverse estime les paramètres contenus dans les coefficients de cette EDP à partir de ses solutions. Ce processus est répété de manière itérative jusqu'à ce que les valeurs du champ d'onde u propagé dans le modèle m aux positions des capteurs coïncident avec les données enregistrées d.

Lorsque la totalité du champ d'onde est injectée dans le processus d'optimisation, la FWI a potentiellement la capacité de fournir des modèles du sous-sol plus réalistes en termes de résolution et de caractérisation physique que les autres méthodes d'imagerie. Ses fondements théoriques ont été établis en géophysique dans les années quatre vingt (Lailly, 1983; Tarantola, 1984a), et depuis la compréhension de ses potentialités et de ses limites s'est affinée au gré des applications réalisées grâce aux moyens modernes d'acquisition et de calcul haute performance (Virieux and Operto, 2009). Les challenges méthodologiques à surmonter pour une mise en œuvre pertinente de la FWI ont néanmoins été clairement posés dès les premières études publiées sur le sujet. Par exemple, les premières analyses sur des cas synthétiques indiquent que le coût calcul de la FWI constitue un frein important à son

application à des cas d'étude réels (Gauthier et al., 1986; Crase et al., 1990; Luo and Schuster, 1991; Crase et al., 1992). Plus fondamentalement, contrairement à la plupart des autres méthodes d'imagerie, la FWI est un problème inverse non linéaire et les premiers tests numériques montrent que le processus d'optimisation non linéaire échoue souvent car la fonctionnelle à minimiser ϕ_0 est fortement non convexe faisant converger l'inversion vers un minimum local inacceptable quand le modèle initial est trop éloigné du modèle réel (Gauthier et al., 1986; Mora, 1989; Luo and Schuster, 1991). Bien que la FWI soit supposée reconstruire un large spectre de longueurs d'onde du modèle à partir des champs d'ondes réfléchis et transmis, les premiers essais (Devaney, 1984; Wu and Toksöz, 1987; Mora, 1988, 1989) ont uniquement fourni une image de la réflectivité c'est-à-dire des courtes longueurs d'onde tel que l'aurait fourni une méthode de migration. Cela motiva plusieurs analyses de résolution de la FWI pour comprendre quelle partie du spectre des nombres d'onde dans l'espace des modèles pouvait être reconstruit pour une géométrie d'acquisition et pour une bande passante de source (Jannane et al., 1989; Mora, 1989). Ces difficultés furent identifiées lors des investigations pionnières de Gauthier et al. (1986); Wu and Toksöz (1987); Mora (1988, 1989); Luo and Schuster (1991), et restent aujourd'hui des sujets de recherche d'actualité en FWI. Dans cette thèse, j'aborderai certains aspects de la FWI portant sur la réduction de son coût calculatoire et de sa non linéarité.

Bien que la puissance de calcul soit devenue suffisante dans les années 80 pour résoudre l'équation d'onde pour des modèles du sous-sol d'hétérogénéité arbitraire, par des méthodes de différences finies dans le domaine temps-espace notamment, les applications de la FWI sur des cas d'étude réels restaient hors de portée. Un premier pas ayant permis de réduire considérablement le coût de la FWI a résidé dans l'utilisation de la méthode de l'état adjoint pour calculer le gradient de la fonctionnelle (Lailly, 1983; Tarantola, 1984a). Une deuxième étape vers la réduction du coût de la FWI a été franchie en formulant la FWI dans le domaine fréquentiel. Dans les années 90s, le formalisme de la FWI, tel qu'élaboré par Tarantola (1984a), a été transformé dans le domaine fréquentiel pour les équations d'ondes acoustique et élastique (Pratt and Worthington, 1990; Pratt, 1990; Pratt and Shipp, 1999). L'équation d'onde dans le domaine fréquentiel est une forme généralisée de l'équation d'Helmholtz pouvant s'écrire sous forme matricielle comme

$$A(m,\omega)u(s,m,\omega) = s(\omega), \tag{3}$$

où les coefficients de $A(m,\omega)$, matrice résultant de la discrétisation de l'opérateur de l'équation d'onde en domaine fréquentiel. Ce système linéaire peut être résolu par des méthodes directes ou itératives. Les méthodes directes reposent sur des techniques d'élimination de Gauss telles que les approches fondées sur des décompositions triangulaires supérieure/inférieure de la matrice ($A = LU$). La factorisation LU de la matrice dans l'équation (23) dépend uniquement de m et de la fréquence ω. Par conséquent, pour chaque itération non-linéaire du problème inverse, une seule factorisation LU est effectuée par fréquence pour calculer le gradient. Les champs d'onde calculés pour chacune des sources sont efficacement calculés à partir des facteurs LU par substitutions directe et inverse. Les limites des approches fréquentielles fondées sur des solveurs directs résultent du calcul et du stockage des facteurs LU. Pour des applications élastiques 3D, l'approche par solveur direct semble aujourd'hui inaccessible en raison de la nature vectorielle de l'équation d'onde élastique nécessitant le calcul d'au moins trois composantes de vitesse particulaires et des faibles longueurs d'ondes propagées en relation avec la vitesse de propagation des ondes de cisaillement. Dans cette configuration, seules les modélisations en domaine fréquentiel avec des solveurs itératifs ou des modélisations en domaine temporel semblent possibles.

Une réduction du coût calcul est envisageable si le nombre de problèmes directs est diminué. Comme pour des approches fréquentielles fondées sur des solveurs itératifs ou des modélisations temporelles, le coût des problèmes directs est proportionnel au nombre de sources (équations (18) et (23)), la manière la plus naturelle de réduire le coût est de diminuer le nombre de sources. Une méthode courant d'accélération, appelée encodage des sources, consiste à effectuer une acquisition sismique conventionnelle (en émettant une source à la fois). Néanmoins, au lieu de résoudre le problème direct, eqrefech:chintroforprob ou (23), pour chaque source, des super-sources sont formées par combinai-

son linéaire des sources pondérées par des facteurs aléatoires (Romero et al., 2000; Neelamani et al., 2008; Krebs et al., 2009; Ben Hadj Ali et al., 2011; Schuster et al., 2011; van Leeuwen et al., 2011; Huang and Schuster, 2012). Pour encoder les données de la sorte, il est nécessaire que les sources soient enregistrées par le même ensemble de récepteurs, désignant ainsi les acquisitions d'extension fixe (fixed-spread en Anglais) comme les acquisitions de fond de mer. Le volume de données généré par N_s sources et enregistré par N_r récepteurs est compressé en une seule collection de traces à tir commun.

Contributions de cette thèse

Généralement, le gradient de la fonction coût de la FWI est implémenté avec la méthode de l'état adjoint à partir de l'équation d'onde du second ordre auto-adjointe. En revanche, l'équation d'onde formulée sous forme d'un système hyperbolique d'ordre 1 en vitesse-contrainte ne conduit pas à une forme auto-adjointe ($A \neq A^\dagger$), nécessitant une implémentation dissociée de l'opérateur direct (équation d'état) et de l'opérateur adjoint. Ma contribution a consisté à développer un formalisme permettant d'exprimer les équations de l'élastodynamiques du premier ordre sous forme d'un opérateur auto-adjoint facilitant ainsi grandement l'implémentation du gradient. Ceci est implémenté via un changement de variable appliqué aux contraintes normales, fourni par la décomposition en vecteurs propre de la matrice de raideur. Ce changement de variable permet de reformuler l'équation d'onde sous une forme pseudo-conservative au sein de laquelle les paramètres du milieu sont regroupés dans une matrice diagonale factorisée aux dérivés temporelles du système hyperbolique (Castellanos et al., 2011).

Bien que les méthodes d'encodage de sources aient été beaucoup utilisées en FWI ou en migration par renversement temporel (Neelamani et al., 2008; Krebs et al., 2009; Ben Hadj Ali et al., 2011; Schuster et al., 2011; van Leeuwen et al., 2011; Huang and Schuster, 2012), elles ont principalement été combinées avec des algorithmes d'optimisation de gradient, reproduisant ainsi les principes des méthodes de gradient stochastiques (Robbins and Monro, 1951; Spall, 2003). Dans le but de réduire encore le coût des calculs et améliorer le facteur d'accélération fourni par les méthodes d'encodage j'ai combiné des algorithmes d'optimisation de quasi-Newton et de Newton avec de l'encodage de source. Dans le cas des données réelles du champ de Valhall, l'inversion avec différentes méthodes d'optimisation ne convergent pas vers le même minimum local. Ceci résulte probablement du fait que, au voisinage du modèle initial, la fonction coût contient plusieurs minimum locaux. L'estimation de l'accélération fournie par l'encodage de sources et les méthodes d'optimisation du second d'ordre est donc biaisée car la qualité des modèles finaux n'est pas identique. Je montre que, avec et sans encodage de sources, les méthodes de Newton suivent la direction de descente qui converge vers les modèles du sous sol pour lesquels la fonction coût a la valeur la plus faible. Pour une meilleure compréhension des artefacts d'interférence générés par l'encodage des sources, une estimation est donnée de la variance du gradient avec encodage de sources pour des données sans bruit à partir de trois types d'assemblages de sources. L'analyse de ces variances peut être utilisée pour concevoir des stratégies optimisées d'encodage de sources. Une partie des résultats présentés ont été soumis pour publication dans Castellanos et al. (2013).

Beaucoup de travaux ont été consacrés à l'exploration de nouvelles normes dans l'espace des données (Djikpéssé and Tarantola, 1999; Guitton and Symes, 2003; Ha et al., 2009; Pyun et al., 2009; Brossier et al., 2010), mais seulement récemment l'influence de la norme utilisée dans l'espace des modèles (la régularisation) a été analysée (Burstedde and Ghattas, 2009; Ramírez and Lewis, 2010; Anagaw and Sacchi, 2012; Guitton, 2012). Je compare dans cette thèse deux normes dans la régularisation, la norme l_2 et la variation totale (TV) à partir d'un test synthétique réaliste (le modèle BP-2004 salt) et les données d'OBC enregistrées sur le champ de Valhall. En l'absence de bruit, le modèle final obtenu avec la norme TV est considérablement meilleur. Pour le cas d'étude réel, les modèles de vitesse finaux obtenus avec les normes l_2 et la TV sont comparables. Néanmoins, la stratification induite par les couches géologiques est mieux restituée dans les modèles construits avec la régularisation par variation totale.

L'algorithme de débruitage de Rudin-Osher-Fatemi (ROF) (Rudin et al., 1992) élimine le bruit d'une image en minimisant la variation totale de l'image, tout en maintenant l'image débruitée aussi similaire que possible à l'image initiale. Néanmoins, les limites du débruitage par TV résident dans le fait qu'il peut éliminer la texture (les structures de petites dimensions) dans l'image. Cela a motivé un certain nombre de travaux dédiés à la conception d'algorithmes de débruitage local par TV qui préserve la texture de l'image (Bertalmio et al., 2003; Vese and Osher, 2003). J'applique un débruitage local par TV où j'incorpore une information a priori sur la réflectivité fournie par une image migrée. L'algorithme de débruitage local par TV n'a aucune action aux positions de l'espace où la migration a positionné des réflecteurs et débruite les autres parties de l'image. L'algorithme de débruitage local par TV remplit alors sa mission tout en préservant les principaux réflecteurs.

Contents

INTRODUCTION

1 INTRODUCTION 17

1.1 Full waveform inversion : the challenges of non-linear seismic imaging. 22
1.1.a *Resolution* 23
1.1.b *Computational Cost* 24
1.1.c *Non-linear optimization* 27

1.2 Introduction to this thesis and description of main results 31

2 INTRODUCTION (FR) 38

2.1 L'inversion des formes d'ondes complètes : le défi de l'imagerie sismique non linéaire. 43
2.1.a *Analyse de résolution* 44
2.1.b *Coût de calcul* 45
2.1.c *Optimisation non linéaire* 48

2.2 Présentation de la thèse et description des principaux résultats 54

Chapter 1 Fundamentals of inverse problem theory

0.3 Linear Inverse Problems 64

0.4 Non linear Inverse Problems 65

1 Ill Posed Inverse Problems 66

2 Optimization 70

2.1 Line Search 70

2.2 Search Directions 71
2.2.a *Steepest Descent* 72
2.2.b *Linear conjugate gradient* 72
2.2.c *Newton methods* 73
2.2.d *BFGS* 74
2.2.e *Limited memory BFGS* 75
2.2.f *Errors with the approximate Hessian, and restart procedures* 76

2.3 Preconditioner . . . 78

Chapter 2 Full Waveform Inversion

1 Seismic imaging with full waveform inversion . . . **79**

1.1 Full waveform inversion . . . 79

1.2 The gradient . . . 81

1.2.a *Preconditioner* . . . *83*

1.2.b *Physical interpretation of the gradient* . . . *84*

1.3 The Hessian . . . 86

1.3.a *Physical interpretation* . . . *90*

1.4 Solving FWI numerically in the frequency or time domain . . . 94

1.5 Image resolution analysis . . . 102

Chapter 3 Speed-up of FWI with source encoding

1 Introduction . . . **112**

2 Speed-up FWI : state of the art . . . **117**

3 Method . . . **119**

3.1 Full waveform inversion problem . . . 119

3.2 Second-order optimization methods . . . 120

3.3 Preconditioner . . . 122

4 Source encoding . . . **123**

4.1 Optimization algorithms with source encoding . . . 123

5 Numerical Examples . . . **124**

5.1 Synthetic example . . . 127

5.1.a *Illustration of the problem : Convergence rates of stochastic versus deterministic algorithms* . . . *127*

5.1.b *Synthetic data without noise* . . . *128*

5.1.c *Synthetic data with noise* . . . *130*

5.1.d *Stochastic gradient* . . . *132*

5.1.e *Summary of the BP-2004 salt experiment* . . . *133*

5.2 Real data example . . . 133

6 Estimation of the variance with source encoding, without noise. . . . **136**

7 Conclusion . . . **141**

8 Tables . . . **142**

9 Figures . . . **144**

Chapter 4 Regularization techniques for FWI

1 Regularization . . . 163

1.1 Tikhonov regularization . . . 164
1.1.a *Choice of λ for the known noise level case* . . . 165
1.1.b *Choice of λ for the unknown noise level case* . . . 166
1.1.c *Example 1: No regularization and overfitting the data* . . . 166
1.1.d *Example 2: Regularization and use of the L-curve* . . . 167
1.1.e *Norm for the regularization term* . . . 167
1.1.f *Norm for the data* . . . 169

1.2 Total Variation Regularization . . . 170

1.3 TV denoising algorithms for seismic imaging : A modified ROF method. . . . 171

1.4 Numerical Implementation of regularization and denoising (l_2 and TV norms) . . . 173
1.4.a *TV Regularization : numerical examples* . . . 174
1.4.b *Denoising : numerical examples* . . . 186
1.4.c *Conclusions* . . . 187

2 The model null space and the spectral content of the data . . . 194

2.1 Frequency response . . . 194

2.2 Spectral content of the data in the BP-2004 Salt Model . . . 202

Conclusions and perspectives

1 Conclusions and perspectives . . . 210

2 Conclusions et perspectives (FR) . . . 212

APPENDICES . . . 209

1 Imaging conditions for other seismic imaging methods . . . 215

1.1 Time Reversal Mirror Imaging Condition . . . 215

1.2 Imaging with noise cross correlations . . . 220

1.3 Imaging condition with adjoint methods . . . 223
1.3.a *Reverse time migration imaging condition* . . . 224

1.4 Time reversal imaging conditions : conclusions . . . 224

2 Travel time tomography . . . 225

3 The gradient of the misfit function : computation with the adjoint state method . . . 228

3.1 The Adjoint State Method . . . 228
3.1.a *Adjoint operators* . . . 229
3.1.b *Directional derivatives* . . . 229
3.1.c *The Lagrangian and the adjoint state equations* . . . 230

4 The gradient computation for the velocity-stress elastodynamic wave equations without attenuation in conservative form, with the adjoint-state method 232

4.1 Introduction . . . 232

4.2 Forward Equations . . . 232
4.2.a *Lagrangian and adjoint variables in the time domain* . . . 234

4.3 The Adjoint State Equations . . . 236
4.4 The symmetric Pseudo-Conservative Form . . . 236
4.4.a *Adjoint State Equations for the Conservative Formulation* . . . *239*
4.4.b *State equations* . . . *240*
4.4.c *Adjoint state equations* . . . *240*
4.4.d *Gradient of the misfit function* . . . *241*
4.4.e *Algorithm* . . . *242*

5 THE HESSIAN : COMPUTATION WITH THE ADJOINT STATE METHOD . . . **242**

6 SOURCE ENCODING WITH DIRAC NOTATION . . . **246**
6.1 Misfit function . . . 246
6.2 Gradient . . . 248

BIBLIOGRAPHY . . . 250

INTRODUCTION

1 INTRODUCTION

In several scientific and industrial applications we are interested in knowing the composition of the earth's subsurface. For cost and feasibility reasons, it is not possible to directly observe the subsurface composition (as one would do, for example by drilling). This has driven the development of much cheaper and easily implementable approaches, generically known as seismic imaging techniques. They consist of different algorithms allowing to reconstruct the subsurface structure indirectly from physical measurements at the surface. The data can correspond to the recordings of seismic waves, electromagnetic waves, gravity anomalies, amongst others. The quality of the reconstruction of the earth's composition depends on the ***imaging technique*** used and ultimately on the ***measured data*** that are available. The purpose is therefore to reconstruct the subsurface parameters based on the measured data as illustrated in Figure 15. This thesis will focus on imaging techniques based on seismic recordings.

Data acquisition : transmission and reflection energy

On a global scale, earthquakes generate waves that travel through the earth's interior and the particle velocity of the earth's surface is measured on arrays of receivers around the earth as shown schematically in Figure 16a. Most of data used on global scale imaging is known as transmission data because the earthquakes (sources) generate waves that travel through the portion of the earth that needs to be imaged, and the ***transmitted energy*** is recorded on the other side [1]. In reservoir exploration, used to find natural resources (hydrocarbons, gas, water), the perturbations in a delimited zone are created by controlled explosions either on land or marine environments. An example of a streamer acquisition is shown in Figure 16b. The sources are controlled explosions generated by moving boats. The sources and receivers move together and allow for the exploration of a wider area. The interest is to image the subsurface of the earth below the ocean bottom. When the sources and receivers are placed the surface of the earth (even if the boats are on the surface of the water), this is referred to as a surface acquisition. Other configurations, such as cross-hole acquisitions, schematically consist of at least two wells drilled

1. This does not mean that there are no reflected waves in the recorded data. Reflected waves are created between the surface and the shallow structures.

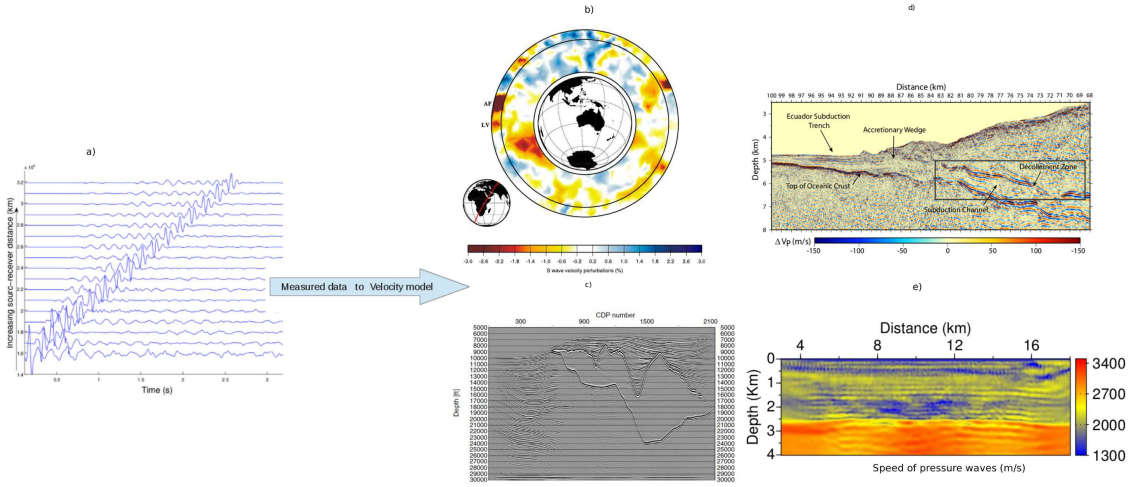


Figure 1: Seismic imaging methods reconstruct the earth's subsurface parameters from measured data of the earth's displacement at the surface. a) Measurements at different sensor positions (y-axis), with respect to time (x-axis), called seismograms. The distance from the source to the receiver is increasing in the y-axis. b) Global image of the shear wave velocity perturbation (δv_s) with respect to a background model using travel time tomography (Montelli et al., 2004). c) Exploration scale image of the position of the reflectors in a salt dome (Liu et al., 2011), using migration techniques. d) Regional pressure wave velocity perturbation (δv_p) with respect to a background model using ray + Born inversion (Ribodetti et al., 2011). e) Exploration scale image of the pressure wave velocity v_p using full waveform inversion of the Valhall oil field (this thesis).

into the earth with the target area to be imaged in the middle. The sources are generated from one well and the wave measurements are recorded on the opposite well. Surface acquisitions and cross-hole acquisitions do not provide the same type of data. The cross-hole acquisition measures principally transmitted energy (similar to the data recorded with teleseismic waves), because there are no receivers on the same side of the source to measured the reflected energy. With a surface acquisition the situation is the opposite. With the sources and receivers being placed on the same side of the target, the main information being recorded is ***reflected energy***. Nonetheless, if the surface acquisition has a long offset coverage (large distance between sources and receivers) the data may contain transmitted energy as well associated with diving waves and super-critical reflections. This may be the case for fixed-spread acquisition (ocean bottom seismic acquisition). The acquisition geometry will thus determine if the recorded data contains more transmitted or reflected energy.

A seismic record (receiver gather) collected during an OBC (ocean bottom cable) survey in the Valhall oil field in the North Sea is shown in Figure 17a. One receiver, whose position is labelled by the 0 km offset (horizontal axis), recorded a collection of seismograms generated for each source along a profile with source-receiver offset ranging from $-12km$ to $4km$. The vertical axis is the time evolving since the source was triggered. In Figure 17b shows the initial physical 2D model, where the colors represent the magnitude of the propagation speed of the pressure waves (v_p) in the earth. The main paths that the waves follow when the source is triggered are indicated by the rays. The white paths are refracted (transmitted) waves, the red paths denote the reflected waves by the red boundary and the blue paths show the waves reflected by the blue boundary. In this example, there are no transmitted waves below $\approx 2km$. Figure 17c shows the wavefield measured for the shot placed approximately at $x = 2.5km$. The first arrival on

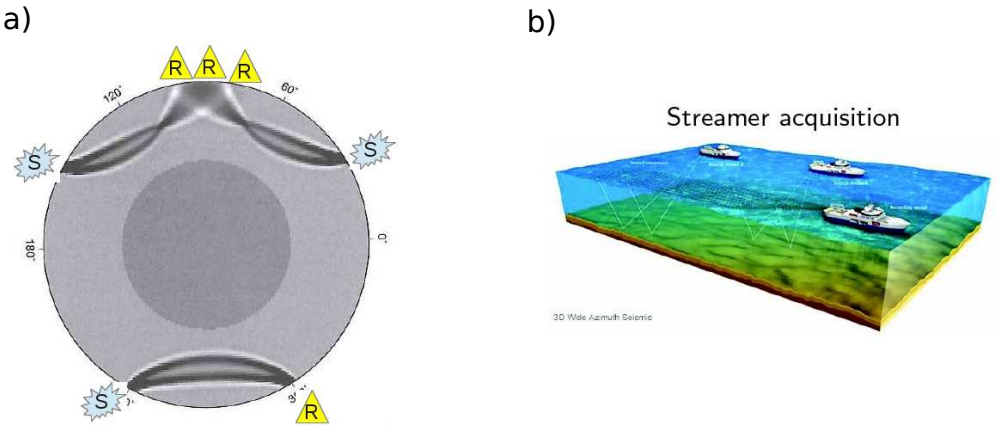


Figure 2: a) Figure adapted from (Nolet, 2008). The earthquakes (S) generate teleseismic waves that travel through the earth's interior and are recorded at seismic receiver stations (R). b) Figure showing a marine streamer acquisition for exploration seismology.

the seismogram would correspond to a refracted wave (white path in Figure 17b) and the latter arrivals are the reflected waves on the blue and red interfaces. With Figure 17a, it is possible to broadly see that the travel times of the transmitted waves $t = o/v$ (o is offset and v is speed) provide information on the slope in the gather, which gives information on the average wave-speed v along the propagated paths. The travel-times of the reflected waves, $t = \sqrt{o^2/4 + z^2}/v$, depends both on the average wave-speed of the medium and the depth of the reflectors z.

Imaging techniques

The acquisition geometry determines the content of the data, and different imaging techniques use different parts of the data in the seismograms. In global scale imaging, the data is primarily transmission data and the information that is used in tomography techniques is usually the *arrival time of the direct wave* (pressure) because the first arrival contains the main model properties related to the transmission of the waves. Travel time tomography assumes a known background model m_0, and small perturbations to the model δm are searched such that the numerical propagation in $m_0 + \delta m$ will match the measured arrival time of the first wave. An example of global model perturbations δv_s found to fit the arrival times of the shear waves is shown in Figure 15b. When imaging on a regional scale with a surface acquisition, the data is primarily reflection data. Migration techniques exploit only reflected data. The seismograms are preprocessed and the direct and refracted (transmitted) waves are removed from the seismograms, such that *only the (single) reflected data* are used to construct the image. Migration techniques produce images as the one shown in Figure 15c, showing the position of the interfaces where the energy is reflected (known as reflectors). Using only purely reflected data, it is also possible to find the position of the reflectors and the magnitude of the reflection perturbations δm with respect to a background model. Figure 15d shows an example of δv_p using and imaging technique known as ray-Born migration (Lambaré et al., 1992). On both global and regional scales, it is also possible to use the *information of the entire wavefield* consisting of transmitted and reflected energy. However, as has been explained, in global tomography the natural acquisition system is such that the transmitted energy will be predominant and on regional scales with surface acquisitions the reflected energy will be predominant. An example of a regional

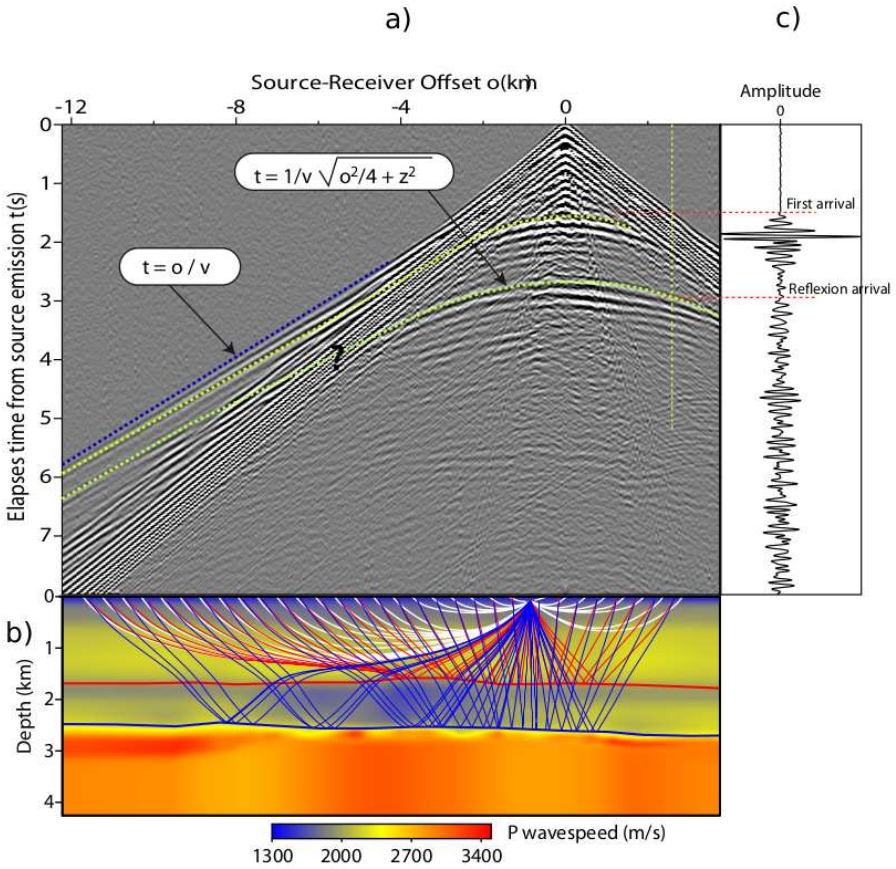


Figure 3: a) An example of a real ocean bottom cable receiver shot gather in the Valhall oil field . b) The initial guess of the 2D model of one model parameter, showing some wave paths. c) The wavefield as a function of time measured at one receiver position (seismogram).

scale image of a model for the wave-speed propagation of the pressure waves using the whole waveform is shown in Figure 15e. The imaging technique used is full waveform inversion (Lailly, 1983; Tarantola, 1984a). Note that there are difference between all the images. If a Fourier transform is applied to these images, the global scale tomographic image would have principally low wavenumbers and the migration images would show high wavenumber content.

In this work, we focus on studying **Full Waveform Inversion** (**FWI**). This imaging technique aims to take into account in the inversion process all the information present in the waveform, that is both transmitted and reflected energy (Lailly, 1983; Tarantola, 1984a; Virieux and Operto, 2010). More precisely, it is a non-linear inverse problem, aimed to minimize the difference between the observed and computed data,

$$\min_m \phi_0 = \min_m \|Pu(m) - d\|_2^2, \tag{4}$$

where u is the computed wavefield, d is the recorded data and m is the function of model parameters. P is a projection operator of the computed wavefield on the same positions where the recorded wavefield is defined. The computed wavefield is the solution of the wave equation. For example, assuming a propagation of only acoustic waves, u is the solution to

$$\left(\nabla^2 - \frac{1}{v^2}\frac{\partial^2}{\partial t^2}\right) u(x,t) = s(x,t) \qquad (x,t) \in \Omega \times [0,\infty), \tag{5}$$

where $v(x) = 1/\sqrt{\rho(x)\kappa(x)}$, κ is the compressibility, ρ is the density, v is the wave velocity. Here, the model m consists of one physical parameters : $m = \{1/v^2(x)\}$. Solving (18) is known as the forward problem, and (17) is known as the inverse problem. Using FWI as an imaging technique is therefore an alternating process. The forward problem is solved to find u for a given model m, and then this wavefield u is used in the inverse problem to determine to model update. In other words, the direct problem consists in solving a partial differential equation (PDE), and the inverse problem estimates the parameters of the PDE. This process is repeated iteratively until the wavefield u propagated in the model m coincides with the observed data d. This work flow is schematically represented in Figure 18.

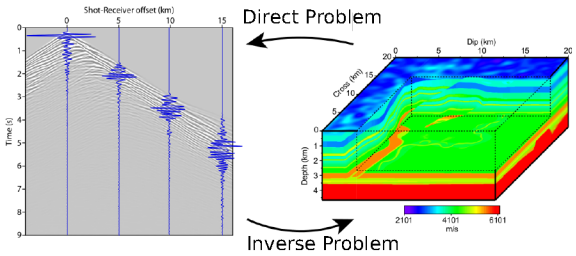


Figure 4: FWI iterates the solution of a direct problem (18) and an inverse problem (17). The direct problem allows to pass from the model to the data space, by using the current model m to solve the equation and find the wavefield u. The inverse problem allows to go from the data space to the model space by using the wavefield u to determine the model that minimizes a misfit ϕ_0.

As the whole waveform information is used, FWI has the potential to be very accurate compared to other imaging techniques. Its theoretical fundamentals in the time domain were introduced in geophysics in the 1980's (Lailly, 1983; Tarantola, 1984a), and since then the comprehension has evolved and been used with success (Virieux and Operto, 2009). At the same time, FWI faces some important challenges. For example, early synthetic experimental results indicated that the ***computational cost*** of FWI represented a true limitation if one wished to apply it to real data (Gauthier et al., 1986; Crase et al., 1990; Luo and Schuster, 1991; Crase et al., 1992). Furthermore, unlike previous imaging techniques, FWI is a non-linear inverse problem and early numerical tests revealed that the ***non-linear optimization*** process often failed because ϕ_0 is not convex, by converging to unsatisfactory local minima when the initial model was far from the true model (Gauthier et al., 1986; Mora, 1989; Luo and Schuster, 1991). Although FWI is expected to update the transmission and the reflectivity information of the model, first attempts (Devaney, 1984; Wu and Toksöz, 1987; Mora, 1988, 1989) only provided the reflectivity and only resolved the high vertical wavenumbers of the model (similar to migration techniques). This prompted several authors to analyze the ***resolution*** power of FWI, and explore which part of the image wavenumber spectrum can be obtained in an inversion for a given acquisition geometry and source bandwidth (Jannane et al., 1989; Mora, 1989). Other setbacks appeared in multi-parameter inversion. Results showed that the resolution of the reconstructed parameters varies from one parameter to the next depending on the local aperture illumination resulting from the acquisition geometry, the cross-talk between parameters (interaction between parameters) and relative weight of the sensibility of different parameters in the data (Tarantola et al., 1984; Mora, 1987)[2]. These disadvantages that were pointed out early in time (Gauthier et al., 1986; Wu and Toksöz, 1987; Mora, 1988, 1989; Luo and Schuster, 1991), currently remain as challenges and subjects of research in FWI.

2. A parametrization is understood as a set of independent parameters that fully describe the model space.

In this PhD thesis we will address some aspects related to the computation costs and non-linearity questions. But, before presenting in more detail the scope of this work, let us briefly go over some details over each of the mentioned points.

1.1 Full waveform inversion : the challenges of non-linear seismic imaging.

Imaging condition

The *imaging condition* is the operator acting from the data space to the model space, that indicates the model parameters that have a sensitivity on the computed wavefield. Therefore, the imaging condition will reveal all the model parameters that, when subject to small variations, generate variations in the data. Consider a reflection wave arrival in the measured data d at time $t = T$, that is not present in the computed wavefield u. The purpose of the imaging condition is to determine where in the image space the model should be modified so that projected wavefield u at the receiver position will also contain the reflection arrival. For one source receiver pair (S-R), the only restriction is that the location x of the diffracting point in the subsurface should satisfy

$$t_{S,x} + t_{x,R} = T. \tag{6}$$

For an homogeneous background model of velocity v, all the points x that satisfy (19) form an *ellipse with focal points at the source and receiver positions*, like that shown in Figure 19a. This can be obtained by explicitly writing the time for the wave to travel from S to x, $t_{S,x}$, and time to travel from R to x, $t_{x,R}$, where D is the distance between the source and the receiver,

$$\begin{aligned} t_{S,x} &= \sqrt{(x/v)^2 + (z/v)^2} \\ t_{x,R} &= \sqrt{(D-x)/v)^2 + (z/v)^2}. \end{aligned}$$

Replacing this back into (19), leads to the following equation of an ellipse,

$$\sqrt{x^2 + z^2} + \sqrt{(D-x)^2 + z^2} = vT,$$

with focal points $f = 0, D$, major axis $a = vT/2 + D/2$ and minor axis $b = \sqrt{(vT)^2 + D^2}/2$.

Summarizing, for one source-receiver pair and a reflection arrival at time T, any point in the ellipse could correspond to the true position of the diffracting point because they all satisfy (19). In the example in Figure 19a, $v = 2m/s$ the reflection arrival time is $T = 5s$, $x_S = 0$ and $x_R = 8\ m$, resulting in an offset of $D = 8m$. To eliminate the ambiguity and determine where on the ellipse the reflector is located, the sum of the ellipses of all source-receiver pairs in the model space will create constructive interference due to energy arriving in phase, and reveal the true reflector position as shown schematically in Figure 19b for a horizontal reflector.

The imaging condition (19) in the model space can be written in an equivalent way, for one source and all its corresponding receivers as,

$$g(x) = \frac{\partial \phi_0(u(x,t,m))}{\partial m} = \left(P\frac{\partial u}{\partial m}\right)^{\dagger} (Pu(x,t,m) - d(x,t)) \tag{7}$$

$$= -\int_0^T \frac{\partial^2 u(x,t,m)}{\partial t^2} \lambda(x, T-t, m) dt, \tag{8}$$

where g stands for gradient, u is the solution of the direct problem, $\lambda(T-t)$ is a back-propagated wavefield that is computed using the residual data at the receiver positions as sources. In essence, the imaging condition is proportional to the correlation of the direct and back-propagated

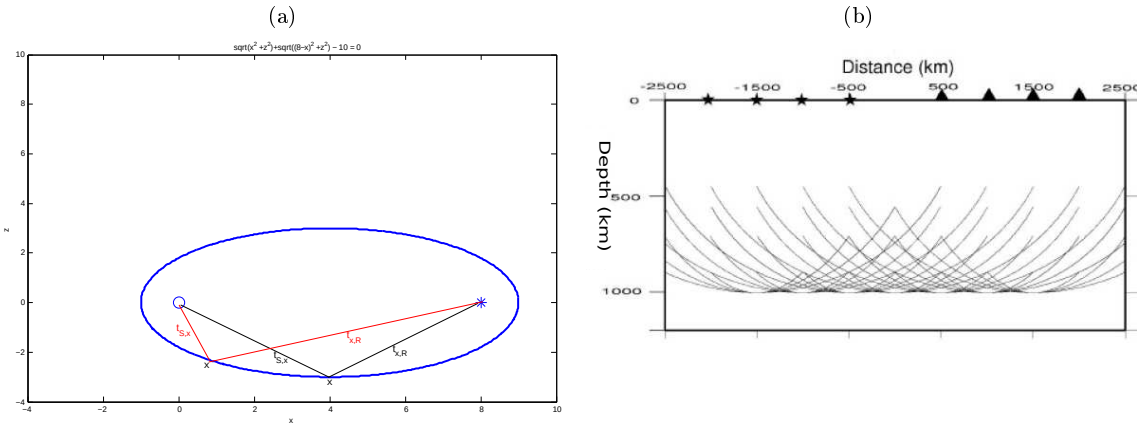


Figure 5: a) Example of the points satisfying the imaging condition $t_{S,x}+t_{x,R} = T$, for a reflection residual at time $T = 5s$, in a model with $v = 2m/s$, with a source at $x_S = 0$ and a receiver at $x_R = 8\ m$, resulting in an offset of $D = 8m$. The points form an ellipse with focal points at the source and receiver positions. b) Figure from Agudelo (2005). Imaging a horizontal reflector. Intersection of ellipses create coherent interference, and an image is created at the depth of the reflector.

wavefield on all points of the domain, integrated over time. By applying a Fourier transform, the gradient can also be expressed in the frequency domain,

$$g(x) \quad = \quad \int_{-\infty}^{\infty} \omega^2 u(x,\omega,m)\lambda^*(x,\omega,m)d\omega. \tag{9}$$

More details about the interpretation of the imaging condition can be found in Chapter 2, section1.2.b. The imaging condition of FWI is related to the imaging condition of other seismic imaging techniques, as is explained in more Appendix 1, via the time reversal procedure.

1.1.a *Resolution*

On a theoretical level, the quality of the images of the earth parameters provided by FWI will supersede those provided by previous imaging techniques because the full waveform content is used in the inversion process and because the model used to solve the forward problem (18) is updated in every iteration. The breakthrough with respect to previous techniques was due to the way the wave equation is solved. Without sufficient computational power, the wave equation was solved analytically and asymptotically using *ray theory*, which required smooth models. It was therefore necessary to use a *linear approximation of the wave equation*, and separate the model m into a smooth background model m_0 and a small unknown model perturbation δm : $m = m_0 + \delta m$. Once this is done, the solution of the wave equation is found using first order perturbation theory, known as the *Born approximation*. The total wavefield u is approximated as $u(m_0 + \delta m) \approx u_0 + (\partial u/\partial m)\delta m$, where u_0 is the solution of the wave equation with a model m_0 and $(\partial u/\partial m)\partial m$ is the scattered wavefield by a model perturbation δm. Note that the Born approximation ignores higher order diffraction terms (i.e waves diffracted two times are not modelled). This distinction between m_0 and δm imposes a *scale separation between a smooth background model and the perturbation model*, as shown in Figure 20a,b. When the computational power became sufficient, the wave equation could be solved numerically using *finite differences* for an arbitrary complex model (Virieux, 1984). Without the need to linearize the wave equation, there was no longer need to make a scale separation between a smooth background model and

the perturbation model, and the waves were propagated in complex models containing both the smooth and reflectivity properties. If the acquisition system allows to measure both transmitted and reflected data and if the source has a sufficiently broad bandwidth, FWI is theoretically amenable to the imaging of a broad and continuous spectrum of wavenumbers. In practice, however, it has been difficult to update simultaneously both the transmission and reflection characteristics of the model. The first inversions only partially reconstructed the image (Devaney, 1984; Wu and Toksöz, 1987). The question asked at the time, and that still remains is, can the inversion algorithm retrieve the low and high wavenumbers simultaneously (Mora, 1989)? A horizontal and vertical resolution capacity of the data should be performed to estimate which wavenumbers in the image domain may be reconstructed. A more detailed explanation of the resolution analysis and limitations are explained in Chapter 2, section 1.5.

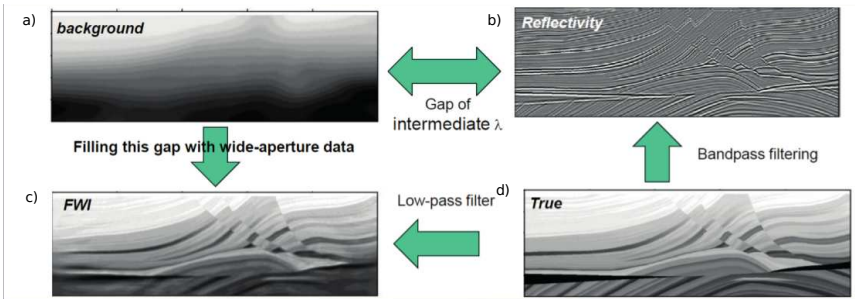


Figure 6: Figure d) shows the true velocity model m. In a linear approximation, the model m is separated in a smooth background model m_0 (Figure a) and a perturbation model δm (Figure b). FWI does not impose a scale separation if transmission and reflection data are available, a low pass version of the wavenumber image spectrum can be reconstructed (Figure c).

1.1.b *Computational Cost*

Although the computational power became sufficient in the 1980's to solve the wave equation for an arbitrary model with finite differences in the time domain, FWI was still out of reach for real data applications.

Adjoint method

A first step that dramatically decreased the computational cost was the use the *adjoint state method* to compute the gradient (Lailly, 1983; Tarantola, 1984a). Using equation (20), it is necessary to evaluate $\partial u/\partial m_i$ which are the partial derivatives of the computed wavefield with respect to *all* the model parameters, $i = 1, ..., N \times N_p$. Here N represents the number of grid points in the discretization of the domain, and N_p the number of physical parameters to be reconstructed (ρ, v_p, v_s). It is possible to compute the partial derivatives using finite difference schemes, $\partial u/\partial m_i = (u(m + \Delta m_i) - u(m)/\Delta m_i$, which represents a computational cost of $C = 2 \times \mathbf{N} \times \mathbf{N_p} \times N_s$, where N_s is the number of sources. Using the adjoint state method (Lailly, 1983; Tarantola, 1984a), the gradient is computed via equation (21). The computational cost is $C = 2 \times N_s$ because, for each source, one wave equation is solved to find u and another to find the back-propagated wavefield λ. The computation of the gradient with the adjoint state method is explained in Appendices 3,4.

Frequency domain

A next step in reducing the computational cost of FWI was taken by working in the *frequency domain*. In 1990 the equations for FWI (Tarantola, 1984a) were transformed to the frequency domain, for the acoustic and elastic forward problem (Pratt and Worthington, 1990; Pratt, 1990; Pratt and Shipp, 1999). Applying a Fourier transform to the forward problem (18), the discretization with finite differences (FD) is performed directly (the forward modeling operator in the frequency domain is sometimes referred to as Helmholtz operator). The resulting forward problem is

$$A(m,\omega)u(s,m,\omega) = s(\omega), \tag{10}$$

where $A(m,\omega)$ is the discretization of the wave equation operator in the frequency domain, and depends on the frequency $\omega = 2\pi f$. Equation (23) is a linear system where A is a matrix of dimensions $N \times N$, and u and s are vectors of dimensions $N \times 1$. This linear system can be solved with direct or iterative methods. With iterative methods, the number of iterations needed to find the solution depends on the conditioning of matrix, and therefore adequate preconditioners are needed (Erlangga, 2005; Plessix, 2007; Erlangga and Nabben, 2008). Each source function requires an independent iterative solution of the system. Direct methods to solve linear systems are based on Gauss elimination and include methods such as the $A = LU$ factorization of the matrix. The LU factorization of A in (23) only depends on m and the frequency ω. Therefore, for each iteration of the inverse problem, one LU factorization is performed per frequency. For all the source functions, the wavefield is easily found by forward and backward substitutions. If the number of frequencies is less than the number of sources, using the LU factorization to solve (23) allows to solve the direct problems efficiently because the LU factorization, which represents the highest computational effort in the solution of the forward problem, is only done once per source. In FWI this represents a considerable gain (Pratt and Worthington, 1990; Pratt, 1990) because in a typical 2D seismic experiment the number of sources is of the order of $10^2 - 10^3$ and in a 3D seismic experiment it is around $10^3 - 10^4$. Nonetheless, for a certain period it was thought that working in the frequency domain would not represent a computational gain in real applications, specially with short offset data, because all the frequencies had to be taken into account in the inversion, with $\Delta f \geq 1/2T_{max}$ (Freudenreich and Singh, 2000). Based on a resolution analysis of wave number coverage, it was shown that the number of frequencies needed were much less than those dictated by the Nyquist criteria and indeed FWI could be successful with a few frequencies (Sirgue and Pratt, 2004).

Besides reducing the number of forward problems to be solved, working in the frequency domain also has computational advantages in the inverse problem in terms of the memory required to store the wavefields. To evaluate the gradient (21) in the time domain, the correlation of the direct and back-propagated wavefields is performed at each time step t. Since the back-propagated wavefield λ is computed only after the the direct wavefield has been computed to its final time, the cross-correlation of the wavefields can be done in two ways. Either by storing in memory the N_t direct wavefield $u(x,t), x \in \Omega$, where N_t is the number of time steps in $[0,T]$, or by recomputing u_c as it is needed to perform the cross-correlation with the back-propagated wavefield. The first method is rarely used, and only the second method remains efficient, but requires the re-computation of wavefields. On the other hand, if the inverse problem is performed in the frequency domain, the gradient is proportional to the multiplication of the direct and back-propagated field, for each frequency. Therefore only two wavefields, $u(x,\omega)$ and $\lambda(x,\omega)$, need to be stored in memory to compute the gradient. More details can be found in Chapter 2, section 1.4.

Source encoding

The limitations of working in the frequency domain with direct solvers is that the LU factorization has to be stored in memory. With a finite difference grid, the matrix A is a sparse

banded matrix. For 2D applications there are 3 bands and the distance between bands is N. For 3D applications there are five bands and the distance between bands is N^2. Unlike A, the L and U matrices are full and thus are memory demanding. For *3D acoustic* applications it may still be feasible to apply direct solvers (Operto et al., 2007). The first real data applications of 3D acoustic FWI the forward problem was solved in the frequency domain with iterative solvers (Plessix, 2009; Plessix and Perkins, 2010) or time marching algorithms (Sirgue et al., 2010). For *3D elastic* applications it seems almost certain that direct solvers may face difficulties. The reason is that finite difference schemes will require a minimum number of discretization points per wavelength to avoid numerical dispersion. Typically, in the earth $3000\ m/s < v_p < 5000\ m/s$, and $1300\ m/s < v_s < 3000\ m/s$. Therefore, for a fixed frequency $\lambda_s < \lambda_p$ and modeling shear waves will require a much finer grid, greatly increasing N in each dimension. The size of L and U are therefore too large to store in memory, and iterative solvers must be employed.

Computational gains have been attempted to reduce the number of the direct problems, which is often the most computationally expensive part of FWI. Note that with iterative or direct solvers the number of direct problems is proportional to the number of sources (equations (18) and (23)). However, when working with direct solvers, the computational burden is greatly relieved because the same LU factorization may be used for all sources. Therefore, to reduce the number of direct problems, the number of right hand sides of the direct problem have to be reduced. In *source blending* the acquisition of the data is modified. Several sources are shot simultaneously, or with some time delays between them, and the recorded data by the sensors therefore has the information of each source, and the interference effects between them (Beasley, 2008; Berkhout, 2008). Source blending reduces the time of collection data in the field and also the number of sources that are processed when solving the direct problem. Another popular speed-up technique, called *source encoding*, consists in performing a traditional seismic acquisition experiment using only one source at time, for N_s sources. However, instead of solving the direct problem (18) or (23) for each source, encoded super-sources $\tilde{s}$ are formed by creating random linear combinations of the sources (Romero et al., 2000; Neelamani et al., 2008; Krebs et al., 2009; Ben Hadj Ali et al., 2011; Schuster et al., 2011; van Leeuwen et al., 2011; Huang and Schuster, 2012),

$$\tilde{s} = \sum_{i=1}^{N_s} \alpha_i s_i, \qquad \tilde{d}_o = \sum_{i=1}^{N_s} \alpha_i d_{o_i}, \tag{11}$$

where α are random coefficients. The observed data is also encoded into $\tilde{d}_o$. To be able to encode the data in this way, it is necessary for all the sources to share the same set of receivers (referred as fixed-spread acquisitions). The volume of data produced by N_s sources and measured at N_r receivers is encoded into only one recorded data set at the receiver positions.

Even though source encoding with gradient descent algorithms reduces the computational cost per iteration (2 direct problems, instead of $2 \times N_s$), more iterations need to be performed because each model update is less accurate due to interference effects between sources (referred to as cross-talk), as illustrated in Figure 21a. In the end, there will be a computational gain (speed-up) if the number of direct problems solved to attain a predefined value of misfit function is less with source encoding than using the sources individually in the standard way, as illustrated in Figure 21b. Source encoding techniques have been applied successfully to real data sets (Baumstein et al., 2011; Routh et al., 2011; Bansal et al., 2013; Schiemenz and Igel, 2013).

1.1.c *Non-linear optimization*

Early applications of FWI showed that, although the accuracy of the forward problem was greatly improved and FWI seemed feasible, there were several additional difficulties with the

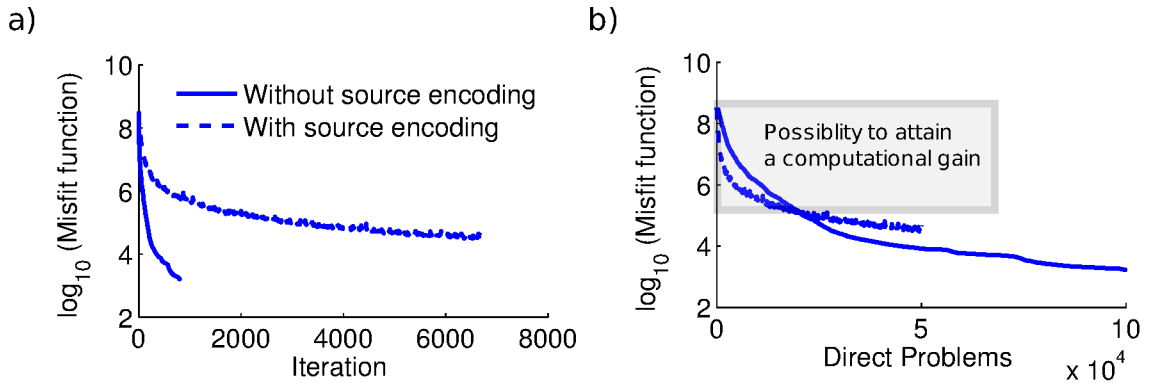


Figure 7: FWI inversion convergence curves with source encoding (dashed lines) and without source encoding (solid lines). (a) Comparison between the convergence rates. Less iterations are required without source encoding to attain the same misfit value. (b) Comparison of the computational efficiency (direct problems). If the reduction of the misfit function remains in the gray box, a computational gain can be attained because less direct problems are solved with source encoding than without. This example is presented in more detail in Chapter 3.

optimization problem (Gauthier et al., 1986; Luo and Schuster, 1991). The misfit function (17) is non-linear with respect to m. This means that small changes in the model parameters can generate large changes in the misfit function ϕ_0, and the variations occur in a non-linear fashion. In an acoustic 2D cross-well (transmission) data inversion Luo and Schuster (1991) compared travel time inversion and FWI. The numerical inversions showed that the resolution power of FWI is higher. However, travel time inversion was more successful because FWI converged to local minima.

Non-convex optimization

A numerical illustration of the difficulties in minimizing the l_2 norm of the waveform misfit (17) which is non-convex compared to a travel time misfit which is convex, is illustrated in Figure 22. Travel time tomography defines the misfit ϕ_0 as the difference between the observed and calculated arrivals times of the maximum of the wavefield

$$\min_m \phi_0 = \min_m \|T_c(m) - T_o\|_2^2, \tag{12}$$

where T_o represents the observed arrival time of the maximum and T_c represents the calculated arrival time (See Appendix 2 for a more precise description.). Consider a wavefield whose displacement can be described by $d(t) = e^{(t-2.5)^2} \sin(\omega t)$. For a low frequency (long period) $T = 2\ s$, the wavefield is plotted in dark blue in Figure22a. Imagine the computed wavefield $u(t)$ is equal to the true wavefileld but arriving τ seconds in advance, $u(t) = d(t - \tau)$. The computed wavefield $u_c(t)$ for $\tau = 0.5s$ is plotted in dashed red. The maximum of the computed wavefield T_c and observed wavefield T_o are marked with light blue stars. Figure 22d shows the misfit function as a function of the delay time error τ in the computed wavefield. For the range of values plotted here, the travel-time misfit function (light blue) increases as the delay increases. As can be appreciated graphically, the optimization algorithm will be able to locate the minima. The waveform misfit (dark blue) shows that only if $\tau < 1s$, a local optimization algorithm will provide the right solution. In particular, for $\tau = 0.5$ which corresponds to the red wavefield 22a, the minimization will converge because the the red star in Figure 22d shows that this point is in the valley of attraction. The optimization of the full waveform difference becomes more difficult as the frequency increases (period decreases). For Figure 22e where $T = 1s$, the initial delay must

satisfy $\tau < 0.5\ s$ to converge, and in Figure 22f where $T = 0.5\ s$, the initial delay must satisfy $\tau < 0.25s$ to converge. As can be deduced from the figures, the condition to avoid converging to secondary minima is that $\tau < T/2$. In reference to Figure 2.16, this means that the model should provide delays within the first maximum of the waveform. When this condition is not satisfied, *cycle-skipping* occurs. For example, for a frequency of $f = 2Hz$ and $\tau = 0.5s$, Figure 22f reveals that using the l_2 norm of the waveform misfit, the inversion will remain at a local minimum, where cycle skipping occurred. The non-linearity therefore requires that the starting model be sufficiently accurate to be in the valley of the global minimum and avoid cycle skipping.

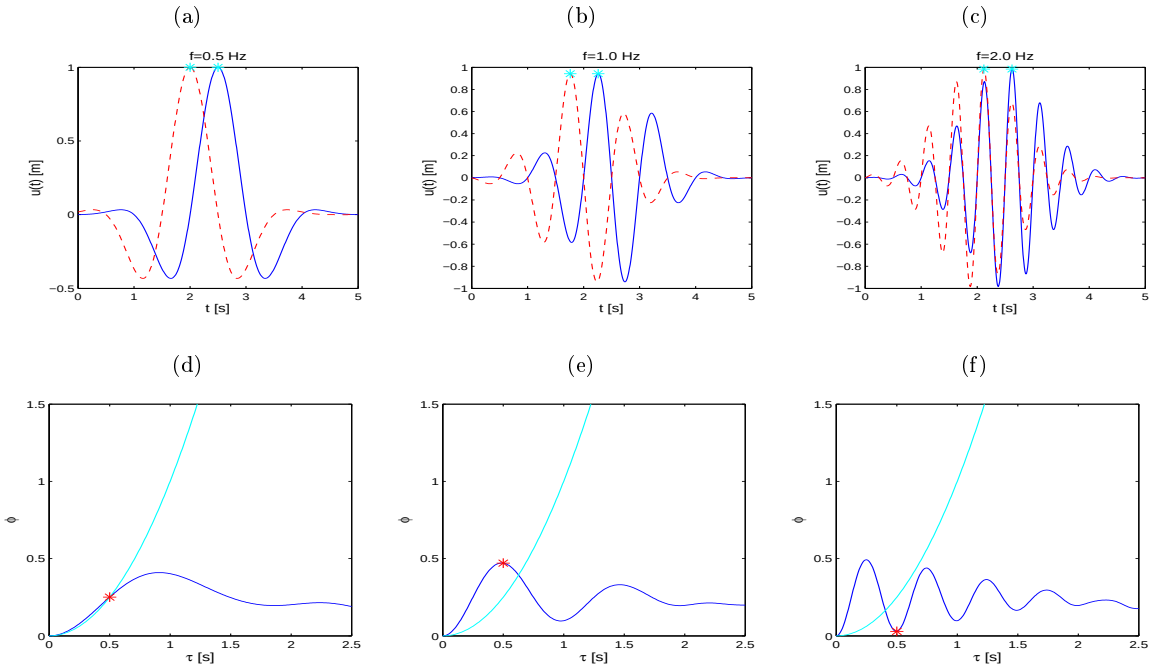


Figure 8: The first row plots the observed wavefield $d(t) = e^{(t-2.5)^2} \sin(2\pi f t)$ in blue, and the computed wavefield in red, $u(t) = d(t-\tau)$ for $\tau = 0.5s$. The wavefields are plotted for a) f= 0.5 Hz b) f=1.0 Hz, d) f=2 Hz. The second row plots the travel time misfit function (25) in light blue as a function of the delay τ, and the full waveform misfit function (17) in dark blue. The misfit functions are plotted for d) f= 0.5 Hz e) f=1.0 Hz, f) f=2 Hz. For the values plotted here, the travel time misfit function is convex (light blue). The misfit function measuring the difference in the full waveform (dark blue) is non-convex. For a delay of $\tau = 0.5Hz$, the optimization will be successful for $f = 0.5\ Hz$ but will converge to a secondary minima for $f = 2hz$. This is called cycle-skipping.

Alternate definitions of the misfit function have been proposed to overcome this difficulty both in the time and in the frequency domain (Shin et al., 2002; Sheng et al., 2006; Shin and Min, 2006; Pyun et al., 2007; Shin et al., 2007; Shin and Ha, 2008; van Leeuwen et al., 2010; Hale, 2013). The purpose is to create misfit functions that are less-sensitive to cycle skipping, which is attained by creating misfit functions with a less oscillatory behaviour (as in Figure22a), at the expense of loosing information in the waveform (for example, use the envelope of the misfit function, time damping or cross correlating the wavefields). Alternate misfit functions may be particularly useful in the absence of low frequencies in the data, to avoid premature cycle-skipping and to obtain an initial model that will be then fed to FWI with an l_2 norm misfit function.

Hierarchical inversion

With the same train of thoughts, the inversion in the time or frequency or frequency domain can be performed in *sequential hierarchical* way, using increasing frequency information (Bunks et al., 1995; Pratt and Shipp, 1999; Sirgue and Pratt, 2004). The algorithm consists in low-pass filtering the data, and performing a complete FWI. Let us denote the final model m_1. The data is then passed by a low pass filter or band-passed filter with a higher cut-off frequency, using as a starting model m_1. The inversions are continued sequentially to contain all the frequency content range of the data. In the example in Figure 22, the hierarchical approach would consist in first minimizing the misfit function in Figure 22a, and using the final model of the inversion to minimize the misfit function shown in Figure 22b.

The waveform difference of low frequency data provides misfit functions that are more convex, and therefore are easier to minimize and require a less accurate starting point. The waveform difference of high frequency data is highly non-convex and requires an accurate initial point for the minimization to converge to the global minima. The hierarchical approach in FWI therefore provides a better starting model as the frequencies increase. An illustration of the application of a hierarchical inversion in FWI is shown in Figure 23. The first FWI is performed only with $3Hz$ data. The final model of FWI for the $3Hz$ data is used as a starting model for FWI with 5 Hz data. The sequential inversion process is continued with $8Hz$ and $20Hz$. Note that low frequencies add big characteristics to the image and as the frequency increases (wavelength decreases), more details are added to the models and the resolution improves.

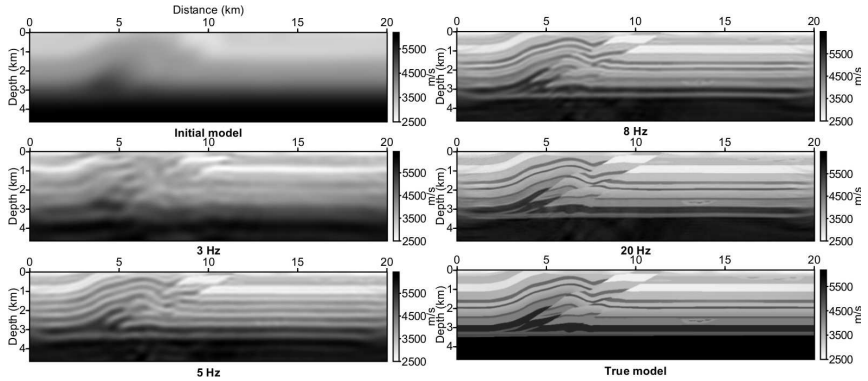


Figure 9: Sequential FWI problems solved from low to high frequencies. The initial problem is on the top left, and the final model is on the bottom right. Nine frequencies between $3Hz$ and $20Hz$ are inverted. The final model of the FWI of $3Hz$ data is used as a starting model for FWI using $5Hz$ data. This process is continued sequentially, until arriving at $20Hz$. The image wavenumber spectrum is progressively contributing with high wavenumber content as the frequency increases.

The Hessian

Other attempts to improve the optimization results have been done by including the second derivatives of the misfit function (the *Hessian*) in the minimization process (Pratt and Worthington, 1990). Using *Newton methods*, the direction of descent is the inverse of the Hessian times the gradient. In one of the first applications of (non-linear) FWI, the results were compared to (linear) diffraction tomography ,and there was the clear intuition that the action of Hessian should help to converge faster, because the inverse Hessian would act as a focusing

operator (Pratt and Worthington, 1990). However, because of the computational cost, it was not until a few years later that a clear demonstration of the benefits of the Newton methods was made. With a second order approximation of the misfit function, the Hessian consists of two terms, one related to the correlation of scattered wavefields $\partial u/\partial m$ and another related to second order scattering $\partial^2 u/\partial m^2$. It was shown that the smearing artefacts that appear due to the ambiguity in the imaging condition (Figure 2.16), can be reduced with the deconvolution action of the inverse Hessian under the Gauss-Newton approximation (using only first order scattered wavefields) (Pratt et al., 1998). The full Newton approximation is beneficial to correct for second-order scattering terms. That is, the solution of the full wave equation models multiple diffracted waves. However, there is an inconsistency because the imaging condition (20), only images first order scatter terms $\partial u/\partial m$. Therefore, the parts in the residual wavefield due to higher-order scattering will be wrongly considered as first-order scattering and will be imaged as additional diffraction points, providing artefacts in the image. The Hessian provided by the full-Newton approximation will correct these artefacts (Pratt et al., 1998). In theory, when the Hessian is not used, the artefacts should disappear anyway with gradient algorithms as iterations proceed. However, in practice, due to the non-linearity, the inversion with gradient algorithms may fall into a local minima due to the imaging artefacts, and never recover. For example, to recover the true velocity model in Figure 24a with a frequency of $7Hz$ is difficult because the frequency is very high and there is a high chance of cycle skipping and falling in local minima. In addition, there are double diffracted wavefields between the two spherical inclusions. The inversion using full Newton in Figure 24e, provides the best solution. The other advantage of Newton methods is to reduce the number of iterations required to reach a misfit value, with the set back that the computational price is higher. Real data applications using Newton and quasi-Newton(l-BFGS) optimization algorithms are shown in (Brossier et al., 2009a,b; Plessix et al., 2012; Métivier et al., 2014).

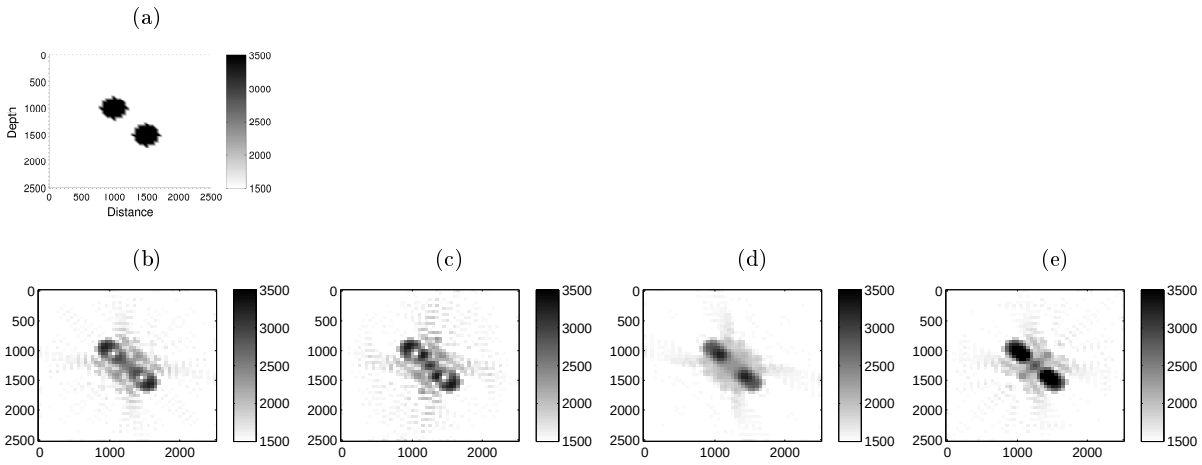


Figure 10: Inversion with a frequency $f = 7Hz$, with regularization, using different optimization methods. Due to the presence of local minima and an inaccurate initial velocity model, all methods provide different final models. a) True velocity model v_p b) v_p using steepest descent c) v_p using l-BFGS, d) v_p using Gauss Newton e) v_p using Full Newton. For more details on this example, go to Chapter 1.3.a.

Multi parameter inversion

The benefits of including the Hessian extend to multi-parameter FWI. Early simultaneous multi-parameter inversion results showed that the resolution of the reconstructed of the parame-

ters depends on the parametrization (Tarantola et al., 1984; Mora, 1987; Pratt and Worthington, 1990). There are coupling effects between the different physical parameters (v_p velocity, v_s velocity, density, attenuation, parameters to describe anisotropy (ϵ, δ)), which makes the optimization process challenging because the gradient of with respect to one parameter, contains the influence of other parameters. That is, a model perturbation for one parameter is in fact a linear contribution of the gradients with respect to all the parameters. In addition, the sensitivity of the data to each parameter $\partial u/\partial m$ may have different orders of magnitude. As a result, the inversion will only update the parameters with the highest partial derivatives. A re-parametrization of the physical parameters in the inversion, will help to rescale and decouple the physical parameters. A re-parametrization of m will modify $\partial A/\partial m$, known as the radiation pattern, present in the gradient expression. To reduce the cross talk between parameters, re-parametrizations can be searched such that the radiation patterns of different parameters have the smallest intersection possible. The Hessian will also help to decouple the effects and to rescale the magnitude of the parameters. For example, if ρ and v_p are the parameters to be reconstructed, the Gauss-Newton approximation of the Hessian will contain four main blocks with the cross-correlations $(\partial u/\partial v_p)^\dagger (\partial u/\partial v_p)$, $((\partial u/\partial v_p)^\dagger (\partial u/\partial \rho)$, $(\partial u/\partial \rho)^\dagger (\partial u/\partial v_p)$, $(\partial u/\partial \rho)^\dagger (\partial u/\partial \rho)$. Each block will have a different magnitude that will, theoretically, balance the magnitudes of different entries of the gradient so as to update both v_p and ρ. The Hessian will depend on the parametrization. Different parametrizations will provide different weights to the partial derivatives. An adequate parametrization will lead to a well conditioned Hessian, and each of the blocks will become similar in magnitude. In addition, in the same way that the Hessian corrected the smearing ambiguity artefacts and focused the energy on the reflector (Pratt et al., 1998), with multi-parameter inversion this will be useful to decrease the interaction between parameters, and refocus the energy on one parameter. An asymptotic (infinite frequency and Born approximation) study of the eigen-vectors of the Hessian, can help to determine if a parametrization is suitable and allows to decouple the physical parameters (Forgues and Lambaré, 1997; Plessix and Cao, 2011; Operto et al., 2013). Real data elastic anisotropic applications studying the effect of the parametrization have been performed (Gholami et al., 2013b,a; Prieux et al., 2013a,b).

Regularization of the inverse problem

The difficulties in the optimization process and the under-determination of the inverse problem require *regularization* terms to constrain the space of solutions and guide the inversion towards a final model that a geophysicist considers acceptable. For example, noisy models are discarded and the smoothest models are preferred. The regularization term can be included in the misfit function,

$$\begin{aligned} \phi &= \phi_0 + \lambda \|R(m)\|_2^2 && (13) \\ &= \|Pu_c(m) - d_o\|_2^2 + \lambda \|R(m)\|_2^2, && (14) \end{aligned}$$

where $R(m)$ is the regularization term and $\lambda\|R(m)\|_2^2$ is a penalty term via a Lagrange multiplier. The optimization will seek to find the model that best explains the data *and* the regularization term simultaneously. Traditionally, smooth models are favored and thus $R(m) = \nabla m$ so as to minimize the gradient. If there is a-priori information about the values of the model at certain positions, this can also be included in $R(m)$ (Asnaashari et al., 2013). There have been very few works using other regularization terms or norms. For example, it is possible to change the norm and minimize the total variation (TV) of models (Ramírez and Lewis, 2010; Anagaw and Sacchi, 2012; Guitton, 2012),

$$\phi = \|Pu_c(m) - d_o\|_2^2 + \lambda \|\nabla(m)\|_1, \quad (15)$$

or minimize the number of coefficients of the model under a change of basis W (Candes and

Romberg, 2005; Loris et al., 2010; Herrmann and Li, 2012),

$$\phi \quad = \quad \|Pu(m) - d\|_2^2 + \lambda \|W(m)\|_1. \tag{16}$$

Other regularization schemes have been proposed, where the regularization term is not added to the misfit function, but instead it is multiplied (Abubakar et al., 2002, 2004). Theoretically, multiplicative regularization has the advantage that the relative weights between the data gradient and the model gradient are naturally determined.

1.2 Introduction to this thesis and description of main results

The brief discussion of the main questions and challenges in FWI shows why this field is an active area of research since the 1980s. This thesis aims to add to the growing understanding of FWI by addressing some points concerning the computational cost and the non-linear optimization process.

We start by presenting in Chapter 1 the inverse problem theory and optimization algorithms for a general setting, without restraining ourselves to the specific case of seismic imaging. We introduce the concepts of linear and non-linear inverse problems, and ill-posed problems. As inverse problems are generally formulated as minimization problems, we introduce the optimization algorithms that we make reference to throughout this thesis. For the specific application of FWI, this chapter is not essential.

Chapter 2 complements the introduction to FWI. A physical interpretation is given to the two main objects that appear in the minimization algorithms: the gradient and the Hessian. For a small size problem we compute the full Hessian and we illustrate the action of one line on the inverse Hessian on the gradient to get an understanding on how the artefacts are corrected. The physical interpretation of the imaging conditions of other seismic methods that also make use of time reversal are briefly reviewed in Appendix 1. The equations for the gradient and the Hessian found in the literature with the adjoint-state method (Lions, 1972; Chavent, 2009; Plessix, 2006; Métivier et al., 2013a) are used here, and re-derived in Appendix 3 and 5.

Self-adjoint formulation of the elastic isotropic wave equation

Generally, in FWI, the gradient computation with the adjoint state method is implemented with a second order self-adjoint expression of the wave equation. For the elastic wave equation, the fist-order velocity-stress formulation does not provide a self-adjoint formulation ($A \neq A^\dagger$), implying that the forward and adjoint modeling operator have to be implemented independently. Our contribution, presented in Appendix 4, lies in developing a formalism to recast the isotropic first order elastic velocity-stress equations in a self-adjoint fashion. This is done through linear transformations based on a change of basis into the eigen-vector space of the operators. This allows to compute the gradient with the only one implementation of modeling operator. This work can be found in Castellanos et al. (2011).

Source encoding with second order optimization methods

With iterative and direct solvers, the number of forward problems are proportional to the number of sources. With the LU factorization, the forward problem is solved more efficiently because the same factorization can be used for all sources, and the wavefield solution to each source is simply found by substitutions. However, to perform 3D elastic FWI in time or frequency domain with the current computational resources, time marching and iterative solvers appear to

be the most feasible solution, and source encoding can greatly reduce the computational burden. Although source encoding techniques have been widely used (Neelamani et al., 2008; Krebs et al., 2009; Ben Hadj Ali et al., 2011; Schuster et al., 2011; van Leeuwen et al., 2011; Huang and Schuster, 2012), they have been mostly combined with gradient optimization algorithms, resembling stochastic gradient algorithms (Robbins and Monro, 1951; Spall, 2003). With the purpose of reducing even more the computational cost and improving the gain (increasing the gap in the curves with and without source encoding in Figure 21), we combine quasi-Newton and Newton optimization methods with source encoding techniques, which had not been done in the context of FWI. In other application areas such as machine learning, this exploration has also recently begun (Schraudolph et al., 2007). The lack of convergence proofs of second-order stochastic methods (Bottou and Le Cun, 2005), and the lack of efficient formulations to compute the Newton descent directions (Métivier et al., 2013b, 2014) are some of the reasons. The equations for the efficient Hessian computation are shown in Appendix 5. However, the computational cost per iteration of the truncated Newton methods (Gauss-Newton and Full Newton) is higher per iteration, because the Newton descent direction requires the solution of an additional linear system. In the case of $l-$BFGS, the Hessian approximation needs to be periodically restarted. Therefore, a priori, it is not clear whether the computational savings can be further improved with second order optimization methods.

In Chapter 3, we predefine stopping criteria based on the reduction of the misfit function, and compare the convergence (iterations) and computational efficiency (forward problems) of four optimization methods (non-linear conjugate gradient, $l-$BFGS, Gauss Newton and Full Newton), when they are implemented in an efficient frequency-domain FWI with random source encoding. We compare the convergence, costs and final velocity models with the results obtained with the individual sources. This work-flow is first applied on a realistic synthetic experiment inspired by the geology of the Gulf of Mexico both for noise-free data and noisy data. Then, we assess the benefit provided by random source encoding and second-order optimization methods when applied on a *2D* real ocean-bottom-cable (OBC) dataset recorded from the Valhall oil field. Even though the maximum advantage of source encoding can be seen with time marching or iterative solvers, we work with direct solvers. However, the conclusions we draw are based on the number of direct problems solved and therefore are directly applicable to iterative solvers in the frequency domain. Part of these results were submitted for publication (Castellanos et al., 2013).

Conclusions on source encoding with second order optimization methods with synthetic data

Our FWI results on the synthetic case when the initial model is sufficiently accurate and the data does not contain noise, indicate that the computational gain is boosted with source encoding combined with second order methods. The highest computational efficiency is provided by $l-$BFGS and Gauss-Newton, and the highest convergence rate is attained by Gauss-Newton. The savings in computational cost for our numerical test are shown in Figure 25. When very noisy data is considered (the power of the noise is 25% of the power of the data), the convergence rates of all optimization algorithms without source encoding is similar. This suggests that the action of the inverse Hessian on the gradient does not considerably improve the descent direction, due to the noise in the data. With source encoding, there is a similar situation and the convergence rates and computational cost of all methods are comparable. However, Newton methods show the lowest statistical variance in the final model, when many FWI inversions are performed. The computational savings for our numerical test with noise are lower compared to the noiseless scenario.

Conclusions on source encoding with second order optimization methods with real

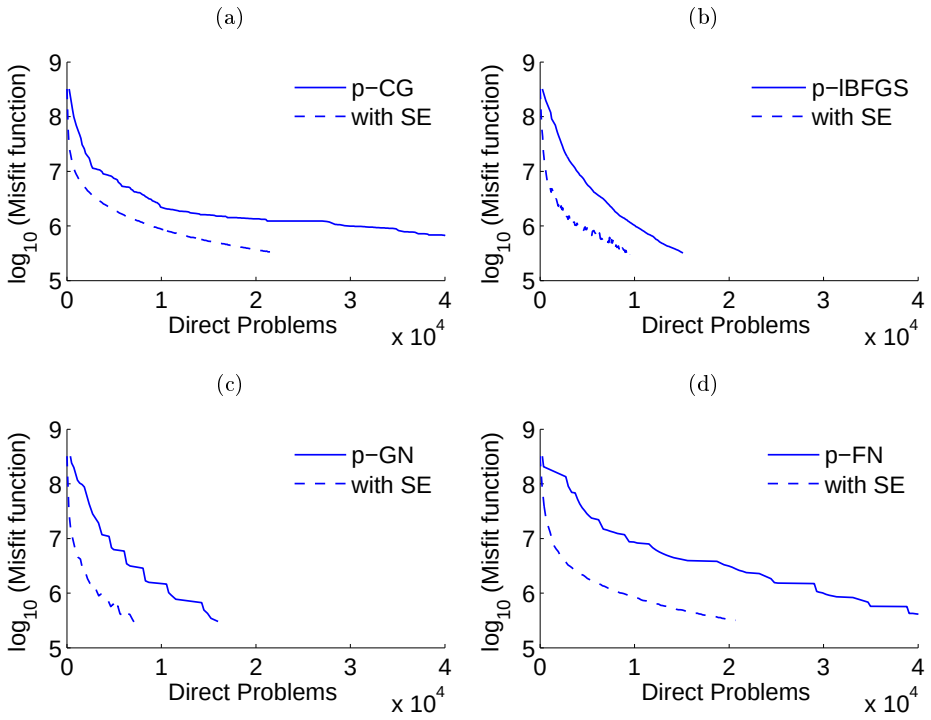


Figure 11: BP case study without noise. Assessment of the computational efficiency provided by source encoding - Reduction of the misfit function as a function of the direct problems. Optimization algorithm a) Conjugate Gradient - (Steepest descent with source encoding) b) l-BFGS c) truncated Gauss Newton d) truncated Full Newton. The solid lines represents the computational cost without source encoding and the dashed lines represent the computational cost with source encoding. More details in Chapter 3, Section 5.1.b.

data

When we use the real data set of the Valhall oil field, the inversion with different optimization methods does not converge to the same local minima. This is probably due to the fact that near the initial model, the misfit function has several local minima. The assessment of the computational gain on the real data set is therefore biased, because the quality of the final models is not exactly the same. We observe that with and without source encoding, Newton methods provide the descent direction that result in models with the lowest data misfit. As in the synthetic case with noisy data, the lowest statistical variance is provided by the Newton methods. Because there are considerably more sources in the real data set than in the synthetic test, the computational savings provided by source encoding are $\approx 90\%$. In all our numerical tests, the final quality of the models with source encoding are very close to the models obtained with the full set of sources. We verify that the quality of our final velocity models is adequate by comparing the migrated images computed in the different FWI models with and without source encoding.

Estimation of the variance of the encoded gradient

To have a better understanding of the cross-talk, in section 6 of Chapter 3 we estimate the variance of the encoded gradient with noisless data, with three ways to create the super sources. The variance estimations may used design strategies to encode the sources. We find that when the gradients generated by several sources are similar, which may occur when there is a very

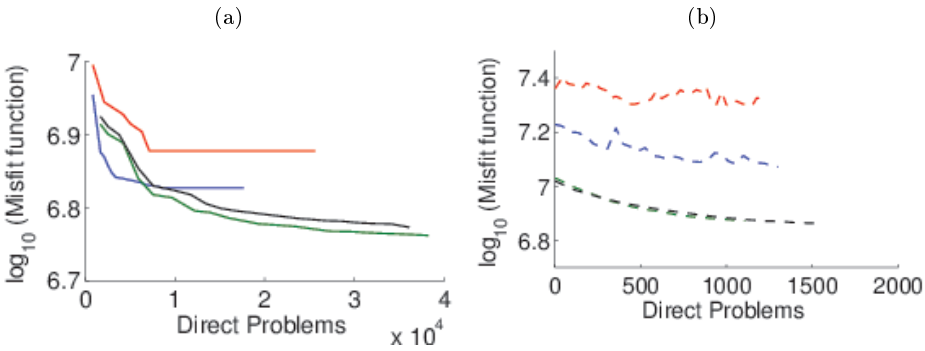


Figure 12: Valhall case study. Assessment of the computational efficiency for the fourth frequency group. Reduction of the misfit function as a function of the direct problems a) without source encoding (solid lines) b) with source encoding (dashed lines). Blue : Conjugate Gradient, Red: l-BFGS, Green: Gauss Newton, Black : Full Newton. The difference in the number of direct problems is of one order of magnitude. More detail in Chapter 3, Section 5.2.

dense acquisition of sources and receivers and there is a lot of redundancy in the data, assembling all the sources in one super source provides a large variance. In this case when there is a lot of redundancy, it is preferable to subsample the sources. On the other hand, if the gradients generated by each source are very different, subsampling the sources with provide a large variance in the encoded gradient, and it is better to assemble all the sources in one supersource.

General remarks on source encoding with second order optimization methods

The computational savings of source encoding can be increased by combining it with quasi-Newton or Newton optimization methods. However, as the noise level in the data increases, the computational savings obtained with second order methods are reduced, and become similar to the savings obtained with gradient descent algorithms. However, even if this is the case and the computational savings are reduced and similar to those obtained with gradient algorithms, Newton methods provide the most robust direction of the descent and provide the final velocity models with the lowest statistical variance.

Stochastic optimization algorithms have lower convergence rates than deterministic optimization algorithms. Therefore, the maximum computational gain is obtained in the early part of the inversion and decreases as the inversion proceeds. In other words, the computational gain depends on the data misfit value. To assure a reasonable computational gain with source encoding, stopping criteria have to be defined based on a tolerance error of the data misfit value.

Despite the cross-talk generated by the encoding of the data and sources, the quality of the final velocity models is satisfactory. Moreover, we illustrate with an example that when we force the misfit function to be strongly non convex by reducing the accuracy of the initial model and including higher frequencies, source encoding can help to steer the inversion toward an improved minimum of the misfit function thanks to a broader exploration of the model space.

Assuming there is no noise in the data and a steepest descent algorithm approach, we derive some formulas for the estimation of the variance of the encoded gradient. The analysis of the variance suggests that, in absence of information about the gradients produced by each source, choosing all sources in the super source is perhaps the best strategy. However, in the knowledge of similarity amongst gradients (because of the geometric disposition, for example) perhaps more

clever strategies to create the super sources can be defined.

Total variation regularization

Chapter 4 is devoted to a few aspects concerning the regularization of the inverse problem. Indeed, several works have studied the effect of changing the norm of the data (Djikpéssé and Tarantola, 1999; Guitton and Symes, 2003; Ha et al., 2009; Pyun et al., 2009; Brossier et al., 2010), but only recently the effects of changing the norm of the regularization have been explored (Burstedde and Ghattas, 2009; Ramírez and Lewis, 2010; Anagaw and Sacchi, 2012; Guitton, 2012). We compare the effect of two different regularization norms, l_2 (equation 27) and Total Variation (equation 28), using realistic synthetic data (the BP-2004 salt model) and the Valhall real OBC data set. For the synthetic BP-2004 model we used an initial model that was good enough to avoid cycle skipping, but as far away as possible from the true model so that the regularization had more importance. Due to the reflecting boundary close to the surface, there are some artefacts that appear in that region. For both the case with and without noise, the near surface oscillations are reduced with TV regularization. Without noise, the final velocity model with the TV norm is considerably better, as shown in Figure 27. For the real data set application, the final velocity models obtained with the l_2 and TV norm are comparable, with similar overall characteristics. Nonetheless, the step-like behavior obtained using the TV norm is clear in the velocity logs. In conclusion, our results show that using the TV norm in the regularization term is appropriate for earth models, and provide finally velocity models with sharper boundaries.

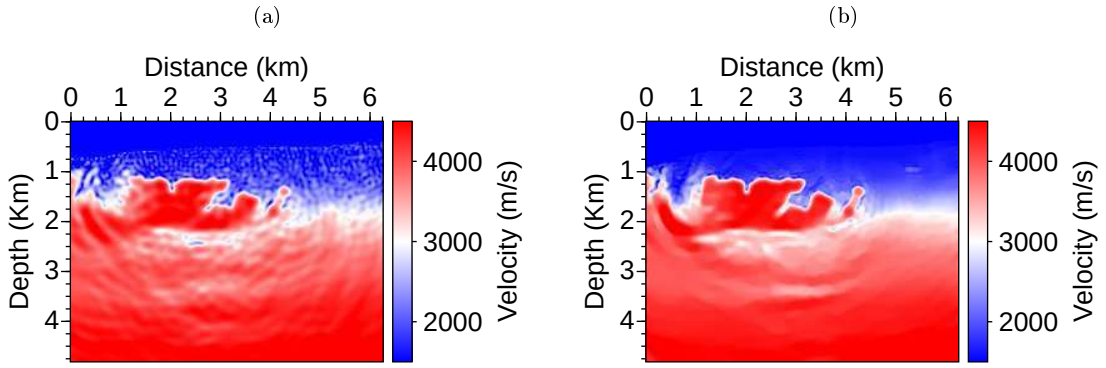


Figure 13: Data without noise. Comparison of FWI results with a regularization term with a) $\|\nabla m\|_2^2$ and b) T V regularization $\|\nabla m\|_1$. More details in Chapter 4, Section 1.4.a.

Local total variation denoising using prior migration information

The Rudin-Osher-Fatemi (ROF) denoising algorithm (Rudin et al., 1992) removes noise from an image by minimizing the total variation of the image, while trying to maintain the denoised image as similar as possible to the original image. This denoising algorithm has shown remarkable success and is quite popular (Osher et al., 2005; Caselles et al., 2011), for example in biomedical imaging. However, the weakness of TV denoising is that it removes texture (small features) of the images. Therefore, there have been works oriented towards finding local TV denoising algorithms that allow to preserve texture (Bertalmio et al., 2003; Vese and Osher, 2003).

In section 1.3 of Chapter 4, we apply the TV denoising to a final model after FWI has finished the inversion. In particular, we use a model obtained from the real Valhall data set. As is to

be expected, when too much denoising is performed, the algorithm removes small structures. One solution is to stop earlier the denoising algorithm. Alternatively, we propose a local total variation denoising by incorporating information of the reflectivity provided by a migration image. The local TV denoising algorithm does not do anything where the migration image detects reflectors, and denoises the other parts of the image. The modified TV denoising algorithm succeeds in denoising the model but preserving the important reflectors, as shown in Figure 28.

The difference of the spectral content of reflection and transmission data

To close chapter 4, we consider some aspect related to the inversion in the frequency domain in the presence of highly contrasted media that generate strong reflected waves. We provide some observations regarding the spectral content of the measured data for the BP-2004 salt model using a surface acquisition. We identify a difference in the spectrum of the short offsets, receivers close to the reflecting boundaries, and receivers far from the reflecting boundaries. We see the reflected and transmitted waves have a different spectral content, and we observe a gap in the reflected wave spectrum. The impact this has on the inversion is yet to be studied with more detail, but we believe that this gap enlarges the model null space that may be otherwise reduced by including a wider range of frequencies in the inversion.

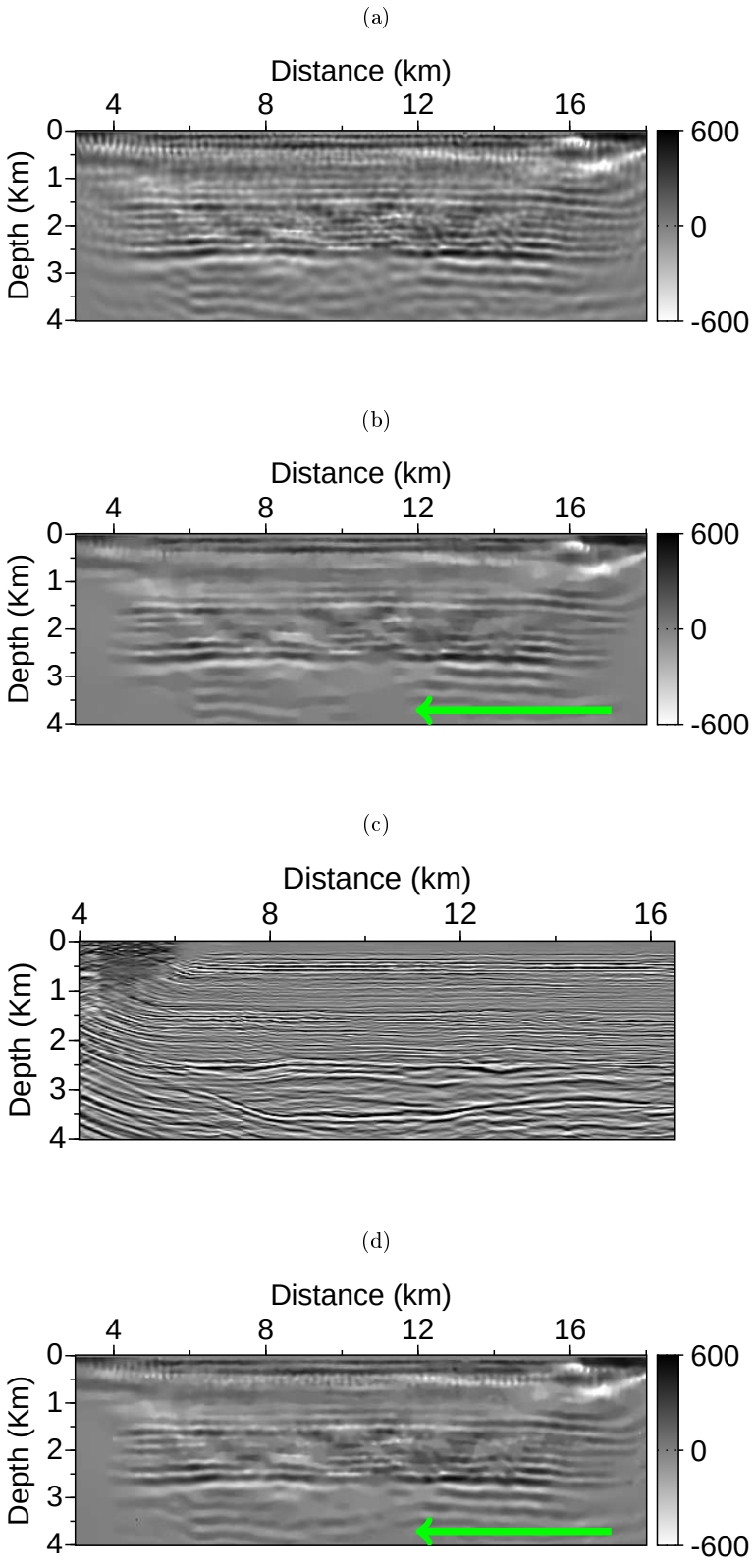


Figure 14: Illustration of the local denoising algorithm with total variation. a) Original noisy model perturbation model. b) perturbation model after TV denoising algorithm. c) Migration image. d) perturbation model after Local TV denoising algorithm with the same parameters as b) and using the information of the migration image. More details in Chapter 4, Section 1.4.b.

2 INTRODUCTION (FR)

Connaître la composition de l'intérieur de la Terre est d'un intérêt particulier pour de nombreuses applications académiques et industrielles. Pour des raisons de coûts et de faisabilité technique, il n'est pas possible d'observer directement la composition du sous-sol autrement que localement par forage. Cette difficulté a motivé le développement d'approches plus économiques et dont la mise en œuvre est plus aisée, connues sous le terme de méthodes d'imagerie géophysique. Ces méthodes constituent une panoplie d'algorithmes permettant de reconstruire la structure de l'intérieur de la Terre à partir d'enregistrements effectués au voisinage de la surface de certaines mesures physiques. Ces mesures peuvent correspondre à l'enregistrement d'ondes sismiques, d'ondes électromagnétiques, d'anomalies gravimétriques, entre autres. La qualité de l'imagerie du sous-sol dépend de la méthode d'imagerie utilisée et in fine des données enregistrées disponibles. L'objectif est par conséquent de reconstruire des paramètres décrivant les propriétés du sous-sol à partir d'enregistrements d'observables géophysiques tel qu'illustré sur la Figure 15. Cette thèse porte sur les méthodes d'imagerie géophysique fondées sur l'enregistrement des ondes sismiques.

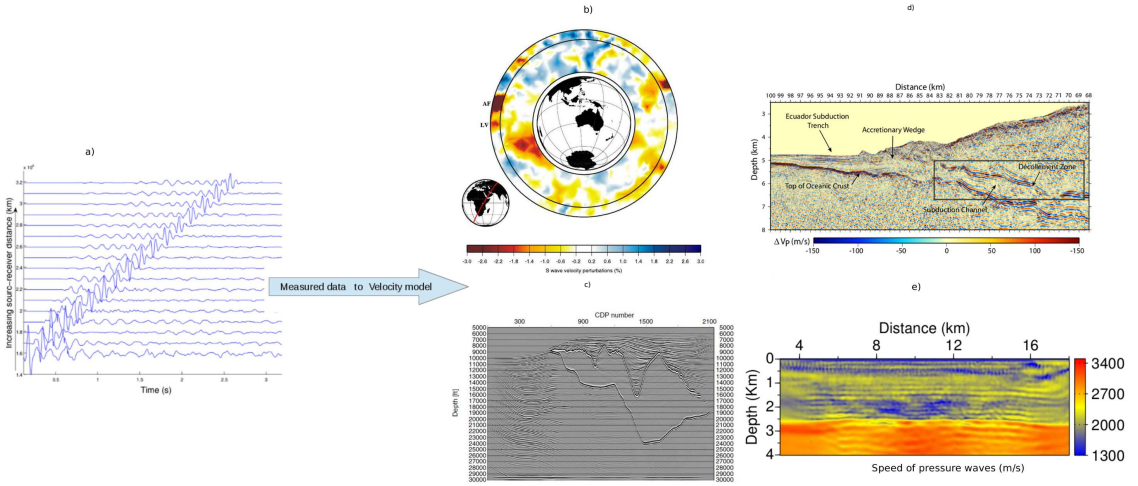


Figure 15: Les méthodes d'imagerie sismique reconstruisent les paramètres du sous-sol à partir de l'enregistrement des perturbations du sous-sol (déplacements, perturbation de pression dans l'eau) engendrées par l'arrivée des ondes sismiques. a) Enregistrements à différentes positions de capteurs (axe y) en fonction du temps de propagation (axe x), appelés sismogrammes. b) Imagerie de la vitesse de propagation des ondes cisaillantes à l'échelle globale de la Terre par tomographie des temps de trajet (Montelli et al., 2004). c) Imagerie des réflecteurs dans une zone de dôme de sels à l'échelle de l'exploration pétrolière par migration profondeur (Liu et al., 2011). d) Image migrée quantitative paramétrée par des perturbations de vitesse de propagation des ondes de compression du chenal de subduction dans le bassin de Guayaquil (Equateur) obtenue par migration/inversion rai+Born (Ribodetti et al., 2011). e) Modèle de vitesse des ondes de compression à l'échelle de la prospection pétrolière dans le champ pétrolier de Valhall par inversion des formes d'ondes complètes (cette thèse).

Acquisition sismique: transmission et réflexion

A l'échelle globale, les tremblements de terre génèrent des ondes qui se propagent dans l'intérieur de la Terre avant d'être enregistrées par des réseaux de stations sismologiques dé-

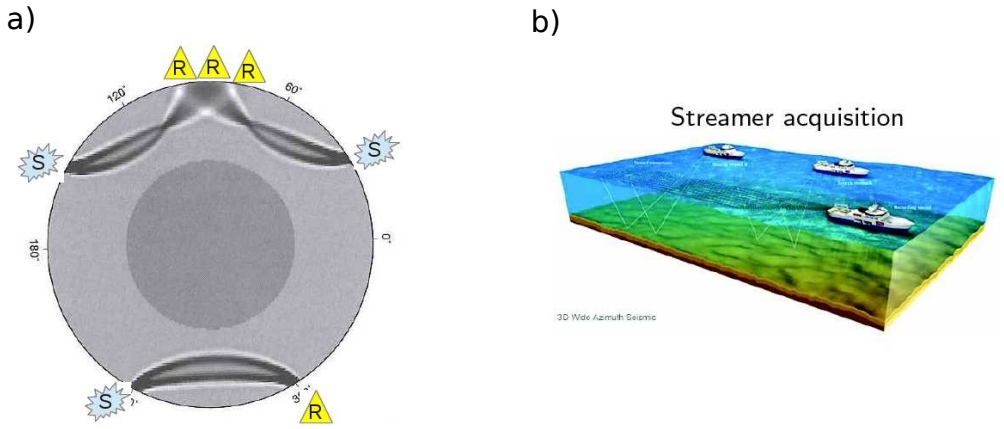


Figure 16: a) Les tremblements de terre (S) génèrent des ondes télé-sismiques qui se traversent l'intérieur de la Terre et sont enregistrées à la position des stations sismologiques (R) (adaptée de Nolet (2008)). b) Illustration d'une acquisition de sismique réflexion multitrace en prospection sismique.

ployés à la surface des continents comme illustré schématiquement sur la Figure 16a. La plupart des données sismologiques utilisées à l'échelle globale se sont propagés suivant un régime de propagation en transmission, car les ondes utilisées traversent de manière pro-grade les structures à imager avant d'être enregistrées par des stations situées à de grandes distances de la source sismique. En prospection sismique dont la finalité est la détection de ressources naturelles (pétrole, gaz, eau) en domaines marin ou terrestre, les ondes sont émises par des sources artificielles dont la position et l'heure d'émission sont parfaitement contrôlées. Un exemple d'acquisition de sismique réflexion multitrace est illustrée sur la figure 16b. Les sources sont des explosions émises depuis un navire boute feu. Les sources et les récepteurs (situés au sein d'une flûte sismique) se déplacent de manière synchrone permettant l'exploration de vastes domaines. Lorsque les sources et les capteurs sont situés au voisinage de la surface, on parlera d'acquisition de surface, la cible à imager étant située sous les dispositifs d'émission et de réception. D'autres dispositifs tels que des dispositifs d'entre-puits sont constitués d'au moins deux puits forés dans la Terre, la cible à imager étant située entre les puits. Un puits est instrumenté avec les sources, tandis que les capteurs sont déployés dans le puits en vis-à-vis. Les acquisitions de surface et d'entre-puits fournissent des données de nature différente. Les acquisitions d'entre puits enregistrent principalement des ondes transmises et des ondes réfléchies avec des angles de réflexion élevés, tandis que les acquisitions de surface enregistrent principalement les ondes réfléchies car les sources et les récepteurs sont situés du même coté relativement à la cible d'étude. Néanmoins, si les acquisitions de surface disposent de forts déports horizontaux (offset) entre les sources et les capteurs, les données contiennent aussi des arrivées transmises correspondant aux ondes plongeantes, les ondes réfractées ou coniques et les ondes réfléchies aux incidences super-critiques. Cette configuration est celle des acquisitions dite grand-angle pour lesquelles le dispositif de capteur est fixe par rapport au dispositif de sources (fixed-spread acquisition en Anglais) correspondant aux acquisitions de fond de mer en domaine marin ou à certaines acquisitions terrestres. Par conséquent, la géométrie du dispositif d'acquisition détermine la nature des ondes enregistrées et la proportion d'arrivées transmises et réfléchies.

Un enregistrement sismique (collection de sismogrammes à récepteur commun) enregistré du-

rant une campagne d'acquisition de fond de mer (avec des câbles de fond de mer) sur le champ pétrolier de Valhall en mer du Nord est présenté sur la Figure 17a. Un récepteur, dont la position est indiquée par l'offset 0 km en abscisse, enregistre une collection de sismogrammes générés par chaque source disposée le long d'un profil situé au droit du dispositif de capteurs avec une gamme d'offsets s'échelonnant entre -12km et +12km. L'axe vertical est le temps de propagation depuis le temps d'émission de la source. La Figure 17b montre un modèle physique du sous-sol en deux dimensions, où l'échelle de couleur représente la vitesse de propagation des ondes de compression (v_p) dans le sous-sol. Les principaux trajets suivis par les ondes entre les source et le capteur sont figurés par des rayons, trajectoires perpendiculaires aux fronts d'ondes. Les trajets blancs correspondent à des ondes plongeantes (ou ondes transmises), les trajets rouges et bleus représentent les ondes réfléchies au toit et à la base d'une zone à faible vitesse correspondant à une accumulation de gaz. Pour la gamme d'offsets enregistrés, il n'y a pas d'ondes transmises à des profondeurs supérieures à $\approx 2km$. La Figure 17c montre le sismogramme enregistré pour un tir situé approximativement à $x = 2.5km$. La première arrivée dans le sismogramme correspond à une onde transmise (trajet blanc sur la Figure 17b) et les arrivées tardives correspondent principalement aux ondes réfléchies (trajets rouges et bleus sur la Figure 17b). La Figure 17a montre que les temps de trajet des ondes transmises satisfont au premier ordre l'équation d'une droite $t = o/v$ (o est l'offset et v est la vitesse de propagation) dont la pente donne une information sur la vitesse moyenne du milieu dans laquelle l'onde s'est propagée. Les temps de trajet des ondes réfléchies satisfont au premier ordre l'équation d'une hyperbole $t = \sqrt{o^2/4 + z^2}/v$ (z est la profondeur du réflecteur) qui dépend de la vitesse moyenne du milieu situé au dessus du réflecteur et de la profondeur du réflecteur. A grand-offset (aux incidences super-critiques), l'équation de l'hyperbole tend vers l'équation d'une droite fournissant la vitesse v tandis qu'à offset nul il y a une indétermination entre la vitesse et la profondeur de l'interface. Cette analyse schématique illustre les deux échelles caractéristiques généralement considérées en prospection sismique: les grandes longueurs d'onde du milieu représentées par les vitesses de propagation des ondes v et les courtes longueurs d'onde associées aux réflecteurs paramétrés ici par le paramètre z. Les approches d' imagerie reposent généralement sur des approches hiérarchiques alternant la mise en œuvre des vitesses par tomographie et l'imagerie des réflecteurs par migration.

Méthodes d'imagerie

La géométrie du dispositif d'acquisition contrôle la nature de l'information contenue dans les données sismiques et différentes méthodes d'imagerie sismiques sont conçues pour utiliser une partie spécifique de cette information. A l'échelle globale, les données sont fortement dominées par le régime de propagation en transmission et l'attribut sismique le plus naturellement utilisé dans ce contexte d'imagerie globale est le temps d'arrivée de la première arrivée (l'onde P). Dans ce cadre applicatif de petites perturbations, la tomographie par inversion des temps de trajet repose sur l'hypothèse qu'un modèle de référence est connu et de petites perturbations de ce modèle de référence sont recherchées telles que les différences entre les temps observés et calculés dans m_0 soient minimisées. Un modèle global de perturbation des vitesses de propagation des ondes de cisaillement est présenté sur la Figure 15b a titre d'illustration. A contrario, quand l'imagerie est effectuée à l'échelle régionale à partir d'un dispositif d'acquisition de surface, les données sont dominées par les ondes réfléchies. Ce type de données sont majoritairement traitées par des méthodes de migration. Les sismogrammes sont pré-traités pour éliminer les ondes directes et réfractées ainsi que les réflexions multiples pour ne conserver dans les données que les ondes réfléchies (primaires). Les méthodes de migration fournissent une image des réflecteurs (Figure 15c) sur lesquels les ondes se sont réfléchies. Des méthodes de migration quantitative formulées dans le cadre théorique des problèmes inverses fournissent également des informations sur l'amplitude des perturbations δm au niveau des réflecteurs ou sur la réflectivité. A titre d'illustration, la Figure 15d présente un modèle de perturbation des vitesses de propagation des ondes P obtenu par migration asymptotique rai+Born (Lambaré et al., 1992). Que ce soit à

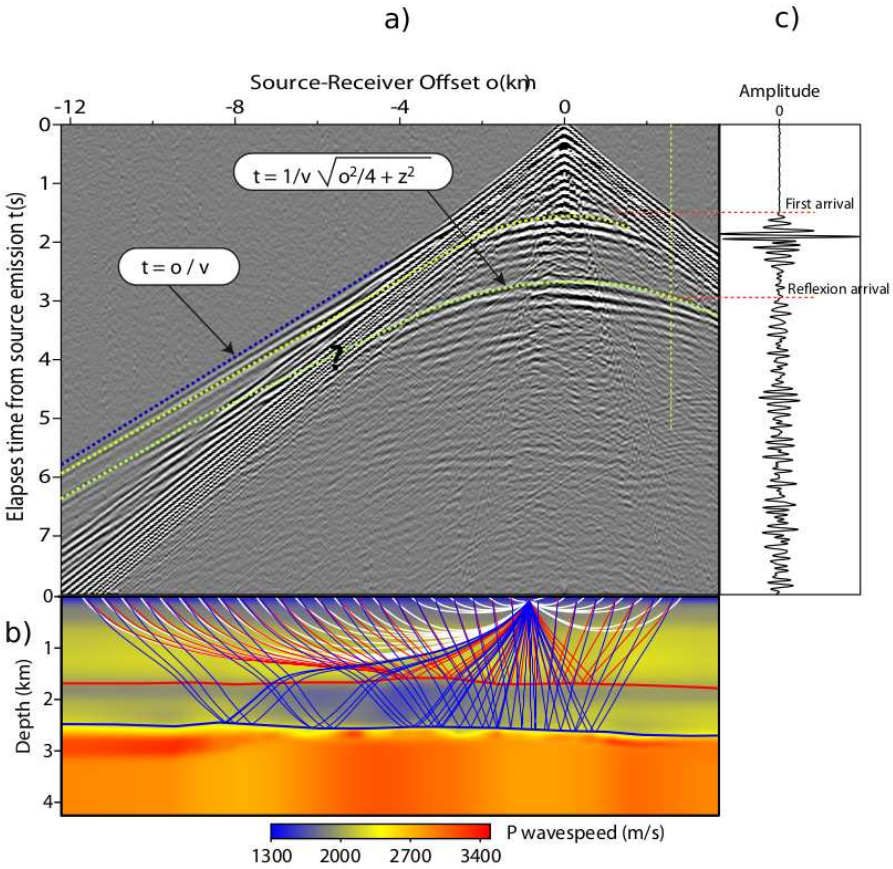


Figure 17: a) Un exemple de collection de sismogrammes enregistrés par un capteur d'une campagne de fond de mer sur le champ de Valhall. b) Un modèle de vitesse lisse du sous-sol sur lesquels sont superposés plusieurs trajets d'ondes. c) Exemple d'un sismogramme enregistré par le capteur pour un tir de l'expérience.

l'échelle globale ou régionale, il est aussi théoriquement possible d'utiliser l'information contenue dans la totalité du champ d'onde (ondes transmises et réfléchies). Un exemple de modèle de vitesse des ondes P à l'échelle régionale obtenu par inversion du champ d'onde total est présenté sur la Figure 15e. Ce type de méthode d'imagerie est généralement appelée inversion des formes d'ondes complètes ("full waveform inversion" en anglais) (Lailly, 1983; Tarantola, 1984a). Les différentes images présentées jusqu'alors diffèrent notamment par leur contenu spectral. Les modèles obtenus à partir des ondes transmises représentent les grandes longueurs d'onde du milieu, tandis que les images migrées fournissent une image des courtes longueur d'onde. L'inversion des formes d'ondes complètes vise à combler la séparation d'échelles existant entre ces deux gammes de longueurs d'onde sous réserve que la géométrie du dispositif d'acquisition et la bande passante de la source soient adaptées et que l'imagerie puisse être poussée jusqu'à des fréquences suffisamment élevées.

Cette thèse porte sur l'étude de la méthode d'inversion des formes d'ondes complètes(**FWI**) (**?**Tarantola, 1984a; Virieux and Operto, 2010). La FWI est un problème inverse non linéaire visant à minimiser une distance entre les données observées et les données modélisées par résolution complète de l'équation d'onde.

$$\min_m \phi_0 = \min_m \|Pu(m) - d\|_2^2, \tag{17}$$

où u est le champ d'onde modélisé, d représente les données enregistrées et m sont les paramètres décrivant les propriétés du sous-sol que l'on cherche à imager. L'opérateur de projection P qui extrait la valeur du champ d'onde modélisé aux positions de l'espace où les données ont été enregistrées. Le champ d'onde modélisé est solution de l'équation d'onde. Par exemple, dans le cas de l'approximation acoustique, u est solution de

$$\left(\nabla^2 - \frac{1}{v^2}\frac{\partial^2}{\partial t^2}\right) u(x,t) = s(x,t) \qquad (x,t) \in \Omega \times [0,\infty), \tag{18}$$

où $v(x) = 1/\sqrt{\rho(x)\kappa(x)}$, κ est le module d'incompressibilité, ρ est al densité, v est la vitesse de propagation des ondes P. Ici, les paramètres m représentent aux différentes positions de l'espace une seule propriété physique, la lenteur au carré: $m = \{1/v^2(x)\}$. La résolution de l'équation (18) est dénommée problème direct, et la résolution de l'équation (17) est le problème inverse. La FWI nécessite la résolution de ces deux problèmes de manière alternée. Le problème direct fournit u pour un modèle initial m donné, et ce champ d'onde u est utilisé dans un deuxième temps par le problème inverse pour mettre à jour le modèle du sous-sol. En d'autres termes, le problème direct consiste à résoudre une équation aux dérivés partielles (EDP), et le problème inverse estime les paramètres contenus dans les coefficients de cette EDP à partir de ses solutions. Ce processus est répété de manière itérative jusqu'à ce que les valeurs du champ d'onde u propagé dans le modèle m aux positions des capteurs coïncident avec les données enregistrées d. Cette séquence de taches est représentée schématiquement sur la Figure 18.

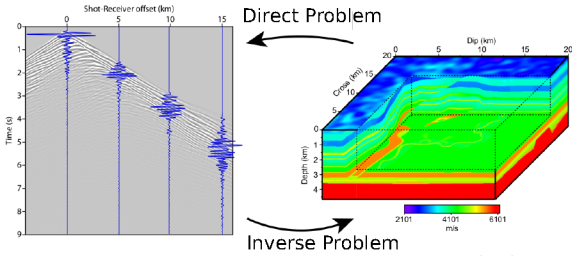


Figure 18: FWI résous de manière itérative le problème direct (18) et le problème inverse (17). Le problème direct permet le passage de l'espace des modèles à l'espace des données en utilisant le modèle du sous-sol m pour résoudre l'équation d'onde fournissant u. Le problème inverse permet le passage de l'espace des données à l'espace des paramètres en utilisant le champ d'onde u pour déterminer le modèle du sous-sol qui minimise un écart de formes d'ondes ϕ_0.

Lorsque la totalité du champ d'onde est injectée dans le processus d'optimisation, la FWI a potentiellement la capacité de fournir des modèles du sous-sol plus réalistes en termes de résolution et de caractérisation physique que les autres méthodes d'imagerie. Ses fondements théoriques ont été établis en géophysique dans les années quatre vingt (Lailly, 1983; Tarantola, 1984a), et depuis la compréhension de ses potentialités et de ses limites s'est affinée au gré des applications réalisées grâce aux moyens modernes d'acquisition et de calcul haute performance (Virieux and Operto, 2009). Les challenges méthodologiques à surmonter pour une mise en œuvre pertinente de la FWI ont néanmoins été clairement posés dès les premières études publiées sur le sujet. Par exemple, les premières analyses sur des cas synthétiques indiquent que le coût calcul de la FWI constitue un frein important à son application à des cas d'étude réels (Gauthier et al., 1986; Crase et al., 1990; Luo and Schuster, 1991; Crase et al., 1992). Plus fondamentalement, contrairement à la plupart des autres méthodes d'imagerie, la FWI est un problème inverse non linéaire et les premiers tests numériques montrent que le processus d'optimisation non linéaire échoue souvent car la fonctionnelle à minimiser ϕ_0 est fortement non convexe faisant converger l'inversion vers un minimum local inacceptable quand le modèle initial est trop éloigné du modèle

réel (Gauthier et al., 1986; Mora, 1989; Luo and Schuster, 1991). Bien que la FWI soit supposée reconstruire un large spectre de longueurs d'onde du modèle à partir des champs d'ondes réfléchis et transmis, les premiers essais (Devaney, 1984; Wu and Toksöz, 1987; Mora, 1988, 1989) ont uniquement fourni une image de la réflectivité c'est-à-dire des courtes longueurs d'onde tel que l'aurait fourni une méthode de migration. Cela motiva plusieurs analyses de résolution de la FWI pour comprendre quelle partie du spectre des nombres d'onde dans l'espace des modèles pouvait être reconstruit pour une géométrie d'acquisition et pour une bande passante de source (Jannane et al., 1989; Mora, 1989). D'autres difficultés inhérentes aux inversions multi-paramètres sont apparues. Les résultats montrent que la résolution avec laquelle chaque classe de paramètre est reconstruire varie d'un paramètre à l'autre en fonction de l'illumination angulaire locale fournie par la géométrie du dispositif d'acquisition, les couplages entre paramètres et l'influence variable de chaque paramètres dans les résidus des données (Tarantola et al., 1984; Mora, 1987)[3]. Ces difficultés furent identifiées lors des investigations pionnières de Gauthier et al. (1986); Wu and Toksöz (1987); Mora (1988, 1989); Luo and Schuster (1991), et restent aujourd'hui des sujets de recherche d'actualité en FWI.

Dans cette thèse, j'aborderai certains aspects de la FWI portant sur la réduction de son coût calculatoire et de sa non linéarité. Mais avant d'entre plus en détail dans mon travail de thèse, je présente un panorama général sur les principales caractéristiques de la FWI.

2.1 L'inversion des formes d'ondes complètes : le défi de l'imagerie sismique non linéaire.

Condition d'imagerie

La condition d'imagerie peut être défini comme l'opérateur agissant de l'espace des données vers l'espace des modèles et fournissant l'information sur les paramètres ayant une influence sur le champ d'onde calculé. Par conséquent, la condition d'imagerie révèle l'ensemble des paramètres qui, soumis à de faibles perturbations, génèrent des perturbations dans les données. A titre d'illustration, considérons une arrivée réfléchie dans les données d enregistrée au temps $t = T$, qui n'est pas prédite par les données modélisées u. La contribution de la condition d'imagerie est de déterminer les positions de l'espace où le modèle du sous-sol doit être modifié tel que le champ d'onde modélisé projeté aux positions des récepteurs contienne cette arrivée réfléchie. Pour un couple source-récepteur (S-R), la seule condition est que la position x du paramètre permette de vérifier l'équation

$$t_{S,x} + t_{x,R} = T. \tag{19}$$

Pour un modèle de référence homogène de vitesse v, toutes les positions x qui permettent de vérifier (19) sont situés sur une ellipse dont les points focaux sont situés aux positions de la source et du capteur, comme illustré sur la Figure 19a. Cela se montre aisément en écrivant explicitement le temps de trajet d'une onde se propageant de S à x, $t_{S,x}$, et de R à x, $t_{x,R}$,

$$\begin{aligned} t_{S,x} &= \sqrt{(x/v)^2 + (z/v)^2} \\ t_{x,R} &= \sqrt{(D-x)/v)^2 + (z/v)^2}, \end{aligned}$$

où D est la distance entre la source et le capteur. Substituer ces expressions dans (19) fournit l'équation suivante de l'ellipse,

$$\sqrt{x^2 + z^2} + \sqrt{(D-x)^2 + z^2} = vT,$$

3. La paramétrisation du sous-sol est ici définie comme un jeu de paramètres physiques indépendants caractérisant les propriétés du sous-sol.

dont les points focaux sont aux positions $f = 0, D$, le grand axe est de longueur $a = vT/2 + D/2$ et le petit axe de largeur $b = \sqrt{(vT)^2 + D^2}/2$.

En résumé, pour un couple source-récepteur et une arrivée réfléchie enregistrée au temps T, tout point de l'ellipse représente une position potentielle où serait situé le diffractant réel ayant généré l'onde réfléchie car tous ces points permettent de vérifier (19). Sur l'exemple de la Figure 19a, $v = 2m/s$ le temps d'arrivée de la réflexion $T = 5s$, $x_S = 0$ et $x_R = 8\ m$, donnant un déport source-récepteur de $D = 8m$. Pour éliminer cette ambiguïté sur la position du réflecteur, la sommation des ellipses associées à tous les couples source-récepteur la position du réflecteur par interférence constructive de toutes les ellipses à cette position. Ce processus de construction par interférences constructives est illustré sur la Figure 19b pour un réflecteur horizontal.

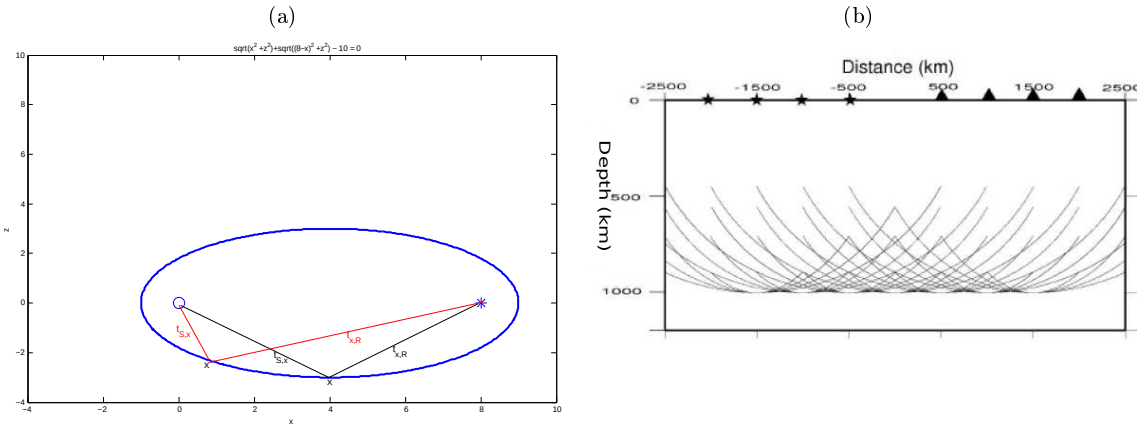


Figure 19: a) Exemple de points diffractants vérifiant la condition d'imagerie $t_{S,x} + t_{x,R} = T$, pour une onde réfléchie enregistrée au temps $T = 5s$, dans un modèle homogène de vitesse $v = 2m/s$, pour une source située à $x_S = 0$ et un récepteur situé à $x_R = 8\ m$. Les points forment une ellipse dont les points focaux coïncident avec la position du tir et du capteur. b) Figure tirée de Agudelo (2005). Imagerie d'un réflecteur horizontal. L'interférence constructive des ellipses à la position du réflecteur forme l'image de celui-ci.

La condition d'imagerie (19) dans l'espace des modèles peut être écrite de manière équivalente, pour une source et un ensemble de capteurs comme,

$$g(x) = \frac{\partial \phi_0(u(x,t,m))}{\partial m} = \left(P\frac{\partial u}{\partial m}\right)^{\dagger}(Pu(x,t,m) - d(x,t)) \quad (20)$$

$$= -\int_0^T \frac{\partial^2 u(x,t,m)}{\partial t^2}\lambda(x,T-t,m)dt, \quad (21)$$

où g est le gradient de la fonctionnelle, u est la solution du problème direct, $\lambda(T-t)$ est le champ d'onde rétro-propagé qui est calculé en utilisant les données résiduelles aux positions des récepteurs comme une source. Par construction, la condition d'imagerie correspond à la corrélation pondérée à décalage nul du champ direct avec le champ rétro-propagé en tout point du milieu, intégré sur le temps. La sommation sur le temps peut être représentée de manière équivalente par une sommation sur les fréquences,

$$g(x) = \int_{-\infty}^{\infty} \omega^2 u(x,\omega,m)\lambda^*(x,\omega,m)d\omega. \quad (22)$$

Plus de détails sur l'interprétation de la condition d'imagerie sont fournis dans le chapitre 2, Section 1.2.b. La condition d'imagerie de la FWI est analogue à celle d'autres méthodes

d'imagerie reposant sur une procédure de renversement temporel, comme cela est expliqué dans l'annexe 1.

2.1.a *Analyse de résolution*

D'un point de vue théorique, la qualité des images fournies par la FWI sera supérieure à celle fournie par d'autres méthodes d'imagerie car la FWI utilise toute l'information contenue dans les formes d'onde et le modèle du sous-sol utilisé pour résoudre le problème direct (18) est mis à jour à chaque itération. La percée fournie par la FWI réside dans la manière dont l'équation d'onde est résolue. Sans puissance de calcul suffisante, l'équation d'onde était résolue analytiquement ou asymptotiquement avec la théorie des rais qui nécessite des modèles lisses. Il était alors nécessaire d'utiliser une approximation linéarisée de l'équation d'onde et de séparer le modèle du sous-sol m en un modèle de référence lisse m_0 et un modèle de petites perturbationsδm : $m = m_0 + \delta m$. Une fois cette séparation d'échelle effectuée, la solution de l'équation d'onde est trouvée par la théorie des perturbations à l'ordre 1, connue sous le nom de l'approximation de Born du premier ordre. Une approximation du champ d'onde total u est défini comme $u(m_0 + \delta m) \approx u_0 + (\partial u/\partial m)\delta m$, où u_0 est la solution de l'équation d'onde dans le modèle m_0 et $(\partial u/\partial m)\partial m$ est le champ d'onde diffracté par la perturbation du modèle δm.

Notons que l'approximation de Born d'ordre 1 néglige les termes de diffraction multiple. La distinction explicite entre m_0 et δm induit une séparation d'échelles entre le modèle de référence lisse et le modèle de perturbations, comme illustré sur la Figure 20a,b. Quand les ressources de calcul devinrent suffisantes, l'équation d'onde pu être résolue de manière "exacte" avec des méthodes numériques volumétriques telles que la méthode des différences finies pour des modèles complexes du sous-sol d'hétérogénéité arbitraire (**?**). En s'affranchissant du besoin de linéarisable l'équation d'onde, la séparation d'échelle entre le modèle de référence et le modèle de perturbations devient caduque car la propagation des ondes peut être calculée dans des modèles de complexité arbitraire. Si la géométrie d'acquisition permet l'enregistrement d'ondes transmises et réfléchies et si la source a une bande passante suffisamment large, FWI a le pouvoir de résolution théorique pour imager un spectre large et continu de nombres d'ondes. En pratique, il s'est néanmoins révélé difficile de mettre à jour simultanément toutes les composantes spectrales du modèle du sous-sol (Devaney, 1984; Wu and Toksöz, 1987; Mora, 1989). Une description plus détaillée de cette analyse de résolution est fournie dans le chapitre 2, Section 1.5.

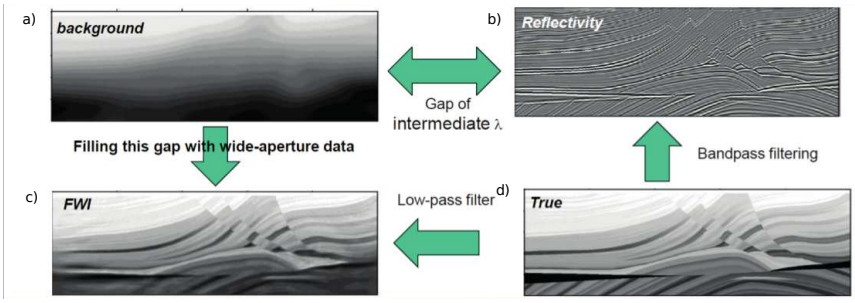


Figure 20: Illustration schématique des différentes échelles possibles impliquées en exploration sismique. (d) Le modèle exact du sous-sol m. Dans des approches linéarisables, le modèle m est décomposé en un modèle de référence lisse m_0 (a) et un modèle de perturbation δm (b). La FWI ne requiert pas cette séparation d'échelle si les données contiennent à la fois des arrivées transmises et réfléchies. Dans ce cas, le modèle du sous-sol reconstruit (c) est une version passe bas du vrai modèle (d).

2.1.b *Coût de calcul*

Bien que la puissance de calcul soit devenue suffisante dans les années 80 pour résoudre l'équation d'onde pour des modèles du sous-sol d'hétérogénéité arbitraire, par des méthodes de différences finies dans le domaine temps-espace notamment, les applications de la FWI sur des cas d'étude réels restaient hors de portée.

Méthode de l'état adjoint

Un premier pas ayant permis de réduire considérablement le coût de la FWI a résidé dans l'utilisation de la méthode de l'état adjoint pour calculer le gradient de la fonctionnelle (Lailly, 1983; Tarantola, 1984a).

L'équation (20) nécessite le calcul des dérivés partielles du champ d'onde modélisé par rapport à la totalité des paramètres du modèle, $i = 1, ..., N \times N_p$. Ici, N représente le nombre de points de grille dans le domaine de calcul et N_p le nombre de classes de paramètres à reconstruire, par exemple, $(\rho, v_p, v_s) \rightarrow N_p = 3$. La méthode des différences finies fournit une approche brutale pour calculer les dérivés partielles, $\partial u/\partial m_i = (u(m + \Delta m_i) - u(m)/\Delta m_i$, impliquant un coût calcul proportionnel à $C = 2 \times \mathbf{N} \times \mathbf{N_p} \times N_s$, où N_s est le nombre de sources sismiques. Avec la méthode de l'état adjoint (Lailly, 1983; Tarantola, 1984a), le gradient de la fonctionnelle est calculé avec l'équation (21) pour un coût proportionnel à $C = 2 \times N_s$. En effet, l'implémentation de l'équation (20) nécessite deux résolutions de l'équation d'onde par source, une pour calculer le champ d'onde incident u et une autre pour calculer le champ adjoint rétro-propagé λ. Le calcul du gradient avec la méthode de l'état adjoint est détaillé dans les annexes 3 et 4.

La FWI dans le domaine fréquentiel

Une deuxième étape vers la réduction du coût de la FWI a été franchie en formulant la FWI dans le domaine fréquentiel. Dans les années 90s, le formalisme de la FWI, tel qu'élaboré par Tarantola (1984a), a été transformé dans le domaine fréquentiel pour les équations d'ondes acoustique et élastique (Pratt and Worthington, 1990; Pratt, 1990; Pratt and Shipp, 1999). Appliquer une transformée de Fourier par rapport au temps à l'équation d'onde transforme un problème d'évolution à un problème stationnaire de conditions aux limites pour chaque composante fréquentielle à modéliser. L'équation d'onde dans le domaine fréquentiel est une forme généralisée de l'équation d'Helmholtz pouvant s'écrire sous forme matricielle comme

$$A(m, \omega)u(s, m, \omega) = s(\omega), \tag{23}$$

où les coefficients de $A(m, \omega)$, matrice résultant de la discrétisation de l'opérateur de l'équation d'onde en domaine fréquentiel, dépendent de la fréquence $\omega = 2\pi f$. L'équation (23) est un système d'équations linéaires où A est une matrice carrée de dimensions $N \times N$ tandis que u et s sont des vecteurs de dimension N. Ce système linéaire peut être résolu par des méthodes directes ou itératives. En ce qui concerne les méthodes itératives, le nombre d'itérations nécessaires à l'estimation de la solution dépend du conditionnement de la matrice. Par conséquent, la conception de pré-conditionneurs performants sont nécessaires et constitue la difficulté de ces approches (Erlangga, 2005; Plessix, 2007; Erlangga and Nabben, 2008). Chaque source sismique (terme de droite) nécessite une nouvelle résolution itérative du système, indépendante de celles effectuées pour les sources précédentes. A contrario, les méthodes directes reposent sur des techniques d'élimination de Gauss telles que les approches fondées sur des décompositions triangulaires supérieure/inférieure de la matrice ($A = LU$). La factorisation LU de la matrice dans l'équation (23) dépend uniquement de m et de la fréquence ω. Par conséquent, pour chaque itération non-linéaire du problème inverse, une seule factorisation LU est effectuée par fréquence pour calculer le gradient. Les champs d'onde calculés pour chacune des sources sont efficacement

calculés à partir des facteurs LU par substitutions directe et inverse. Si l'inversion peut être limitée à quelques fréquences discrètes, des analyses de complexité montrent que les approches par solveur direct sont les plus efficaces (au moins en deux dimensions) en raison de l'efficacité des phases de substitutions et du nombre limité de factorisation à effectuer. Pour des applications de la FWI le gain peut être considérable (Pratt and Worthington, 1990; Pratt, 1990) car le nombre de sources pour des expériences 2D et 3D est respectivement de l'ordre de $10^2 - 10^3$ et de $10^3 - 10^4$. Les domaines d'application privilégiés de l'inversion en domaine fréquentiel concernent les acquisitions dites grand-angle pour lesquelles une analyse de résolution permet de montrer que seules quelques fréquences discrètes peuvent être injectées dans l'inversion grâce à la redondance d'illumination fournie par les fréquences temporelles et les angles d'ouverture (Sirgue and Pratt, 2004). A contrario, les géométries d'acquisition en réflexion fournissent un éclairage angulaire du milieu beaucoup plus étroit qui doit être compensé par un échantillonnage plus fin des fréquences dans la FWI rendant l'approche fréquentielle moins pertinente pour ces configurations (Freudenreich and Singh, 2000).

Au delà du bénéfice fourni par le coût réduit du problèmes direct, les approches fréquentiels permettent la manipulation de volumes de données plus faibles lorsque quelques fréquences discrètes sont inversées. Dans le domaine temporel, le calcul du gradient nécessite la corrélation des champs directs et adjoints à chaque pas de temps t. Comme le champ adjoint rétro-propagé λ ne peut être calculé qu'après que le champ incident ait été complètement calculé (car le terme de source du champ adjoint est formé par les résidus entre les données enregistrées et les valeurs aux positions des capteurs des champs directs modélisés), deux approches sont possibles pour calculer ces corrélations. Une première approche consiste à stocker en mémoire le champ direct à tous les temps $u(x,t), x \in \Omega$, où N_t est le nombre de pas de temps. La deuxième consiste à recalculer le champ incident durant le calcul du champ adjoint soit par de manière pro-grade avec des techniques de point de reprise soit de manière rétrograde à partir des valeurs du champ incident stocké aux frontières du domaine de calcul. La première méthode est rarement utilisé, en particulier en 3D, en raison de son coût de stockage prohibitif. La deuxième approche nécessite de recalculer au moins une fois le champ incident. En domaine fréquentiel, le gradient se résume à un produit pondéré de deux champs d'onde monochromatiques dont le stockage en mémoire ne pose pas de problème. Plus de détails sont fournis dans le chapitre 2, Section 1.4.

Assemblage et encodage de sources

Les limites des approches fréquentielles fondées sur des solveurs directs résultent du calcul et du stockage des facteurs LU. Pour des méthodes aux différences finies, la matrice A est bande diagonale. En 2D, il y a trois bandes et la distance entre bandes est proportionnelle à N. Pour des applications 3D, il y a cinq bandes et la distance entre bandes est proportionnelle à N^2. Contrairement à A, les matrices L et U sont pleines et leur stockage est par conséquent coûteux. Pour des applications acoustiques 3D de dimension réduite, la faisabilité des approches directes a été montrée (Operto et al., 2007). Les premières applications de FWI 3D acoustique en domaine fréquentiel ont été réalisées en domaine fréquentiel à partir de méthodes itératives (Plessix, 2009; Plessix and Perkins, 2010) ou de modélisation temporelles pour le problème direct (Sirgue et al., 2010). Dans ce dernier cas, les champs monochromatiques sont extraits à la volée dans la boucle sur les pas de temps par transformée de Fourier discrète. Pour des applications élastiques 3D, l'approche par solveur direct semble aujourd'hui inaccessible en raison de la nature vectorielle de l'équation d'onde élastique nécessitant le calcul d'au moins trois composantes de vitesse particulaires et des faibles longueurs d'ondes propagées en relation avec la vitesse de propagation des ondes de cisaillement. Dans cette configuration, seules les modélisations en domaine fréquentiel avec des solveurs itératifs ou des modélisations en domaine temporel semblent possibles.

Une réduction du coût calcul est envisageable si le nombre de problèmes directs est diminué.

Comme pour des approches fréquentielles fondées sur des solveurs itératifs ou des modélisations temporelles, le coût des problèmes directs est proportionnel au nombre de sources (équations (18) et (23)), la manière la plus naturelle de réduire le coût est de diminuer le nombre de sources. Lors des acquisitions sismiques, cette approche est implémentée en mélangeant les sources: plusieurs sources sont émises simultanément ou avec un décalage entre elles, et le champ d'onde enregistré contient la contribution mélangée de chaque source en vertu de la linéarité de l'équation d'onde par rapport au terme de source (Beasley, 2008; Berkhout, 2008). Cette technique d'assemblage des sources sismiques réduit considérablement les temps d'acquisition mais aussi le nombre de sources à considérer lors de la résolution du problème direct en FWI. Une autre méthode courant d'accélération, appelée encodage des sources, consiste à effectuer une acquisition sismique conventionnelle (en émettant une source à la fois). Néanmoins, au lieu de résoudre le problème direct, (18) ou (23), pour chaque source, des super-sources sont formées par combinaison linéaire des sources pondérées par des facteurs aléatoires (Romero et al., 2000; Neelamani et al., 2008; Krebs et al., 2009; Ben Hadj Ali et al., 2011; Schuster et al., 2011; van Leeuwen et al., 2011; Huang and Schuster, 2012),

$$\tilde{s} = \sum_{i=1}^{N_s} \alpha_i s_i, \qquad \tilde{d}_o = \sum_{i=1}^{N_s} \alpha_i d_{o_i}. \tag{24}$$

où α désigne des coefficients aléatoires. Les données observées sont encodées de manière analogue, $\tilde{d}_o$ Pour encoder les données de la sorte, il est nécessaire que les sources soient enregistrées par le même ensemble de récepteurs, désignant ainsi les acquisitions d'extension fixe (fixed-spread en Anglais) comme les acquisitions de fond de mer. Le volume de données généré par N_s sources et enregistré par N_r récepteurs est compressé en une seule collection de traces à tir commun.

Même si la méthode d'encodage des sources couplée à des algorithmes optimisation de gradient permet de réduire le coût calcul par itération (2 problème directs au lieu de $2 \times N_s$ problèmes directs), plus d'itérations de l'optimisation doivent être effectuées car chaque mise à jour du modèle du sous-sol est moins précise en raison des interférences entre les différences sources élémentaires formant la super-source (dénommées cross-talk en Anglais), comme cela est illustré sur la Figure 21a. In fine, un gain en termes de coût ou accélération est obtenu si le nombre total de problèmes directs nécessaires pour atteindre une valeur prédéfinie de la fonction coût est plus faible lorsque l'encodage des sources est utilisé, comme cela est illustré sur la Figure 21b. La méthode d'encodage de sources a été appliquée avec succès sur des données réelles par Baumstein et al. (2011); Routh et al. (2011); Bansal et al. (2013); Schiemenz and Igel (2013).

2.1.c *Optimisation non linéaire*

Les applications pionnières de la FWI ont révélé des difficultés supplémentaires associées au problème d'optimisation malgré la précision accrue du problème direct et la faisabilité théorique de la FWI (Gauthier et al., 1986; Luo and Schuster, 1991). La relation entre les données et les paramètres est non linéaire. Cela signifie que de petites modifications des paramètres du modèle peuvent générer de fortes perturbations des données et de la fonction coût ϕ_0, ces variations se comportant de façon non linéaire. Lors d'une inversion acoustique de données entre-puits, Luo and Schuster (1991) ont comparé l'inversion des temps de trajet de la première arrivée et la FWI. Les résultats ont montré le pouvoir de résolution supérieur de la FWI. Néanmoins, l'inversion des temps de trajet a fourni des résultats plus robustes car la FWI convergeait vers des minimums secondaires.

Optimisation non convexe

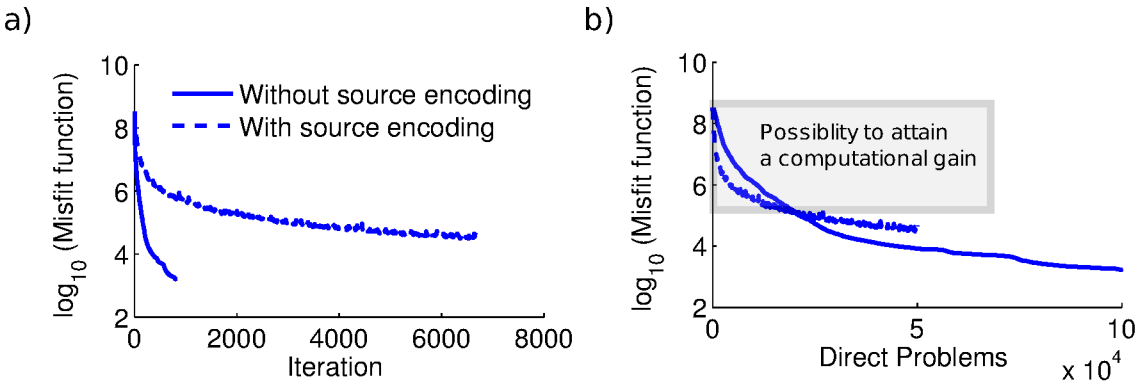


Figure 21: Courbes de convergence de la FWI avec encodage de source (lignes en tireté) et sans encodages de sources (lignes continues). (a) Comparaison entre les taux de convergences. Moins d'itérations sont nécessaires pour atteindre le niveau de convergence souhaité quand l'encodage des sources n'est pas utilisé. (b) Comparaison du coût calcul (mesuré par le nombre de problèmes directs). Si la réduction souhaitée de la fonction coût reste confinée dans le rectangle gris, un gain est obtenu avec l'encodage des sources. Cet exemple est détaillé dans le chapitre 3.

Les difficultés rencontrées lors de la minimisation de la norme L2 de l'écart des formes d'ondes (17) sont illustrées sur la Figure Figure 22. La tomographie des temps de trajet définit la fonction coût ϕ_0 comme la norme L2 de la différence entre les temps observés et calculés.

$$\min_m \phi_0 = \min_m \|T_c(m) - T_o\|_2^2, \tag{25}$$

où T_o représente le temps d'arrivée observé et T_c le temps calculé (voir l'annexe 2 pour une description plus détaillée). Considérons une arrivée sismique dont le mouvement peut être décrit par $d(t) = e^{(t-2.5)^2} \sin(\omega t)$. L'arrivée sismique est représentée à basse fréquence (longue période) $T = 2\ s$ sur la Figure Figure22a. Imaginons que l'arrivée modélisée reproduit la forme de l'arrivée enregistrée mais arrive τ secondes en avance, $u(t) = d(t - \tau)$. L'arrivée modélisée est représentée pour $\tau = 0.5s$ par une ligne rouge tireté. L'amplitude maximale des arrivées modélisées et enregistrées sont situées respectivement aux temps T_c et T_o et sont indiquées par les étoiles bleu clair. La Figure 22d montre la fonction coût en fonction de l'erreur sur le temps de trajet de l'arrivée modélisée. Pour la gamme de valeurs représentées, la fonction coût des écarts des temps de trajet (bleu clair) augmente lorsque l'écart de temps augmente. Comme cela est illustré graphiquement, l'algorithme d'optimisation sera en mesure de localiser le minimum global. L'écart des formes d'onde (bleu foncé) montre que l'optimisation locale convergera vers la bonne solution seulement si $\tau < 1s$. En particulier, pour $\tau = 0.5$ qui correspond à l'arrivée représentée en rouge 22a, l'optimisation convergera parce que l'étoiles rouge sur la Figure 22d est localisée dans le bassin d'attraction. L'inversion des formes d'ondes devient plus problématique quand la fréquence augmente (la période diminue). Pour $T = 1s$ (Figure 22e), l'écart de temps initial τ doit être inférieur à 0.5 s pour permettre à l'inversion de converger vers le minimum global et, pour $T = 0.5\ s$ (Figure 22f), cet écart initial doit être inférieur à 0.25 s. Comme cela est illustré sur les Figures, la condition pour éviter de converger vers un minimum local est $\tau < T/2$. Cela peut s'exprimer de manière équivalente en postulant que le modèle du sous sol doit générer des formes d'onde qui ne s'écartent pas de la forme d'onde observée par plus d'une demi période. Dans le cas contraire, des phénomènes de saut de phase (cycle-skipping) vont se produire. Par exemple, pour une fréquence $f = 2Hz$ et $\tau = 0.5s$, la Figure 22f montre qu'utiliser une norme L2 de l'écart des formes d'ondes, l'inversion restera bloquée dans un minimum secondaire quand un saut de phase est généré. La non linéarité du problème inverse requiert que le modèle initial soit suffisamment précis de manière à être situé dans le bassin d'attraction de la fonction coût.

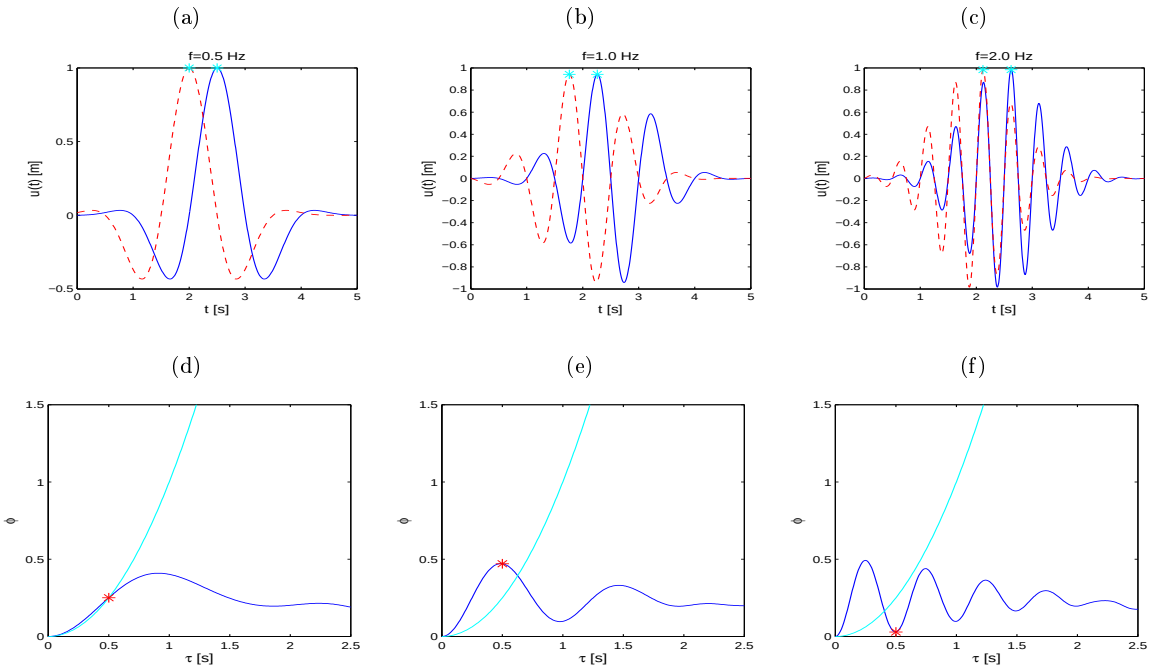


Figure 22: La première ligne représente les formes d'onde enregistrées $d(t) = e^{(t-2.5)^2} \sin(2\pi f t)$ en bleu et les formes d'ondes calculées en rouge, $u(t) = d(t-\tau)$ for $\tau = 0.5s$. Les formes d'ondes sont représentées pour a) f= 0.5 Hz b) f=1.0 Hz, d) f=2 Hz. La deuxième ligne représente la fonction coût des temps de trajet (25) en bleu clair en fonction de l'écart τ et la fonction coût de l'écart des formes d'ondes (17) en bleu foncé. Les fonctions coût sont représentées pour d) f= 0.5 Hz e) f=1.0 Hz, f) f=2 Hz. Pour les valeurs dessinées ici, la fonction coût des écarts de temps de trajet est convexe (bleu clair). En revanche, la fonction coût des écarts de formes d'ondes (bleu foncé) est non convexe. Pour un écart de temps de $\tau = 0.5s$, l'optimisation convergera vers le minimum global pour $f = 0.5\ Hz$ mais convergera vers un minimum local pour $f = 2hz$ en raison de sauts de phase.

D'autres expressions de fonctions coûts ont été proposées pour surmonter cette difficulté à la fois en domaines temporel et fréquentiel (Shin et al., 2002; Sheng et al., 2006; Shin and Min, 2006; Pyun et al., 2007; Shin et al., 2007; Shin and Ha, 2008; van Leeuwen et al., 2010; Hale, 2013). L'objectif est de créer des fonctions coût qui soient moins sensibles aux problèmes de sauts de phase, c'est-à-dire qui présentent un comportement moins oscillant (comme c'est le cas sur la Figure22a), quitte à perdre une partie de l'information portée par la forme d'onde (par exemple, lorsque l'on utilise l'enveloppe du signal, lorsque l'on fenêtre ou amortit le signal en fonction du temps ou lorsque l'on utilise des fonctionnelles fondées sur des corrélations). Ces fonctions coût peuvent se révéler particulièrement utiles en l'absence de basses fréquences dans les données. Dans ce cas, elles peuvent fournir un modèle du sous-sol intermédiaire pouvant être utilisé comme modèle initial de la FWI conventionnelle fondé sur une norme L2 des écarts de formes d'ondes.

Inversion hiérarchique multi-échelles

Dans le même ordre d'idées, l'inversion dans le domaine temporel ou fréquentiel peut être réalisée de manière hiérarchique en utilisant progressivement des données de plus en plus haute

fréquence (Bunks et al., 1995; Pratt and Shipp, 1999; Sirgue and Pratt, 2004). La procédure consiste à appliquer un filtre passe bas aux données avant d'appliquer la FWI. Notons m_1 le modèle final d'une inversion effectuée sur une bande passante donnée. Ce modèle est utilisé comme modèle initial pour l'inversion d'un nouveau jeu de données dont la fréquence de coupure a été augmentée. Cette procédure est poursuivie jusqu'à ce que toute la bande passante a été traitée. Dans l'exemple de la Figure 22, cette approche hiérarchique consisterait à minimiser d'abord la fonctionnelle de la Figure 22a et d'utiliser le modèle final de cette inversion comme le modèle initial pour la minimisation de la fonctionnelle représentée sur la Figure 22b.

L'écart des formes d'ondes à basses fréquences génère des fonctions coûts qui sont plus convexes et donc plus faciles à minimiser et nécessitant un modèle initial moins précis. A contrario, L'écart des formes d'ondes à hautes fréquences est fortement non convexe et requiert un modèle initial précis pour converger vers le minimum global. L'approche hiérarchique de la FWI fournit dès lors un modèle initial plus précis (au sens de mieux résolu) au fur et à mesure que les fréquences augmentent. Cette approche hiérarchique est illustrée sur la Figure 23 à l'aide d'un cas synthétique idéal correspondant à une coupe verticale du modèle overthrust de l'EAGE/SEG. La plus petite fréquence inversée est égale à 3Hz. L'inversion de cette composante fréquentielle fournit le modèle initial de la fréquence suivante de 5 Hz et ainsi de suite jusqu'à une fréquence maximale de 20Hz pour un total de 9 fréquences inversées. Notons que les basses fréquences contribuent à imager les grandes structures du modèle (les grandes longueurs d'onde), plus de détails (courtes longueurs d'onde) étant ajoutées à l'image au fur et à mesure que l'inversion progresse vers les hautes fréquences. Cette approche hiérarchique définit donc une imagerie multi-résolution ou multi-échelles.

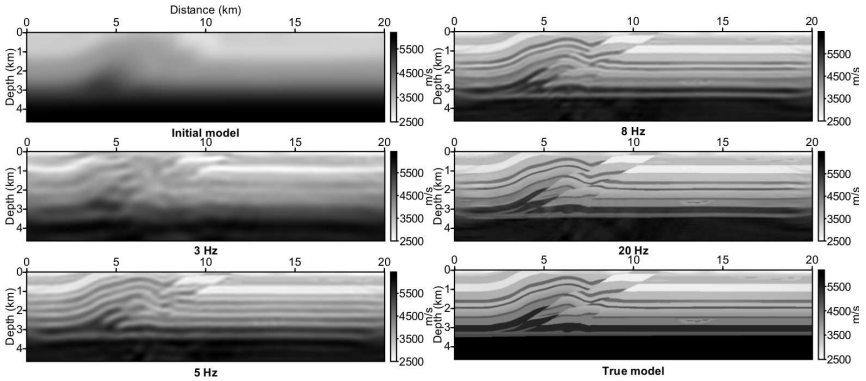


Figure 23: FWI hiérarchique procédant des basses fréquences vers les hautes fréquences. Les modèles initial et final sont respectivement en haut à gauche et en bas à droite. Neuf fréquences entre 3Hz et 20Hz sont inversées. Le spectre des nombres d'onde du modèle est enrichi de nombres d'onde de plus en plus élevés au fur et à mesure que l'inversion progresse vers les hautes fréquences.

Le rôle du Hessien

D'autres approches pour améliorer les performances de l'optimisation numériques ont été proposées en incorporant la dérivé seconde de la fonction coût (la Hessienne) dans le processus de minimisation (Pratt and Worthington, 1990). Lorsque des méthodes de Newton sont utilisées, la direction de descente est l'inverse du Hessien fois le gradient. Lors d'une des premières applications de la FWI (non linéaire), les résultats furent comparés à ceux de la tomographie en diffraction (linéaire) et il y avait une intuition claire que l'action du Hessien contribuait à accélérer la convergence via l'action focalisante (ou déconvoluante) du Hessien (Pratt and Wor-

thington, 1990). Néanmoins, en raison du coût calcul, ce n'est que lors de ces dernières années qu'une démonstration claire des améliorations fournies par les méthodes de Newton fut présentée. Lorsque l'on considère une approximation du second ordre de la fonction coût (fonction coût localement quadratique), le Hessien est constitué de deux termes, un formé par la corrélation entre les champs d'onde diffractés $\partial u/\partial m$ (les dérivés de Fréchet) et l'autre formé par des termes de diffraction double $\partial^2 u/\partial m^2$. Il a été montré que les artefacts d'étalement qui apparaissent en raison de ambiguïté de la condition d'imagerie (Figure 2.16) peuvent être diminués par l'action déconvoluante du Hessien dans l'approximation de Gauss-Newton (qui utilise uniquement les termes de diffraction d'ordre 1) (Pratt et al., 1998). Le Hessien complet fournit une amélioration supplémentaire en corrigeant le gradient des contributions parasites des diffractions d'ordre 2. Plus précisément, la solution complète de l'équation d'onde modélise des ondes diffractées multiples ce qui génère une incompatibilité avec la condition d'imagerie (20) qui considère uniquement les termes en diffraction simple. Cela implique que la partie des résidus dus à de la diffraction double seront interprétés à tort comme des événements simplement diffractés conduisant à une contribution erronée dans l'image. Le terme du second ordre dans le Hessien contribue à supprimer ces artefacts (Pratt et al., 1998). En théorie, quand le Hessien n'est pas utilisé, ces artefacts tendent à disparaître au cours des itérations. Néanmoins, en pratique, en raison de la non linéarité de l'inversion, l'inversion implémentée avec des méthodes de gradient peut rester bloquée dans un minimum local en raison de ces artefacts. Par exemple, reconstruite le modèle de la Figure 24a avec une fréquence de 7 Hz est difficile car la fréquence est élevée relativement à la taille caractéristique des structures présentes dans le modèle initial. Cela peut générer des sauts de phase et la convergence vers un minimum local. De plus, il y a des événements doublement diffractés entre les deux inclusions circulaires. L'inversion fondée sur l'algorithme de Newton fournit clairement la meilleure solution (Figure 24e). L'autre avantage des méthodes de Newton est de réduire le nombre d'itérations nécessaires pour atteindre une valeur donnée de la fonction coût, avec l'inconvénient que le coût d'une itération est plus élevé comparativement à celle d'une méthode de plus grande pente. Des applications sur données réelles avec des algorithmes de Newton et de quasi-Newton (l-BFGS) sont présentées dans Brossier et al. (2009a,b); Plessix et al. (2012); Métivier et al. (2014).

Inversion multi-paramètres

Les bénéfices que l'on peut tirer de la prise en compte du Hessien s'appliquent aussi aux inversions multi-paramètres. Les applications pionnières de la FWI multi-paramètres ont montré que la résolution avec laquelle les paramètres de nature différente sont reconstruits dépend du choix de la paramétrisation du sous-sol (Tarantola et al., 1984; Mora, 1987; Pratt and Worthington, 1990). Il y a des phénomènes de couplages entre les paramètres de nature différente (la vitesse de propagation des ondes de compression v_p et de cisaillement v_s, la densité, l'atténuation, les paramètres décrivant l'anisotropie ϵ et δ, ce qui rend complexe le processus d'optimisation car le gradient de la fonctionnelle par rapport à un paramètre est pollué par l'influence des autres paramètres. Plus précisément, le modèle de perturbation pour un paramètre est en fait une combinaison linéaire des gradients par rapport aux différents paramètres. De plus, la sensibilité des données à chaque paramètre, que l'on peut mesurer par les dérivés de Fréchet $\partial u/\partial m$, peut avoir des ordres de grandeur fortement variable d'un paramètre à l'autre générant un problème d'optimisation mal conditionné. La conséquence est que l'optimisation tendra à ne mettre à jour que les paramètres dont les dérivés de Fréchet sont les plus élevées. Une re-paramétrisation des paramètres physiques dans l'inversion contribuera d'une part à adimensionaliser les dérivés de Fréchet et d'autre part à minimiser les couplages entre paramètres. Une re-paramétrisation du modèle du sous-sol modifiera les dérivés de Fréchet, dont les amplitudes sont contrôlées par le diagramme de rayonnement $\partial A/\partial m$ de la source virtuelle secondaire positionnée à la position du paramètre du modèle (Pratt et al., 1998). Seul ce diagramme de rayonnement différencie

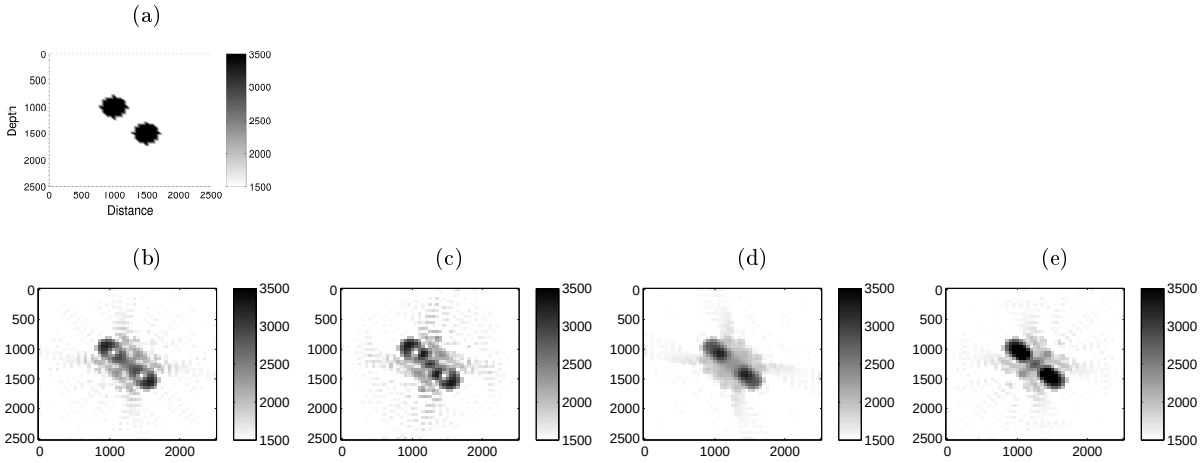


Figure 24: Inversion régularisée à une fréquence de $f = 7Hz$ pour différents algorightmes d'optimisation. En raison de la présence de minimum locaux et de l'imprécision du modèle initial, chaque algorithme fournit un modèle final différent. a) Modèle exact. b) Modèle final reconstruit avec un algorithme de plus grande pente. c) Idem que b) avec un algorithme de quasi-Newton (l-BFGS). d) Idem que b) avec un algorithme de Gauss-Newton. e) Idem que b) avec un algorithme de Full Newton. Pour plus de détails voir le chapitre 1.3.a.

l'expression des gradient de la fonctionnelle par rapport à des paramètres de nature différente. Il est donc important d'avoir une bonne compréhension du rôle joué par ces effets de directivité de source. Pour limiter les effets d'interférence entre paramètres, une paramétrisation du sous-sol peut être sélectionnée telle que les diagrammes de rayonnement des différents paramètres ont une intersection aussi faible que possible en fonction de l'angle de diffraction. Le Hessien contribuera également à corriger de ces effets de couplage et à mettre à l'échelle les gradients de manière à restaurer les bonnes unités physiques dans les modèles de perturbation. Par exemple, si la densité ρ et la vitesse des ondes P v_p sont les paramètres à reconstruire, l'approximation Gauss-Newton du Hessien contiendra 4 blocs formés par les corrélations $(\partial u/\partial v_p)^\dagger (\partial u/\partial v_p)$, $((\partial u/\partial v_p)^\dagger (\partial u/\partial \rho)$,

Chaque bloc contiendra des coefficients d'amplitude différente, dont la fonction sera de corriger les gradients associés à v_p et ρ des effets de couplage entre ces deux paramètres se manifestant pour des angles de diffraction faibles. Les coefficients du Hessien dépendent du choix de la paramétrisation. Différents choix de paramétrisation conduisent à une pondération différente des dérivés partielles. Une paramétrisation adaptée doit fournir un Hessien bien conditionné, au sein duquel chaque bloc a une contribution comparable. Ainsi, de la même manière que le Hessien corrige des effets d'étalement causés par l'ambiguïté dans la condition d'imagerie en focalisante l'énergie à la position des réflecteurs (Pratt et al., 1998), le Hessien contribue à diminuer les effets d'interférence entre paramètres en redistribuant de manière pertinente les amplitudes des gradients dans les modèles de perturbation. Des analyses asymptotiques du Hessien dans le cadre théorique de l'approximation haute fréquence de la théorie des rais et de l'approximation de Born permettent d'effectuer analytiquement sa décomposition en vecteurs propres et d'étudier l'influence de différents choix de paramétrisation sur le conditionnement du Hessien et le poids relatif de chaque vecteur propre (Forgues and Lambaré, 1997; Plessix and Cao, 2011; Operto et al., 2013). Depuis, quelques applications de FWI multi-paramètres à des données réelles ont été présentées dans la littérature (Gholami et al., 2013b,a; Prieux et al., 2013a,b).

Régularisation du problème inverse

La non convexité de la fonctionnelle à minimiser (non linéarité) et le caractère sous-déterminé du problème inverse (non unicité de la solution) nécessitent d'introduire des termes de régularisation pour restreindre l'espace des solutions et guider l'inversion vers un modèle du sous-sol que le géophysicien considérera plausible. Par exemple, des modèles du sous-sol seront rejetés et des modèles lisses seront favorisés dans la recherche. La régularisation peut être introduite dans la fonction coût,

$$\begin{aligned} \phi &= \phi_0 + \lambda \|R(m)\|_2^2 && (26) \\ &= \|Pu_c(m) - d_o\|_2^2 + \lambda \|R(m)\|_2^2, && (27) \end{aligned}$$

où $R(m)$ est le terme de régularisation et $\lambda\|R(m)\|_2^2$ est un terme de paramètre avec λ jouant le rôle de multiplicateur de Lagrange. L'optimisation vise à trouver le modèle qui explique au mieux de manière conjointe les données et le terme de régularisation. Généralement, des modèles lisses sont favorisés ce qui conduit à $R(m) = \nabla m$. Si des informations a priori sur les valeurs du modèle à certaines positions de l'espace sont disponibles, celles-ci peuvent être aussi incorporées dans le terme régularisant (Asnaashari et al., 2013). Relativement peu de travaux ont été publiés sur le choix optimal de régularisation et de normes en FWI. A titre d'exemple, citons néanmoins l'utilisation de régularisations fondées sur la variation totale (TV) des modèles du sous-sol dont l'objet est de favoriser la reconstruction des contrastes dans les modèles du sous-sol (Ramírez and Lewis, 2010; Anagaw and Sacchi, 2012; Guitton, 2012),

$$\phi = \|Pu_c(m) - d_o\|_2^2 + \lambda \|\nabla(m)\|_1, \quad (28)$$

ou des régularisation visant à favoriser le caractère creux de la solution sur des bases adaptées W (généralement d'ondelettes) (Candes and Romberg, 2005; Loris et al., 2010; Herrmann and Li, 2012),

$$\phi = \|Pu(m) - d\|_2^2 + \lambda \|W(m)\|_1. \quad (29)$$

D'autres régularisations ont été proposées, où la régularisation n'est pas additionnée à la fonctionnelle mais est plutôt multipliée (Abubakar et al., 2002, 2004). Théoriquement, les régularisations multiplicatives ont l'avantage que le poids relatif entre le gradient généré par les données et celui généré par la régularisation est déterminé de manière automatique.

2.2 Présentation de la thèse et description des principaux résultats

Cette brève discussion sur les principales questions et les défis soulevés par la FWI montre pourquoi cette technologie est toujours un champ actif de recherche depuis les années 80. Cette thèse a pour objectif d'ajouter une contribution à la compréhension de la FWI en explorant certains aspects spécifiques en rapport avec son coût numérique et sa non linéarité.

Je commence par présenter de manière générale la théorie des problèmes inverses et des algorithmes d'optimisation dans le chapitre 1, sans me restreindre au cas spécifique de l'imagerie sismique. J'introduis les concepts de problèmes inverses linéaires versus non linéaires et des problèmes inverses mal posés. Comme les problèmes inverses sont généralement formulés sous forme d'un problème de minimisation, j'introduis les algorithmes d'optimisation que je serai amenée à manipuler tout au long de cette thèse. Le lecteur familier de ces concepts pourra directement se reporter au chapitre sur la FWI.

Le chapitre 2 complète l'introduction fournie ci-dessus sur la FWI. Une interprétation physique des deux principaux ingrédients intervenant dans les algorithmes d'optimisation est fournie: le gradient de la fonctionnelle et le Hessien.

Pour un problème de petite dimension, je calcule le Hessien complet et illustre l'action d'une ligne de son inverse sur le gradient de la fonctionnelle afin de fournir des éléments d'information sur la mécanique avec laquelle il corrige les artéfacts. L'interprétation physique des conditions d'imagerie associées à d'autres méthodes d'imagerie fondées sur le concept de renversement temporel est passée en revue dans l'annexe 1. Les expressions du gradient et du Hessien qui sont développées dans la littérature avec la méthode de l'état adjoint (Lions, 1972; Chavent, 2009; Plessix, 2006; Métivier et al., 2013a) sont utilisées dans cette thèse et sont redéveloppées dans les annexes 3 et 5 pour le cas spécifique des équations de l'elastidynamique formulées sous forme d'un système hyperbolique du premier ordre.

Formulation auto-adjointe de l'équation d'onde élastique isotrope en vitesse-contrainte

Généralement, le gradient de la fonction coût de la FWI est implémenté avec la méthode de l'état adjoint à partir de l'équation d'onde du second ordre auto-adjointe. En revanche, l'équation d'onde formulée sous forme d'un système hyperbolique d'ordre 1 en vitesse-contrainte ne conduit pas à une forme auto-adjointe ($A \neq A^{\dagger}$), nécessitant une implémentation dissociée de l'opérateur direct (équation d'état) et de l'opérateur adjoint. Ma contribution, présentée dans l'annexe 4, a consisté à développer un formalisme permettant d'exprimer les équations de l'élastodynamiques du premier ordre sous forme d'un opérateur auto-adjoint facilitant ainsi grandement l'implémentation du gradient. Ceci est implémenté via un changement de variable appliqué aux contraintes normales, fourni par la décomposition en vecteurs propre de la matrice de raideur. Ce changement de variable permet de reformuler l'équation d'onde sous une forme pseudo-conservative au sein de laquelle les paramètres du milieu sont regroupés dans une matrice diagonale factorisée aux dérivés temporelles du système hyperbolique. Ce travail a été publié dans Castellanos et al. (2011).

Encodage de sources combiné avec des méthodes d'optimisation du second ordre

En imagerie sismique, le nombre de simulations à effectuer est proportionnel au nombre de source. Dans les approches fréquentielles fondées sur des solveurs directs, le problème direct est résolu de manière plus efficace car la factorisation LU de la matrice est commune à toutes les sources et les solutions sont calculées efficacement par substitution. Néanmoins, pour des applications en trois dimensions de la FWI élastique, les méthodes de modélisations fondées sur des algorithmes explicites d'évolution temporelle ou sur des solveurs itératifs en domaine fréquentiel semblent les plus réalistes. Ces méthodes doivent tirer un bénéfice important des techniques d'encodages de sources car le coût des modélisations est directement proportionnel au nombre de sources contrairement aux approches fondées sur des solveurs directs. Bien que les méthodes d'encodage de sources aient été beaucoup utilisées en FWI ou en migration par renversement temporel (Neelamani et al., 2008; Krebs et al., 2009; Ben Hadj Ali et al., 2011; Schuster et al., 2011; van Leeuwen et al., 2011; Huang and Schuster, 2012), elles ont principalement été combinées avec des algorithmes d'optimisation de gradient, reproduisant ainsi les principes des méthodes de gradient stochastiques (**??**). Dans le but de réduire encore le coût des calculs et améliorer le facteur d'accélération fourni par les méthodes d'encodage (i.e., augmenter l'écart entre les courbes obtenues avec et sans encodage sur la Figure 21), j'ai combiné des algorithmes d'optimisation de quasi-Newton et de Newton avec de l'encodage de source, ce qui n'avait jamais été proposé dans le cadre applicatif de la FWI.

Dans d'autres domaines comme en Machine learning, cette combinaison commence à peine à être explorée (Schraudolph et al., 2007). L'absence de preuve de convergence des méthodes stochastiques du second ordre (Bottou and Le Cun, 2005) et le manque d'algorithmes efficaces pour calculer la direction de descente de Newton (Métivier et al., 2013b, 2014) figurent parmi les raisons ayant freiné ces investigations. Les équations permettant le calcul efficace du Hessien

complet à partir de la méthode de l'état adjoint du second ordre sont présentées dans l'annexe 5. Néanmoins, le coût de calcul par itération de la FWI des méthodes de Newton tronqué (Gauss-Newton et Full Newton) est plus élevé que celui des méthodes de plus grande pente, car le calcul de la direction de descente de Newton fondé sur des approches non matricielles nécessite de calculer plus de problèmes directs. Par ailleurs, dans le cas des méthodes de quasi-Newton fondées sur $l-$BFGS, l'approximation du Hessien a besoin d'être périodiquement ré-initialisé contribuant à ralentir la convergence. Dès lors, il n'est pas clair quel gain réel peut être tiré les méthodes d'optimisation du second ordre en FWI que des techniques d'encodage de sources soient utilisées ou pas.

Dans le chapitre 3, j'introduis un critère d'arrêt des itérations fondé sur le taux de réduction de la fonction coût et je compare la convergence (mesuré par le nombre d'itérations effectuées) et le coût de calcul (mesuré par le nombre de problèmes directs effectué) pour quatre méthodes d'optimisation (gradient conjugué non linéaire, $l-$BFGS, Gauss Newton et Full Newton), lorsque ceux-ci sont implémentés dans la FWI en domaine fréquentiel couplée avec de l'encodage de sources. Je compare la convergence, les coûts et la qualité des modèles du sous-sol obtenus avec source encoding avec ceux obtenus lorsque les tirs sont traités indépendamment. L'analyse est tout d'abord appliquée à un cas d'étude synthétique inspiré de la géologie du Golf du Mexique pour des données sans et avec bruit. Ensuite, je quantifie l'apport de la méthode d'encodage de sources et des méthodes d'optimisation du second-ordre avec des données réelles enregistrées par un dispositif de câbles de fond de mer sur le champ pétrolier de Valhall (mer du Nord). Bien que la méthode d'encodage de sources soit plus adaptée à des méthodes de modélisation fondées sur des solveurs itératifs ou des méthodes explicites d'évolution, j'utilise un code fondé sur un solveur direct pour des raisons d'efficacité calculatoire (les cas d'étude présentés sont 2D et acoustiques). Néanmoins, les conclusions que je tire reposent sur le nombre de problèmes directs effectués (le nombre de phase de substitutions) et s'appliquent donc à l'identique à d'autres méthodes de modélisation. Une partie des résultats présentés ont été soumis pour publication dans Castellanos et al. (2013).

Conclusions sur l'encodage des sources couplée avec des méthodes d'optimisation du second-ordre: cas synthétique

Mes résultats de FWI sur le cas synthétique, lorsque le modèle initial est de précision suffisante et les données ne contiennent pas de bruit, indiquent que le gain calcul fourni par l'encodage des sources couplée avec des méthodes d'optimisation du second-ordre est significatif. La plus grande accélération est obtenue avec l'algorithme de $l-$BFGS et de Gauss-Newton et le meilleur taux de convergence est atteint par Gauss-Newton. Les gains en temps calcul pour ce cas d'étude sont illustrés sur la Figure 25. Quand du bruit est introduit dans les données (la puissance du bruit représente 25% de celle des données), le taux de convergence de toutes les méthodes d'optimisation sans encodage de sources tendent à se niveler. Cela suggère que l'action du Hessien sur le gradient n'améliore plus de manière significative la direction de descente, en raison du bruit dans les données. Avec l'encodage des sources, cette tendance à niveler les performances de chaque méthode d'optimisation est reproduite. Néanmoins, les méthodes de Newton fournissent une variance plus faible des modèles de vitesse obtenus, lorsque plusieurs réalisations de la FWI sont effectuées. L'accélération fournie par l'encodage des sources est plus faible lorsque du bruit est introduit dans les données.

Conclusions sur l'encodage des sources couplée avec des méthodes d'optimisation du second-ordre: cas réel

Dans le cas des données réelles du champ de Valhall, l'inversion avec différentes méthodes

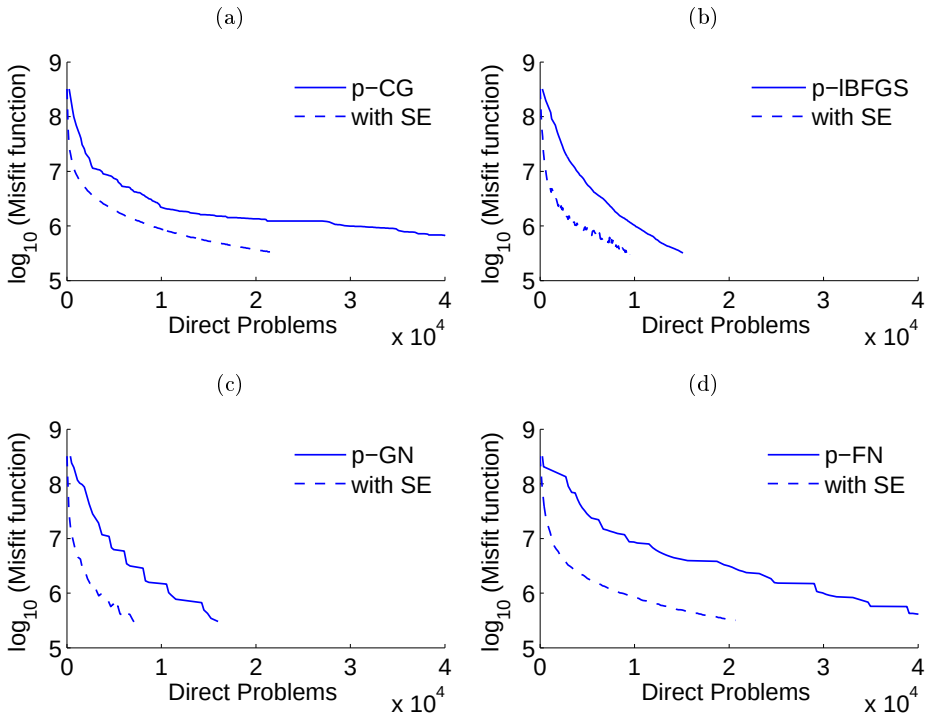


Figure 25: Cas d'étude du modèle BP sans bruit. Estimation de l'accélération fournie par la méthode d'encodage de sources - Réduction de la fonction coût en fonction du nombre de problèmes directs. Algorithmes d'optimisation. a) Gradient conjugué - (plus grande pente si l'encodage des sources est utilisé) b) l-BFGS c) Gauss Newton tronqué d) Newton tronqué. Les lignes continues et tirettes représentent respectivement le coût calcul sans et avec encodage de sources. Plus de détails sont fournis dans le chapitre 3, section 5.1.b.

d'optimisation ne convergent pas vers le même minimum local. Ceci résulte probablement du fait que, au voisinage du modèle initial, la fonction coût contient plusieurs minimum locaux. L'estimation de l'accélération fournie par l'encodage de sources et les méthodes d'optimisation du second d'ordre est donc biaisée car la qualité des modèles finaux n'est pas identique. Je montre que, avec et sans encodage de sources, les méthodes de Newton suivent la direction de descente qui converge vers les modèles du sous sol pour lesquels la fonction coût a la valeur la plus faible. En accord avec les résultats des cas synthétiques, la variance la plus faible est atteinte pour les méthodes de Newton. Comme le nombre de sources est très supérieur dans le cas réel que dans le cas synthétique, l'accélération fournie par l'encodage de sources est plus important jusqu'au point d'atteindre $\approx 90\%$. Dans tous les tests effectués, la qualité des modèles du sous-sol reconstruits avec l'encodage des sources est très proche de celle obtenue quand chaque source est traitée indépendamment. J'ai validé la qualité des modèles de vitesse obtenus par FWI en comparant les images migrées calculées dans les différents modèles de vitesse.

Estimation de la variance du gradient calculé avec encodage de sources

Pour une meilleure compréhension des artefacts d'interférence générés par l'encodage des sources, j'ai estimé dans la section 6 du chapitre 3 la variance du gradient avec encodage de sources pour des données sans bruit à partir de trois types d'assemblages de sources. L'analyse de ces variances peut être utilisée pour concevoir des stratégies optimisées d'encodage de sources. J'ai conclu que les gradients générés par plusieurs sources étaient comparables, ce qui peut se com-

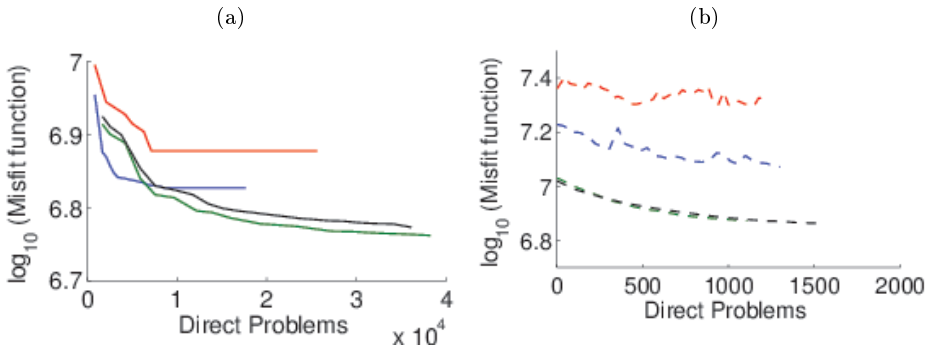


Figure 26: Cas d'étude de Valhall. Estimation de l'accélération pour le dernier groupe de fréquences inversées. Réduction de la fonction coût en fonction du nombre de problèmes directs a) sans encodage de sources (lignes continues) b) avec encodage de sources (lignes discontinues). Bleu : Gradient conjugué, Rouge: l-BFGS, Vert: Gauss Newton, Noir: Newton. La différence entre le nombre de problèmes directs est d'un ordre de grandeur. Pour plus de détails voir le chapitre 3, Section 5.2.

prendre dans le cas d'acquisitions extrêmement denses avec une forte redondance d'information. Dans ces configurations, assembler toutes les sources au sein d'une seule super-source conduit à une variance élevée. Dans ce cas où la redondance est forte, il est préférable de sous-échantillonner les sources. A contrario, si les gradients générés par chaque source sont très différents, le sous-échantillonnage des sources fournira une forte variance dans le gradient avec encodage de sources si bien qu'une stratégie fondée sur l'assemblage de toutes les sources au sein d'une seule super-source est préférable dans ça cas.

Remarques générales sur l'encodage de sources avec méthodes d'optimisation du second ordre

L'accélération fournie par l'encodage de source peut être augmentée en la combinant avec les méthodes d'optimisation de second ordre de quasi-Newton ou de Newton. Lorsque le niveau de bruit augmente, l'apport des méthodes du second ordre en termes d'accélération devient moins évident comparativement aux méthodes de plus grande pente. Néanmoins, les méthodes du second ordre fournissent la direction de descente la plus robuste et les modèles de vitesse avec la variance la plus faible.

Les algorithmes d'optimisation stochastiques ont un taux de convergence plus faible que les algorithmes d'optimisation déterministes. Par conséquent, une accélération significative est obtenue lors de la phase initiale de l'inversion et diminue au fur et à mesure que l'inversion progresse vers le minimum de la fonctionnelle. En d'autres termes, l'accélération obtenue dépend de la valeur de la fonction coût à laquelle l'inversion est stoppée. Pour garantie une accélération significative avec la méthode d'encodage de sources, un critère d'arrêt des itérations doit être défini en fonction d'une tolérance d'erreur sur la fonction coût.

En dépit des bruits d'interférence générés par l'encodage des sources, la qualité des modèles de vitesse est satisfaisante. de plus, j'ai illustré, à l'aide d'un exemple de une fonction coût fortement non convexe générée en utilisant un modèle initial moins précis et en incorporant des fréquences plus élevées, que l'encodage de sources pouvait aider à guider l'inversion vers un meilleur minimum local de la fonction coût grâce à une exploration plus exhaustive de l'espace des modèles.

En considérant des données non bruitées et des algorithmes de plus grande pente, j'ai développé l'expression de la variance du gradient avec encodage de sources. L'analyse de la variance suggère que, en l'absence d'information sur les gradients produits par chaque source, regrouper toutes les sources au sein d'une seule super-source est peut être la meilleure stratégie. Néanmoins, si des gradients individuels sont similaires (en raison d'une acquisition sur-échantillonnée), des stratégies plus adaptées d'assemblages de sources peuvent être définies.

Régularisation fondée sur la variation totale

Le chapitre 4 est dédié à la régularisation des problèmes inverses. Bien entendu, beaucoup de travaux ont été consacrés à l'exploration de nouvelles normes dans l'espace des données (Djikpéssé and Tarantola, 1999; Guitton and Symes, 2003; Ha et al., 2009; Pyun et al., 2009; Brossier et al., 2010), mais seulement récemment l'influence de la norme utilisée dans l'espace des modèles (la régularisation) a été analysée (Burstedde and Ghattas, 2009; Ramírez and Lewis, 2010; Anagaw and Sacchi, 2012; Guitton, 2012). Je compare dans cette thèse deux normes dans la régularisation, la norme l_2 (équation 27) et la variation totale (TV) (équation 28), à partir d'un test synthétique réaliste (le modèle BP-2004 salt) et les données d'OBC enregistrées sur le champ de Valhall. Sur le cas synthétique, j'ai utilisé un modèle initial suffisamment précis pour éviter les sauts de phase mais suffisamment éloigné du vrai modèle pour que la régularisation joue un rôle significatif. En raison des réflexions se produisant au voisinage de la surface, des artefacts sont générés dans la partie superficielle des modèles FWI. Je montre que, pour des données bruitées ou pas, ces artefacts sont fortement réduits quand j'utilise la régularisation par variation totale. En l'absence de bruit, le modèle final obtenu avec la norme TV est considérablement meilleur comme illustré sur la Figure 27. Pour le cas d'étude réel, les modèles de vitesse finaux obtenus avec les normes l_2 et la TV sont comparables. Néanmoins, la stratification induite par les couches géologiques est mieux restituée dans les modèles construits avec la régularisation par variation totale. En conclusion, je fournis dans cette thèse une des premières démonstrations qu'une régularisation fondée sur la variation totale des paramètres est adaptée à la reconstruction de modèles du sous-sol par FWI en restituant de manière plus contrastée les discontinuités lithologiques.

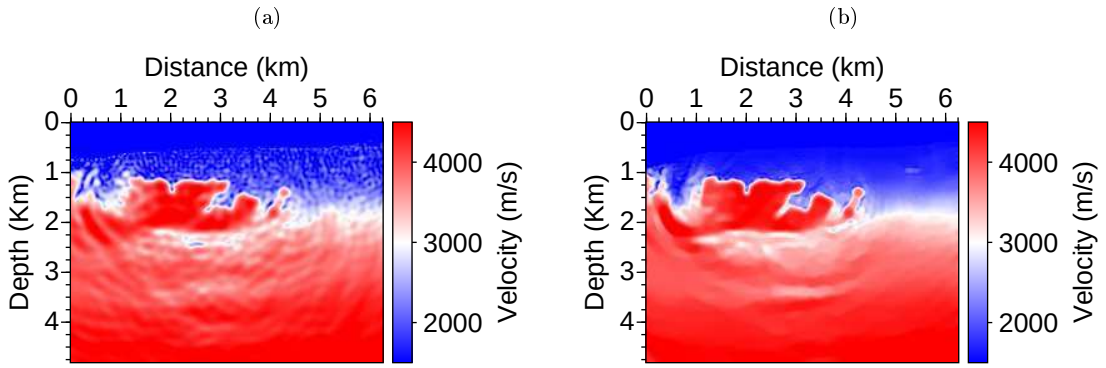


Figure 27: Données sans bruit. Comparaison des résultats de la FWI avec comme terme de régularisation a) $\|\nabla m\|_2^2$ b) variation totale $\|\nabla m\|_1$. Pour plus de détails voir le chapitre 4, Section 1.4.a.

Débruitage par variation totale locale à partir de l'information fournie par la migration

L'algorithme de débruitage de Rudin-Osher-Fatemi (ROF) (Rudin et al., 1992) élimine le bruit d'une image en minimisant la variation totale de l'image, tout en maintenant l'image débruitée aussi similaire que possible à l'image initiale. Cet algorithme de débruitage a connu un remarquable succès et est très populaire (Osher et al., 2005; Caselles et al., 2011), par exemple dans le domaine de l'imagerie médicale. Néanmoins, les limites du débruitage par TV résident dans le fait qu'il peut éliminer la texture (les structures de petites dimensions) dans l'image. Cela a motivé un certain nombre de travaux dédiés à la conception d'algorithmes de débruitage local par TV qui préserve la texture de l'image (Bertalmio et al., 2003; Vese and Osher, 2003).

Dans la Section 1.3 du chapitre 4, j'applique le débruitage par TV au modèle final de la FWI obtenu sur le champ pétrolier de Valhall. Comme cela était prévisible, lorsque un débruitage trop agressif est appliqué, l'algorithme élimines des petites structures de l'image. Une stratégie possible est d'interrompre prématurément l'algorithme. Alternativement, je propose un débruitage local par TV où j'incorpore une information a priori sur la réflectivité fournie par une image migrée. L'algorithme de débruitage local par TV n'a aucune action aux positions de l'espace où la migration a positionné des réflecteurs et débruite les autres parties de l'image. L'algorithme de débruitage local par TV remplit alors sa mission tout en préservant les principaux réflecteurs comme illustré sur la Figure 28.

Sur la différence de contenu spectral des arrivées réfléchies et transmises

Pour clore le chapitre 4, j'ai ré-exploré le problème du choix des fréquences en FWI en domaine fréquentiel lorsque le modèle du sous-sol présente de forts contrastes et génèrent des ondes réfléchies énergétiques dans les données. J'ai illustré à l'aide du modèle BP-2004 salt et pour une acquisition de surface comment le contenu spectral des données réfléchies et transmises pouvait différer en fonction de l'offset sous l'effet de la structure du sous-sol. En particulier, j'ai illustré la présence de gaps dans le spectre des arrivées réfléchies. L'influence de ces variations du spectre du signal sur l'inversion mérite des études plus approfondies. Néanmoins, je montre que ces gaps dans le spectre des données contribuent à élargir l'espace nul et que le remède réside dans l'incorporation d'une bande plus large de fréquences dans l'inversion.

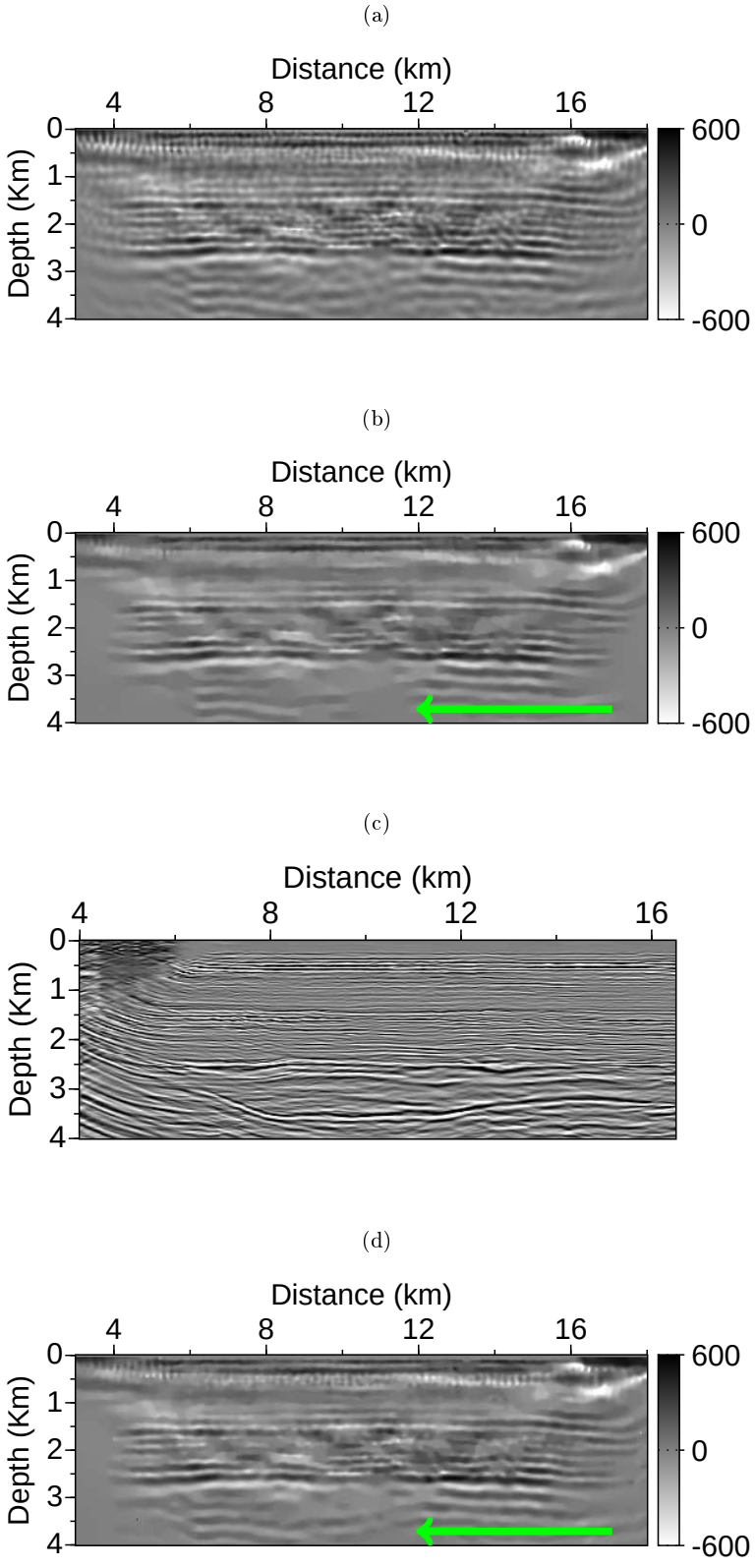


Figure 28: Illustration du débruitage local par variation totale. a) Modèle de perturbation à débruiter. b) Modèle de perturbation après débruitage par variation totale. c) Image migrée. d) Modèle de perturbation après débruitage local par variation totale en utilisant les mêmes paramètres qu'en b) et en utilisant l'information sur la réflectivité fournie par l'image migrée. Pour plus d'informations voir le chapitre 4, Section 1.4.b.

Chapter 1

Fundamentals of inverse problem theory

Inverse problems arise naturally in many physical applications where data of a system is collected in an indirect fashion and one wishes to estimate the values for the model parameters. The applications of inverse problems concerning geophysics fall mainly under the category of tomography [1]. Tomographic methods generally refer to imaging techniques where the information and data are obtained on the surface of the objects or, more generally, through non invasive techniques. In the case of medical applications, some known tomographic methods are CT scans (computarized tomography) and the magnetic resonance imaging (MRI), and in geophysics two well known techniques are travel time tomography and full waveform inversion.

There are two parts in solving an inverse problem. The first step consists in solving a *forward problem* using a physical model m to generate a set of observables d. The relationship between the model $m(x)$ and the data $d(x)$ is through an operator $A(m)$ that contains the physical laws known to govern the process. Depending on the application, $A(m)$ may be an Ordinary Differential Equation (ODE), a Partial Differential Equation (PDE) or, if the problem is linear, a system of algebraic equations. For instance, one may be interested in studying a wave propagation or a temperature diffusion problem, which in this case would lead to an operator $A(m)$ to be described by PDEs.

The *inverse problem* aims to find the model parameters $m(x)$ from the known measured data $d(x)$ (For notational convenience, we will not state explicitly the dependence of m and d with respect to x unless necessary). This type of problem may also be commonly referred to as a parameter estimation inverse problem. The inverse problems are classified into linear or non-linear inverse problems (Menke, 1984). *Linear inverse problems* make reference to problems where the relationship between the measurements and the model parameter is linear : $d \propto m$. *Non-linear inverse problems* arise when the measurements d have a non-linear relation with the unknown parameter m : $d = F(m, x)$, where F is a non linear function in m.

1. The word *tomography* comes from the Greek *tomos* which means section and the English suffix *graphy*

Whether we are dealing with linear or non-linear inverse problems, parameter estimation and inverse problems are generally difficult to solve because they are generally *ill-posed* and thus we are faced with issues where we cannot guarantee the existence of the solution, its uniqueness or a stability in the solution process. Problems with the existence of the solution may arise when the mathematical model that reflects the physics is incomplete and thus we cannot generate a model to explain the data. None-unique problems, also known as undetermined problems, appear commonly with discrete problems when we are faced with the inversion of rank deficient matrices, revealing a non trivial null space. Thus any linear combination of models in the model null space does not change the fit in the data, leading to a situation where a large number of models give the same solution of data fit. Finally, instabilities in the inversion arise often due to the fact that a small change in the measurement lead to big changes in the model. Regularization techniques are frequently employed to generate stable solutions.

The solution strategies to solve inverse problems are therefore going to be dependent and specific to the classification of the problem : linear, non-linear, ill-posed, discrete, or continuous (Tarantola and Valette, 1982; Snieder and Trampert, 2000). After giving a brief description of these classification categories, we will focus on strategies to solve non-linear ill posed discrete inverse problems.

0.3 Linear Inverse Problems

Continuous linear inverse problems can be written in the form

$$\int_{\Omega} a(x,x')m(x')\mathrm{d}x' = d(x), \qquad x \in \Omega, \tag{1.1}$$

where $a(x,x')$ is known as the kernel function. Let $\Omega \subset \mathbb{R}^D$, where $D = 1,2,3$ indicates the dimension of the subspace Ω. In continuous form $d(x) \in \mathcal{D}$ is a function that for any point in the space $x \in \Omega$, $d(x)$ assigns a real or complex value. That is, $d : \Omega \to \mathbb{C}$. In physical applications, $\mathcal{D}$ represents the data space and $d(x)$ the data that may correspond to physical quantities such as the amplitude or phase of a signal, the temperature, displacement, resistivity, amongst many others. Likewise $m(x) \in \mathcal{M}$ is a model function in the model space $\mathcal{M}$. Thus, the model function assigns a parameter value to every point in the domain, $m : \Omega \to \mathbb{R}^{N_p}$, where N_p is the number of physical parameters that are to be reconstructed. For example, if we wish to reconstruct the permitivity and conductivity functions for an electromagnetic problem, then $N_p = 2$.

Equations that have the form of (1.1) are called Fredholm integral equations of first kind, and naturally arise in physics. It is straightforward to see that this problem is linear in m,

$$\int_{\Omega} a(x,x')\left[\alpha_1 m_1(x') + m_2(x')\right]\mathrm{d}x' \;=\; \alpha_1 \int_{\Omega} a(x,x')m_1(x') + \int_{\Omega} a(x,x')m_2(x')\mathrm{d}x'. \tag{1.2}$$

In a compact general form, a linear inverse problem can be written as

$$A(x)m(x) = d(x). \tag{1.3}$$

The operator A maps functions in the model space to functions in the data space, $A : \mathcal{M} \to \mathcal{D}$.

To solve the inverse problem, it is necessary for d to be attainable, that is, for d to be in the range of the transformation, $d \in \mathcal{R}(A)$. If $\forall d \in \mathcal{D}$, d is attainable, then this is equivalent to saying a solution exists . If the null space of A is empty ($\mathcal{N}(A) = 0$), then the solution is unique. We would additionally like the solution to change continuously with the input, which is

equivalent to imposing the continuity, which in turn is equivalent to imposing the boundedness, of the operator A^{-1}. However, this may not always be the case and not all values of d are in the range of A. Gauss and Legendre proposed a more general definition of the solution of an inverse problem based on the least squares method, in which the idea is to minimize the discrepancy expressed as a quadratic functional.

Definition 1.1. (Engl et al., 2000)
Let $A : \mathcal{M} \to \mathcal{D}$ be a bounded operator,
– $m \in \mathcal{M}$ is called the least squares solution of $Am = d$ if

$$||Am - d|| = \inf\{||Az - d|| \mid z \in \mathcal{M}\} \tag{1.4}$$

This is important because, even if the null space is not zero, there may be still interest in finding a solution to (1.3) in some approximate sense, which gives rise to what is called a generalized solution, that can be found through a generalized inverse of A (A^+), or sometimes also referred to as the pseudo inverse of A. The pseudo inverse of a matrix is unique for all matrices, and in the most general scenario, can be found using the singular value decomposition. The pseudo inverse gives the solution x such that A^+x has the minimum difference in the least squares sense to the desired solution y. The Moore-Penrose inverse of an operator A is defined by restricting the domain and range of A such that the restricted operator is invertible (Engl et al., 2000).

A linear inverse problem can therefore be stated as an optimization problem,

$$\min_m \phi(m) = \langle Am - d_o, Am - d_o \rangle = \langle d - d_o, d - d_o \rangle \,, \tag{1.5}$$

where ϕ is the discrepancy or misfit function, $\langle , \rangle$ denotes a scalar product in a Hilbert space $\mathcal{H}$, and we have denoted with d_o the measured (or observed data) to avoid confusion with the data d from (1.3).

In geophysics, an example of a linear inverse problem is *linearized travel time tomography* (Nolet, 2008). The unknown physical parameters are the changes in slowness (ΔS) and the measured data are the travel time differences (Δt) between the observed and calculated travel time, computed from a smooth model. In discrete form:

$$\Delta t = \sum_i l_i \Delta S_i, \tag{1.6}$$

where l_i is the length of the ray segment crossing cell i. The l_i are known values, which have been initially computed through a ray tracing technique using a smooth background model m_0, and the unknowns are ΔS_i. Therefore, this is a linear system of the form $d = Am$, like the ones that have been discussed in this section. The forward problem consists in constructing the matrix A, using any modeling method, such as ray tracing techniques. The inverse problem is solved iteratively, and the background model is always kept constant. This means that the matrix A is not updated therefore the values of l_i that constitute the matrix A are kept constant throughout the inversion. In other words, since the matrix A is not updated it is *independent* of the unknowns ΔS, thus leading to a linear inverse problem. If, on the other hand, we wished to consider the model perturbations ΔS and update the model, and perform the ray tracing to compute A in each iteration, we would be faced with a non linear inverse problem.

0.4 Non linear Inverse Problems

Generally, it may not be possible to write the forward problem in the same form as (1.3), and a more general approach is to formulate the direct and inverse problem as (Tarantola and

Valette, 1982)

$$F(x,m) = u(x,m) \tag{1.7}$$
$$\min_m \quad \phi(m) \quad m \in \mathcal{M}. \tag{1.8}$$

where

$$\phi(m) = \langle F(x,m) - d(x), F(x,m) - d(x) \rangle , \tag{1.9}$$

$d(x) \in \mathcal{D}$, $u \in \mathcal{D}'$ and in general $\mathcal{D}' \neq \mathcal{D}$, and $F(x,m) : \mathcal{M} \to \mathcal{D}'$. An example of a non linear inverse problem arises when the data and the model are not related in a linear fashion. Consider for example

$$A(x,m)u(x) = s(x), \tag{1.10}$$

where $A(x,m)$ represents any non linear differentiable operator acting on a function $u(x)$, and $s(x)$ is a source function. There is no linear relationship between u and m.

Assuming for the moment the operator $A(x,m)^{-1}$ is well defined, it is possible to write the non linear inverse problem in standard form,

$$\min_m \left\langle A(x,m)^{-1}s - d, A(x,m)^{-1}s - d \right\rangle \qquad , m \in \mathcal{M}. \tag{1.11}$$

where we have used

$$F(x) = A(x,m)^{-1}s(x). \tag{1.12}$$

In geophysics, the inversion of the full waveform (FWI) is a non linear inverse problem, where $A(x,m)$ represents the wave equation operator $A : \mathcal{Y} \to \mathcal{W}$, where $\mathcal{Y}$ and $\mathcal{W}$ are all the C^2 functions defined on Ω ($\mathcal{Y} \subset C^2(\Omega, \mathcal{W} \subset C^2(\Omega))$. $s : \mathcal{W} \to \mathbb{R}$ is the external source function, $u : \mathcal{Y} \to \mathbb{R}$ is the unknown wave field and $m(x) \in \mathbb{R}^{D \times N_p}$ are the model parameters. The forward and inverse problem can be synthetically written as

$$A(x,m)u(x) = s(x) \qquad \text{Forward problem} \tag{1.13}$$
$$\min_m \langle Pu(x,m) - d(x), Pu(x,m) - d(x) \rangle . \qquad \text{Inverse problem} \tag{1.14}$$

An operator P has been included in misfit function to project the solutions of the forward problem $\mathcal{Y}$ to the space of the measured data $\mathcal{D}$, $P : \mathcal{Y} \to \mathcal{D}$. For more details on the theory of non linear inverse problems, you may go to Tarantola (2005); Tarantola and Valette (1982); Snieder and Trampert (2000).

Notice that the direct problem in equation (1.13) is a *linear* problem where we assume $A(x,m)$ and $s(x)$ to be know. The inverse problem in equation (1.14) is NOT, however, linear since there is a non linear relationship between $m(x)$ (the physical parameters of the system) and the $u(x)$ (the wavefield).

1 Ill Posed Inverse Problems

For a problem to be well posed, the french mathematician Hadamard, defined three properties that should hold :

For all admissible data, a solution exits (1.15)

For all admissible data, the solution is unique (1.16)

The solution depends continuously on the data (1.17)

If we consider the inverse linear problem $Am = d$ as defined above, condition (1.16) is satisfied if and only if A is injective (one to one). The inverse of A must exist, and the domain of the inverse must coincide with $\mathcal{D}$. Condition (1.17) requires that the inverse operator of A be continuous, or equivalently, bounded. It should be noted that a problem is ill or well posed depending on the definition of the space of solutions, and on the norm that is chosen.

The lack of uniqueness of the solution of an ill-posed problem is one of the biggest concerns. There are many inverse problems in which the purpose is to find a continuous function, for example a model parameter, that is a function of the space variables. This means that the model has an infinite amount of degrees of freedom, but real experiments only provide a finite amount of data. Counting the variables demonstrates that the data cannot have enough information to uniquely determine the model. Another reason for the non uniqueness is that real data are always noisy, and errors in the data are propagated to errors in the estimation of the model parameters Snieder (1998). Considering the linear problem $Am = d$, none-uniqueness arises when the matrix A is rank-deficient and thus has a non trivial null space. All models that lie in the null space are solution to $Am = 0$. Moreover, by superposition, any linear combination of null space models is also in the null space. Therefore, there are an infinite number of models that may be mathematically acceptable and lead to the same misfit in the data. Given a series of possible solutions, one has to decide which one to choose. If possible, one has to include additional information. It is important to keep in mind that in a practical situation, the available data may simply not be enough to uniquely determine the solution. However, the additional information that is introduced may significantly smooth the model, or reduce its contrast, or bias it towards an a priori model. It is not straightforward to measure the bias introduced to the model via the additional information, and a resolution analysis of the model may be needed to help clear out this subject.

Violation of the third condition may lead to numerical instabilities. The inversion can be unstable, specially in the non linear case, and small changes in the data that may be even introduced by noise, may result in big changes in the estimated model. Usually, the solution of the direct problem has smoothing properties that may involve the integration of a function, and the inverse problem requires computing the derivatives and gradients of quantities that may respond in a highly non linear fashion.

Here we will show two relevant examples of Aster et al. (2005) and Engl et al. (2000) that reveal the non uniqueness character and instability that arises in the solution of inverse problems.

Example 1.1. (Aster et al., 2005) Consider a Fredholm integral equations of first kind

$$\int_0^1 g(s,x)m(x)\mathrm{d}x = 0, \tag{1.18}$$

where

$$g(s,x) = s\sin(\pi x). \tag{1.19}$$

We therefore wish to solve the inverse problem

$$\int_0^1 s\sin(\pi x)m(x)\mathrm{d}x = 0. \tag{1.20}$$

Because of the orthogonality of the functions $sin(k\pi x)$,

$$\int_0^1 \sin(k\pi x)\sin(l\pi x)\mathrm{d}x = \delta_{k,l} \qquad k \neq l. \tag{1.21}$$

Therefore the solution to (1.1) is

$$m(x) = \sin(k\pi x), \qquad k = \pm 2, \pm 3, \ldots \tag{1.22}$$

Moreover, since the problem is linear, any linear combination of the possible solutions is also a solution, and thus there are an infinite number of solutions that all fit the data equally well. This is a problem of non-uniqueness.

Example 1.2. Differentiation (Engl et al., 2000)

Solving inverse problems may sometimes require differentiation of the data. Let $f \in C^1[0,1]$, $\delta \in (0,1)$ and $n \in \mathbb{N}(n \geq 2)$. Define

$$f_n^\delta(x) := f(x) + \delta \sin \frac{nx}{\delta} \ , \ x \in [0,1], \tag{1.23}$$

where $f(x)$ represents the exact data and $f_n^\delta(x)$ the perturbed data. Taking the derivative,

$$\left(f_n^\delta\right)'(x) := f'(x) + n \cos \frac{nx}{\delta} \ , \ x \in [0,1]. \tag{1.24}$$

Consequently, for a small error in the data δ the error in the derivative can be arbitrarily large, depending on n. That is,

$$||f - f_n^\delta||_\infty = \delta, \tag{1.25}$$

$$||f' - \left(f_n^\delta\right)'||_\infty = n. \tag{1.26}$$

Therefore, the derivative has the characteristics of an ill posed inverse problem, as it does not depend continuously on the data, with respect to the uniform norm. The corresponding direct problem would be an integration step.

Integration is a smoothing, stable process in which high frequency oscillatory errors in x, are damped out. This can be clearly seen by looking at what happens when we go from (1.24) to (1.23), where the error term $n\cos(nx/\delta)$ is damped to $\delta \sin(nx/\delta)$.

The conclusions do not apply only in this application, and in general when the direct problem smooths perturbations, oscillations will appear in the inverse problem due to small but high frequency data perturbations.

Further examples showing the specific role of non-linearity in inverse problems have been previously shown (Snieder, 1998; Engl et al., 2000). Snieder (1998) show an example of a solution of an inverse problem where role of non linearity is put forward, showing the additional difficulties that arise.

Example 1.3. Non-linearity in seismic tomography (Snieder, 1998)
A cross borehole tomographic experiment is carried out where sources are placed on one side of the heterogeneities and the receivers are placed on the opposite side. In the case where there is an homogeneous velocity model as in Figure 1.1a, the rays travel in straight lines from the source to the sensors and all the receivers from R_1 to R_5 measure the wavefields. When a low velocity heterogeneity is introduced as shown in Figure 1.1b, the rays will bend around the anomaly seeking the path with minimal travel time. In this situation, the anomaly cannot be determined as it is not sampled by the rays, meaning that for a certain range of velocity models, the misfit function will be flat. If we now consider the situation where the a high velocity heterogeneity is present as in Figure 1.1c, a shadow zone will be formed behind the anomaly and no ray will hit receiver R_3. Therefore, there are some rays that have no corresponding measurements, thus we will not be able to reconstruct some model parameters. This accounts for regions in the model space that we are unable to reconstruct.

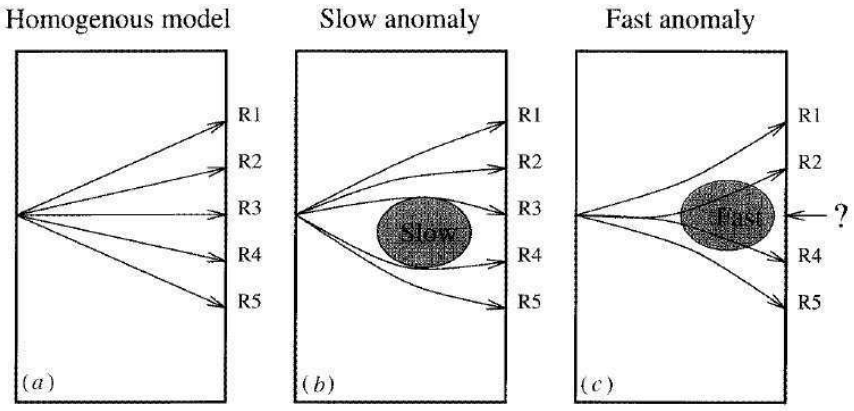


Figure 1.1: Figure from Snieder (1998). A cross bore-hole tomographic experiment where rays join a source to a string of receivers in another well. Three velocity models are considered. **a)** When a homogeneous velocity model is used, the rays travel in straight lines. **b)** When a velocity anomaly is sufficiently low, the rays are curved around the anomaly and it is not sampled by the rays. **c)** When a high velocity anomaly is introduced, the rays are curved and no data is measured by receiver R_3.

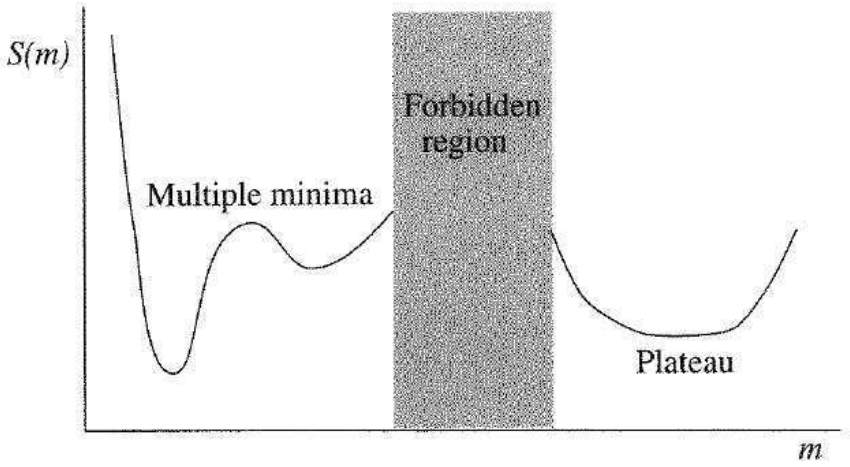


Figure 1.2: Figure from Snieder (1998). A misfit function that shows the multiple minima present due to the non linearity of the inverse problem, the plateau due to the model null space, and the forbidden model regions.

Although this example was illustrated using ray theory, similar situations will generally arise in the solution of inverse problems where some regions of the model space will be poorly sampled, or maybe not at all, because the wavefield amplitude is very small when it reaches these regions. This is what we refer to as the model null space, and creates a flat misfit function (Figure 1.2). Shadow zones may also appear for example in the presence of high velocity contrasting structures. Taking into account second order scattering effects in wave theory is a way to account for this setback.

Solving non linear inverse problems is therefore a challenging task not only due to the the several minima present in the misfit function, but also because when the measured data do not depend on the model parameters, the inverse problem is obviously ill posed and generates flat regions in the misfit function. Additionally, some model parameters may not be recoverable from the data, due to the non linearities, creating some forbidden regions. A possible misfit function may have the form show in Figure 1.2, and reveals the difficultly and complexity of solving non linear inverse problems. The solution of inverse problems is carried out with optimization algorithms, and their success has a clear impact on the quality of the solution and the time it is required to solve it. We will therefore briefly go over a few optimization concepts and algorithms, that will later be used in our numerical experiments.

2 Optimization

Inverse problems are generally formulated as optimization problems. For linear inverse problems, this might be caused by the incapability to invert the direct problem operator due to computational limitations or because the operator is low rank and its inverse does not exist. For non linear inverse problems we may not even have an expression of the operator that we need to invert. The misfit function in the optimization algorithm may be non convex due to the non linearities and presence of noise and possess multiple local minima, or flat regions. Therefore, the successful and efficient resolution of inverse problems depends on the optimization methods and their implementation. It is therefore important to have a clear picture of the optimization methods used, their hypothesis and limitations, as this will have a direct impact on the quality of the solution we expect to obtain.

Optimization methods are different for constrained and unconstrained systems of equations. The equivalence between the constrained and unconstrained formulation has been shown for several forms of objective and regularization functions. However, generally, optimization methods for unconstrained problems are considerably easier to implement and therefore we will mainly focus on these Nocedal and Wright (1999).

Iterative optimization methods generate a sequence $\{x_0, x_1, ..., x_k, x_{k+1}, ...\}$ of solutions where x_0 is the initial condition given by the user. Let us denote the increment in each iteration as

$$x_{k+1} = x_k + \alpha_k p_k, \tag{1.27}$$

where p_k is a descent direction, and α_k is a step length, both of which have to be determined in each iteration. How to find a descent direction p_k will be discussed further below. For now, we will assume the descent direction has been fixed, and focus on finding the best value for step length α which constitutes itself another optimization problem.

2.1 Line Search

Searching the best value for the step length α consists in solving the optimization problem

$$\min_{\alpha>0} \phi(x_k + \alpha p_k). \tag{1.28}$$

It is possible to solve (1.28) with an iterative optimization algorithm. This would imply starting with an initial value α_0 and generating a sequence $\{\alpha_i\}$, until a desired reduction of $\phi(x_k + \alpha p_k)$ is found. However, the descent direction is fixed and we are just searching how far we should move in that direction. Additionally, x_k is just an intermediate value in our sequence of iterates, and it results costly to search extensively for the best value of the step length. One possible approach consists in guessing a value for α, and verifying that the misfit function decreases with this value. Additional conditions can be imposed to make the line search more efficient.

The Wolfe conditions are popularly used in line search methods. A short motivation and explanation for these conditions is given below, inspired and based on (Nocedal and Wright, 1999). At first instance we could require only that

$$\phi(x + \alpha p) < \phi(x).$$

However, to assure that the decrease in the misfit function is not too small, we can ask for $\phi(x + \alpha p)$ to be less than $\phi(x)$ minus a additional quantity. That it, to ensure that the steps are not too small we would like to have a condition such as

$$\phi(x + \alpha p) < \phi(x) + K, \tag{1.29}$$

for some $K < 0$. To find an appropriate value for K, let us begin by doing a Taylor expansion to first order,

$$\phi(x + \alpha p) \quad \approx \quad \phi(x) + \alpha \nabla \phi^T p + O(\alpha^2). \tag{1.30}$$

Note that if p is an acceptable descent direction then $\nabla \phi^T p < 0$ (see Section 2.2). Therefore, we could impose that

$$\phi(x + \alpha p) \leq \phi(x) + c_1 \alpha \nabla \phi^T p, \tag{1.31}$$

where $c_1 \in (0, 1)$. The constant c_1 is included to relax the condition on the decrease of the function. The higher the value of c_1, the greater decrease in the misfit function we require, and probably the harder it will be to find an appropriate value for α.

The second condition, also to rule out very small steps, is the curvature condition,

$$\nabla \phi(x_k + \alpha_k p_k)^T p_k \geq c_2 \nabla \phi_k^T p_k. \tag{1.32}$$

To understand its meaning easier, we can consider a function only of α,

$$y(\alpha) = \phi(x + \alpha p). \tag{1.33}$$

The left hand side of (1.32) is derivative $y'(\alpha)$ and the right hand side is $y'(0)$. Condition (1.32) therefore requires that the slope in the new point be c_2 times greater than the original slope. That is, if we are in a region where the original slope $y'(0)$ has an important negative value, it is likely we can find an α that will give us a slope $y'(\alpha)$ that also has an important negative value. If we are in a region where $y'(0)$ is small, meaning the misfit function is rather flat as a function of α, it is likely the the new slope $y'(\alpha)$ will give us a similar value. Therefore, this second condition generally prevents stopping in the case where $y'(0)$ has an important

value, and it is therefore reasonable to suspect that more progress can be attained by increasing α.

Summarizing, the two Wolfe conditions are

$$\begin{aligned} \phi(x_k + \alpha_k p_k) &\leq \phi(x_k) + c_1 \alpha_k \nabla f_k^T p_k, & (1.34) \\ \nabla f(x_k + \alpha_k p_k)^T p_k &\geq c_2 \nabla f_k^T p_k, & (1.35) \end{aligned}$$

where $0 < c_1 < c_2 < 1$. Nocedal and Wright (1999) report that typical values of c_2 for Newton or quasi Newton methods is $c_2 = 0.9$, and $c_2 = 0.1$ for conjugate gradient methods. Constant c_1 is normally given a very small value, for example $c_1 \approx 10^{-4}$.

2.2 Search Directions

Given iterations of the form,

$$x_{k+1} = x_k + \alpha_k p_k \quad (1.36)$$

we seek directions p_k that will make the misfit function decrease $\phi(x_{k+1}) < \phi(x_k)$. The condition on the descent direction is that $p_k^T \nabla \phi < 0$. It is easy to see why this condition must be satisfied by performing a first order Taylor expansion of the misfit function

$$\phi(x_k + \alpha_k p_k) \approx \phi(x_k) + \alpha p_k^T \nabla \phi_k + o(\alpha_k^2 ||p_k||^2). \quad (1.37)$$

If p_k is a descent direction then

$$\phi(x_k + \alpha_k p_k) - \phi(x_k) < 0, \quad (1.38)$$

which implies $\alpha p_k^T \nabla \phi_k < 0$. Since $\alpha > 0$, we conclude that

$$p_k^T \nabla \phi_k < 0. \quad (1.39)$$

2.2.a *Steepest Descent*

The most natural descent direction is the steepest descent,

$$p_k = -\nabla \phi_k. \quad (1.40)$$

It is natural because to move from x_k to x_{k+1} one chooses the direction in which the maximum decrease of ϕ is found. The descent condition (1.39) is satisfied since $p_k^T \nabla \phi_k = -\|\nabla \phi_k\|^2 < 0$. For the non linear setting, under mild regularity assumptions, gradient descent algorithms can be shown to have sub linear rate of global convergence in terms of the quantities $|\nabla \phi(x^k)|^2$ (Nemirovski, 1999).

2.2.b *Linear conjugate gradient*

Linear and non linear conjugate gradient methods have many variants and are amongst the most commonly used optimization methods because they are faster than steepest descent methods, are very easy to implement and require very little memory storage. For an introduction to this method see for example Shewchuk (1994).

Linear conjugate gradient methods are very useful to solve large scale linear systems of equations. For example, if we wish to solve the linear system $Ax = b$, where A is symmetric positive definite, this can be equivalently posed as an optimization problem

$$\min_x \phi(x) = \min_x \left(\frac{1}{2} x^\dagger A x - b^\dagger x \right)$$

since we know that at the optimal x^*,

$$0 = \nabla\phi(x^*) = Ax^* - b.$$

Now, we say that two directions p_i, p_j are A-conjugate if

$$p_i^\dagger A p_j = 0, \qquad i \neq j. \tag{1.41}$$

The directions $\{p_i\}$ are linearly independent and must span the whole space. Each direction p_n is a linear combination of the steepest descent direction and the previous direction p_{n-1},

$$p_n = -\nabla\phi_n + \beta_n p_{n-1}, \tag{1.42}$$

where p_n and p_{n-1} are A-conjugate and

$$\alpha_n = -\frac{\nabla\phi_n^\dagger \nabla\phi_n}{p_n^\dagger A p_n}, \tag{1.43}$$

$$\beta_{n+1} = \frac{\nabla\phi_{n+1}^\dagger \nabla\phi_{n+1}}{\nabla\phi_n^\dagger \nabla\phi_n}. \tag{1.44}$$

The linear conjugate gradient algorithm will converge in at most N iterations, if N is the dimension of the matrix A. When the matrix is well-conditioned the matrix A the convergence may be attained in fewer iterations.

For general non-convex functions ϕ, an adaptation of the linear conjugate gradient algorithm was done by Fletcher and Reeves (Nocedal and Wright, 1999). The main modification consists in performing a line search to determine the step length α. Otherwise, the algorithm is similar. There are many variations of the conjugate gradient method that provide different expressions for β. A widely known variant is one proposed by Polak and Ribiére where

$$\beta_{n+1} = \frac{\nabla\phi_{n+1}^\dagger \left(\nabla\phi_{n+1} - \nabla\phi\right)}{\|\nabla\phi\|^2}. \tag{1.45}$$

A common practice when using non linear conjugate gradient is to perform restarts periodically every m iterations where β is set to zero and thus a steepest descent direction is taken. Other criteria that are used to restart the algorithm are based on checking the orthogonality of consecutive gradients. When two consecutive directions do not satisfy an orthogonality criteria, the method is restarted. This allows for useless information to be erased, and leads to n-step quadratic convergence (Nocedal and Wright, 1999).

2.2.c *Newton methods*

Amongst other options we find Newton directions which can be derived by doing a Taylor expansion of $\phi(x_k + p)$ up to second order:

$$\phi(x_k + \alpha_k p_k) \approx \phi(x_k) + \alpha_k p_k^T \nabla\phi_k + \frac{1}{2}\alpha_k^2 p_k^T \nabla^2\phi(x_k) p_k. \tag{1.46}$$

Deriving the expansion with respect to p_k, we obtain

$$\left(\nabla^2\phi_k\right)p_k = -\nabla\phi_k, \tag{1.47}$$

or, if the inverse of $\left(\nabla^2\phi_k\right)$ exists,

$$p_k = -\left(\nabla^2\phi_k\right)^{-1}\nabla\phi_k. \tag{1.48}$$

The second derivative $\nabla^2\phi_k$ is known as the Hessian, and is denoted by $H = \nabla^2\phi_k$. As long as the Taylor expansion (1.46) is close to $\phi(x_k + \alpha_k p_k)$, that is, as long as the misfit function can be locally approximated by quadratic function, and as long as $\nabla^2\phi(x_k)$ is positive definite, the Newton direction (1.48) is a reliable descent direction. We can easily verify this by multiplying (1.48) by the transpose of the gradient $\nabla\phi_k^T$, in which case we obtain

$$\nabla\phi_k^T p_k = -\nabla\phi_k^T\left(\nabla^2\phi_k\right)^{-1}\nabla\phi_k. \tag{1.49}$$

If $\nabla^2\phi(x_k)$ is positive definite, then $\nabla\phi_k^T p_k^N < 0$, which guarantees a descent direction. However, if the second derivative is not positive definite several problems may arise. First, its inverse may not be defined and, in addition since the descent direction property $\nabla\phi_k^T p_k^N < 0$ may not be satisfied, therefore p_k^N may not make the function ϕ decrease.

The Hessian measures the change in the gradient when we move slightly in the parameter space. Thus, roughly speaking, including the Hessian information in the optimization methods allows for a the minimization to be less short sighted. The advantages of Newton methods are seen in quadratic convergence rates, and the biggest inconvenient is the need to compute the Hessian, and its inverse. An alternative lies in the quasi Newton methods that do not compute, but rather approximate the Hessian, and still attain super linear convergence rates (Kelley, 1999). The idea behind quasi Newton methods lies on the fact that information about the second derivatives can be extracted from changes in the gradient. Other Newton methods, such as Gauss Newton or truncated Newton, allow for Hessian (or inverse Hessian) approximations to be performed.

2.2.d *BFGS*

The BFGS (Broyden, Fletcher, Goldfarb, Shanno) method is a quasi Newton method that provides and estimate on $\nabla^2\phi(x)$, and updates the approximation as the iterations progress. It can be shown that this methods accumulates the approximate curvature in a sequence of expanding subspaces. This way, the approximate Hessian is represented using a smaller reduced matrix that increases its dimension in every iteration. Therefore, the explicit Hessian, or its inverse, are never stored in memory. Products of the inverse of the Hessian with a vector are formed (Nocedal and Wright, 2006).

Under the approximation that the function ϕ can be locally approximated by a quadratic function, say

$$\psi_k(p) = \phi_k + \nabla\phi_k^T p + \frac{1}{2}p^T B_k p, \tag{1.50}$$

where B_k is a symmetric positive definite (s.p.d) matrix that is updated in every iteration. Taking the derivative, we obtain

$$\nabla\psi_k(p) \quad = \quad \nabla\phi_k^T + B_k p, \tag{1.51}$$

and at $p = 0$, we have

$$\nabla\psi_k(0) \quad = \quad \nabla\phi_k. \tag{1.52}$$

Keeping in mind that $x_{k+1} = x_k + \alpha_k p_k$, it must hold that $\nabla\psi_{k+1}(p = -\alpha_k p_k) = \nabla\psi_k(0)$. From (1.51) we know that,

$$\nabla\psi_{k+1}(p = -\alpha_k p_k) \quad = \quad \nabla\phi_{k+1} - \alpha_k B_{k+1} p_k, \tag{1.53}$$

and using the equality between (1.52) and (1.53), it is found,

$$B_{k+1}\alpha_k p_k = \nabla\phi_{k+1} - \nabla\phi_k. \tag{1.54}$$

Let

$$s_k \quad = \quad x_{k+1} - x_k \tag{1.55}$$
$$y_k \quad = \quad \nabla\phi_{k+1} - \nabla\phi_k, \tag{1.56}$$

where with this notation we arrive the formula that is known as the secant equation,

$$B_{k+1}s_k = y_k, \tag{1.57}$$

which also explains why BFGS is sometimes referred to as a secant method.

☛ It is crucial to satisfy the curvature condition

$$s_k^T y_k > 0, \tag{1.58}$$

which for strong convex functions is always true.

☛ However, for non convex functions the curvature condition may not always hold, so we must impose it as a restriction, in order to maintain the positive definiteness of B. What is impressive, is that for the case of unconstrained optimization, enforcing the curvature condition 1.58, is equivalent to the second Wolfe condition (1.35). For constrained optimization this is may not be the case (Gilbert, 1997). For unconstrained optimization, it is easy to verify the equivalence between the two, by starting with the Wolfe condition (1.35), $\nabla\phi_{k+1}^T s_k \geq c_2\nabla\phi_k^T s_k$, which leads to

$$y_k^T s_k \geq (c_2 - 1)\alpha_k \nabla\phi_k^T p_k. \tag{1.59}$$

The right had side is always positive as long as the descent direction condition holds $\nabla\phi_k^T p_k < 0$, since $c_2 < 1$. Let $H_k = B_k^{-1}$, called the Hessian. We also require the Hessian to be s.p.d, and a secant equation can also be imposed,

$$H_{k+1}y_k = s_k. \tag{1.60}$$

The update formula for H_{k+1} is (Nocedal and Wright, 2006; Kelley, 1999),

$$H_{k+1} = \left(\mathbb{I} - \frac{s_k y_k^T}{y_k^T s_k}\right) H_k \left(\mathbb{I} - \frac{y_k s_k^T}{y_k^T s_k}\right) + \frac{s_k s_k^T}{y_k^T s_k}. \tag{1.61}$$

If one begins with and initial H_0 that is s.p.d, and if the curvature condition is satisfied in each iteration, all the Hessian matrices will remain s.p.d. This is stated in the lemmas below that are taken from (Kelley, 1999), and whose proofs can be found there.

Lemma 2.1. Let H_k be s.p.d , $y^T s \neq 0$, and H_{k+1} be given by 1.61. Then, H_{k+1} is non singular.

Lemma 2.2. Let H_k be s.p.d , $y^T s > 0$, and H_{k+1} be given by 1.61. Then, H_{k+1} is s.p.d.

☛ Lemmas 2.1 and 2.2 allow us to iterate, and be sure that by imposing the curvature condition in each iteration, the updated Hessians will remain s.p.d.

Global convergence is not guaranteed for general non linear function. Theoretical results on global convergence has been shown under the assumption of convex function, or on the certain properties on the iterates. However, local convergence is guaranteed under certain conditions, and the proofs and convergence rates are detailed in (Kelley, 1999; Nocedal and Wright, 2006). To highlight, the local convergence results rely on the assumption that the Hessian matrix is Lipschitz continuous, that the approximate Hessians are not too far from the exact Hessian ($||H^{-1} - \nabla^2\phi(x^*)^{-1}|| \leq \delta_0$), and that the initial point is not too far from the optimum point ($||x_k - x^*|| \leq \delta_0$), then the sequence of BFGS iterates converges to x^* with a super linear rate. However, if x_k is near x^*, but H is far from $\nabla^2\phi(x^*)$, there is no guarantee of the convergence.

2.2.e *Limited memory BFGS*

Quasi Newton methods such as the BFGS in 2.2.d, represent an advantage with respect to pure Newton methods, because the Hessian matrix does not need to be explicitly computed. However, they are still demanding in terms of computational resources because the gradients $\nabla\phi$ and variables x_k need to be stored for *all* previous iterations. In either case, the storage requirements for large scale problems quickly becomes inaccessible as iterations proceed, and modifications or extensions of quasi Newton methods need to be introduced.

Traditional BFGS methods approximate the curvature information in a sequence of expanding subspaces. That is, the approximate Hessian is represented using a smaller reduced matrix that increases its dimension in every iteration. When the number of variables is large, a possible modification consists in limiting the dimension of the subspace of the approximate Hessian to save memory space. Methods such as limited memory reduced Hessians (Gill and Leonard, 2003), or limited memory BFGS (Byrd et al., 1995) methods have been developed.

Numerically it has been proved that limited memory BFGS quasi Newton methods are efficient for the minimization of constrained and unconstrained non-linear functions (Liu and Nocedal, 1989; Byrd et al., 1995) [2], generally due to its low iteration cost. Non linear conjugate gradient methods may be also used, as the memory requirement is reasonable, but they are generally less robust and less efficient on non-linear problems (Gill and Murray, 1979; Nocedal and Wright, 2006). The limited memory BFGS method is based on the idea that the approximate Hessian can be constructed with information regarding only the past m iterations, considering it the most relevant and thus reducing the memory cost. On the other hand, they have the disadvantage of showing a slower convergence than the full BFGS methods, having linear and not quadratic convergence rates. The approximate Hessians may be used in quasi Newton methods, or simply as pre-conditioners, for example in non-linear conjugate gradient methods.

Similar to the update formula for the full BFGS algorithm, for the limited memory BFGS the formula for H_{k+1} is (Nocedal and Wright, 2006),

$$H_{k+1} = \left(\mathbb{I} - \frac{s_k y_k^T}{y_k^T s_k}\right) H_k \left(\mathbb{I} - \frac{y_k s_k^T}{y_k^T s_k}\right) + \frac{s_k s_k^T}{y_k^T s_k}. \tag{1.62}$$

Where $k \in [i-1, i-m]$, where m are is the memory size, and i indicates the current iteration to be updated, $i = k+1$. Expressions for s_k and y_k are given in 1.55,1.56. An efficient computational

2. Theoretically, global convergence on uniform convex function has been shown (Liu and Nocedal, 1989)

algorithm to evaluate (1.62) is given in Nocedal and Wright (2006), that takes advantage of its recursive form. To choose H_k^0, a method that has proved to be useful is $H_k^0 = \gamma_k \mathbb{I}$, where (Nocedal and Wright, 2006),

$$\gamma_k = \frac{s_{k-1}^T y_{k-1}}{y_{k-1}^T y_{k-1}}. \tag{1.63}$$

The scaling γ_k attempts to scale the search direction p_k in such a way that α_k is generally accepted.

Through numerical experiments we may observe that l-BFGS approaches the solution fast initially but then tends to slow down close to the optimum. Perhaps, this is due to the fact that close to the optimum the quadratic approximation of the misfit function is no longer an appropriate model because the effect of the noise is more dominant.

2.2.f *Errors with the approximate Hessian, and restart procedures*

Restart procedures were originally introduced in conjugate gradient algorithms by Beale (1972); Powell (1977), who showed with numerical examples that periodic restarts boosted the efficiency of the algorithm by improving the convergence rate. Criteria were defined to measure whether the orthogonality between two consecutive gradients had been lost and a reset was needed. A general disadvantage of restarting the search direction along the steepest descent is that the reduction in the objective function is usually less than it would be without restarts. Some methods to overcome this setback, propose to restart in a direction proportional to the most recent descent direction (Beale, 1972). Non linear conjugate gradient methods also suffer from a degradation in the computed search direction. Convergence rates are studied by Neumaier (1997), where a restarting procedure is defined based on the line search accuracy and on the deterioration of the conjugate property between gradients. Restarting in a conjugate gradient algorithms with a scaled BFGS preconditioner for unconstrained optimization has also been used (Andrei, 2007), where the restarting criteria is also based on the loss of orthogonality between current and previous gradients.

As with conjugate gradient algorithms, numerical tests with BFGS algorithms have also shown that restarts improve the performance of these optimization methods. As is explained in 2.2.d, certainly, if the Wolfe conditions are not ensured in the line search of the BFGS, and instead the Armijo backtracking line search is implemented, there is no guarantee that the curvature condition will be satisfied. This is the case considered in (Kelley, 1999), where they propose a BFGS-Armijo algorithm. Every time $y^T s$ is not sufficiently positive, the history is of Hessians is thrown away, and the algorithms is restarted. Another possible approach that is suggested is to try to keep as much information as possible, instead of clearing all the memory.

Despite ensuring the Wolfe conditions, which guarantee the approximate Hessian will be spd, there may be cases where BFGS algorithms do not produce satisfactory results. That is, H_k becomes an inaccurate approximation of the Hessian. This could be the case when the product $y_k^T s_k$ is positive but very small. Since this quantity is in the denominator of 1.61 , H_{k+1} may be arbitrarily large. Another reason that could lead to a poor approximate Hessian is the accumulation of numerical errors, for example due to rounding errors, or errors in the the gradient calculation. There are some self correcting properties related to the BFGS methods, but these rely strongly on an efficient line search, and on the quadratic approximation of the misfit function to adequately represent the curvature of the function (Nocedal and Wright, 2006). However, it is also affirmed here that "computational observations strongly suggest that it is more economical,

in terms of function evaluations, to perform a fairly inaccurate line search". Which explains why the values of $c_1 \approx 10^{-4}$ and $c_2 \approx 0.9$ are commonly used in the Wolfe conditions. There are two competing priorities here because on one hand, an adequate line search to stabilize the algorithm the positive definiteness is needed, and the other hand it is not efficient to perform many function evaluations.

The subroutines for limited memory BFGS with bounded constraints provided by Zhu et al. (1997) has included several mechanisms to deal with a deteriorated performance, when rounding errors start to dominate the computation. If the line search performs 20 function evaluations, and the objective function has not decreased, it is considered that the current direction is not useful and all the vectors s_k and y_k are thrown away and the iteration is restarted along the steepest descent direction. If the algorithm also fails along the steepest descent direction, the algorithm terminates. The authors thereby report that this type of failure occurs mainly when rounding errors begin to dominate, which can arise, for example when the required accuracy is high, and that restarting sometimes results in a successful termination, but not always. The other cases when all the limited memory vectors are erased, and the iteration is restarted with a steepest descent direction, is when either the L-BFGS matrix becomes singular or indefinite, or if the descent direction found by the algorithm, does not satisfy $\nabla\phi p < 0$. Since the subroutine that is provided by Zhu et al. (1997) includes bound constraints, in the case where there are variables that actually hit the boundaries satisfying the Wolfe condition, does not mean that the curvature conditions will be satisfied (Gilbert, 1997). In this scenario, to guarantee that the Hessian remains spd, the BFGS update is skipped if

$$\frac{y_k^T s_k}{-g_k^T s_k} \leq \epsilon, \tag{1.64}$$

where ϵ is the machine precision. A small value of $y_k^T S_k$ could lead to the restarting described above.

More recently Schittkowski (2011) also included a restart procedure in a constrained non linear optimization problem using a full BFGS method. The author studies the case of non linear optimization under the influence of noise. To make the inversion robust, a non monotone line search is used, in which the misfit function is allowed to increase in certain cases, and a scaled restart procedure that is performed either periodically, or when the inversion ends with a failed termination. Numerical tests conclude that they are able to solve 90% of in case of extremely noisy function values, opposed to only 30% of success without the restart and monotone line search.

☛ Summarizing, restart methods have long been used because the optimization methods are very sensitive to round-off errors in function or gradient evaluations. Recall that the convergence of limited memory BFGS is local, and relies on having a good initial guess, and that the approximate Hessian is a good approximation of the real Hessian. Since, we can not always satisfy these conditions, the super linear convergence behaviour we expect to obtain, may never occur. In addition, if the true Hessian is singular or severely ill conditioned, the BFGS updates may become more ill conditioned, which augments the effect of round off errors, and deteriorates with iterations. At some point then, the updated matrix may be useless. This explains why restarting may have a good influence, since adopting the steepest descent direction may be better than using a deteriorated direction, damaged because of the Hessian. Another possible reason why restart improves the behaviour is that particularly in non linear conjugate gradient methods and limited BFGS methods, the search direction is found by conjugation properties of gradient and errors may have led to search in a degraded subspace. Recovering from a degradation of this

subspace can be achieved by restarting and searching in a direction proportional to the steepest descent.

2.3 Preconditioner

The convergence of the iterative methods to solve linear systems like (1.47) depends on the condition number κ which is defined as the ratio between the largest and the smallest eigenvalue, and measures the ellipticity (curvature) of the iso surfaces of the misfit function, which determines the complexity of the minimization procedure. As the iso-surfaces become less circular, the gradient points away from minimum which is in the center. These deviations in the minimization directions slow down the optimization process.

To improve the spectral properties, it is possible to multiply the system on the left hand side with an operator $\mathcal{P}$ known as a left preconditioner. There is no general rule to find a good preconditioner as it depends on the properties of the operator to which the preconditioner is applied. If the preconditioner is suitable, the condition number of $\mathcal{P}H$ is better than that of the original operator (we recall that $H = \nabla^2\phi$). For example, ideally $\mathcal{P}H \propto \mathbb{I}$, where all the eigenvalues are all the same, and the condition number is one. Thus, generally $\mathcal{P}^{-1}$ should resemble the original matrix H (or equivalently $\mathcal{P}$ should resemble H^{-1} and, ideally, should not require a great additional computational effort.

The preconditioner $\mathcal{P}$ can also be used with gradient algorithms, where the preconditioner is multiplied with the gradient to provide a better scaling that will allow a faster descent. For l-BFGS optimization methods, the preconditioner can be used in equation (1.62) as the initial estimation H_k^0. This has also been numerically shown to improve the converge of the algorithm (Brossier, 2011; Métivier et al., 2013b).

Chapter 2

Full Waveform Inversion

1 Seismic imaging with full waveform inversion

Seismic imaging is a parameter estimation inverse problem aimed to reconstruct the parameters that govern wave propagation in the earth such as compressional and shear wavespeeds, density and attenuation, from the recorded data of the earth's displacement at the surface Lailly (1984); Tarantola (2005); Nolet (2008). It is therefore often posed as an optimization problem where the purpose is to find the set of model parameters that minimizes a misfit function $\phi(u, m)$, that quantifies the distance between the measurements and the data predicted by the estimated model m. Different definitions of the misfit function lead to different reconstruction methods using different imaging conditions. Data domain methods define a misfit function based on characteristics of data (Tarantola, 1984b). Alternatively, image-domain methods define a misfit based on the quality of the reconstructed images (Shen and Symes, 2008; Symes, 2008). Within the data domain methods, amongst the most popular methods we find least squares migration, travel time tomography and full waveform inversion.

This chapter describes the main components of FWI, and with the help of some simple examples, the physical interpretation of the gradient and the Hessian is given. Other seismic imaging methods based on the same physical principles as FWI are described briefly in Appendix 1. Details of the gradient and Hessian equations derived using the adjoint state method can be found in Appendix 3, 4 and 5.

1.1 Full waveform inversion

Full waveform inversion (FWI) is a data domain imaging method. In comparison to travel time tomography (2), the purpose is to not only fit the phase of one seismic arrival time but to fit the entire waveform (or the entire waveform within a window). As described in section 0.4 of chapter 1, a typical inverse problems consists of two main parts: the direct problem and the inverse problem. In FWI, the direct problem consists in solving the wave equation on a domain

Ω. In the time domain, the acoustic wave equation is,

$$\begin{cases} \left(\nabla^2 - \dfrac{1}{v^2}\dfrac{\partial^2}{\partial t^2}\right) u(x,t) & = s(x,t) \qquad (x,t) \in \Omega \times [0,\infty), \\ u(x,0) & = 0 \\ \dfrac{\partial u(x,t)}{\partial t} & = 0, \end{cases} \tag{2.1}$$

where $u(x,y)$ is the (pressure) wavefield, $s(x,t)$ is the source function, $v(x) = 1/\sqrt{\rho(x)\kappa(x)}$ where κ is the compressibility, ρ is the density, v is the wave velocity. Here, the model m consists of one physical parameter : $m = \{1/v^2(x)\}$. The forward problem (2.1) can be written in compact form,

$$Au = s \tag{2.2}$$

where

$$\boxed{A(x,t,m) = \left(\nabla^2 - m(x)\frac{\partial^2}{\partial t^2}\right).} \tag{2.3}$$

The inverse problem consists in finding the optimal model parameters m that minimize the misfit function ϕ. The misfit function ϕ in FWI is a function of the whole observed waveform d and the whole computed waveform u on the receiver positions. Common choices for ϕ involve cross-correlation functions and $\|\cdot\|_p$ norms ,

$$\phi = Pu \otimes d \tag{2.4}$$

$$\phi = \|Pu - d\|_p, \tag{2.5}$$

with $p = 1, 2$ (Brossier et al., 2010). Using the $\|\cdot\|_2$ norm for the definition of ϕ the inverse problem is

$$\boxed{\min_m \phi(u;m) = \min_m \frac{1}{2}\int_{t=0}^{T}\sum_{s=1}^{N_s}\sum_{r=1}^{N_r}\|Pu_{s,r}(x,t) - d_{s,r}(x,t)\|_2^2,} \tag{2.6}$$

where N_s is the number of sources, N_r is the number of receivers, $d(x,t)$ is the observed data at the receiver positions ($d : \Omega_r \to \mathbb{R}$), u is the solution of the forward problem (53), and P is projection operator $P : \Omega \to \Omega_r$, where Ω_r is the receiver space.

Solving the inverse problem is thus an optimization problem, and we will employ the iterative optimization methods described in section 2 of chapter 1. For a given model in iteration n, the updated model is (equation 1.27)

$$m_{n+1} = m_n + \alpha_n \Delta m_n \tag{2.7}$$

where in its most general form Δm_n is given by the Newton equation (1.48),

$$\Delta m_n = -\left(\nabla_m^2 \phi_n\right)^{-1} \nabla_m \phi_n. \tag{2.8}$$

FWI can also be solved in the frequency domain, by applying a Fourier transform to $u(x,t)$ and $d(x,t)$. The equivalent inverse problem is

$$\boxed{\min_m \phi(u;m) = \min_m \frac{1}{2}\sum_{i}^{N_f}\sum_{s=1}^{N_s}\sum_{r=1}^{N_r}\|Pu_{s,r}(x,\omega_i) - d_{s,r}(x,\omega_i)\|_2^2,}, \tag{2.9}$$

where N_f is the number of discrete frequencies.

1.2 The gradient

With any optimization method (2), finding the set of optimal model parameters requires the derivative of the misfit function ϕ,

$$\boxed{\frac{\partial \phi}{\partial m} = g = \int_{t=0}^{T} \sum_{s=1}^{N_s} \sum_{r=1}^{N_r} \left(P \frac{\partial u_s(x,t)}{\partial m} \right)_r^{\dagger} \left((P u_s)_r(x,t) - d_{s,r}(x,t) \right).} \tag{2.10}$$

We thus need to compute $\partial u(x,t)/\partial m$. In the most general case, this can not be done analytically because we do not have the analytic solution of the wave equation for a general complex media. Therefore, this term will have to be computed numerically. We thus discretize the domain Ω in N grid points. Assuming we only have one physical parameter to invert (for example the density ρ), the discretized model m will also have N components. Numerically, each $\dfrac{\partial u(x,t)}{\partial m_i}$ can be computed using a first order derivative approximation, or taking the derivative of the expression for the direct problem (Au=s) and solving a direct problem:

$$\frac{\partial u(x,t)}{\partial m_i} : \begin{cases} \dfrac{\partial u(x,t)}{\partial m_i} = \dfrac{u(x,t,m+\Delta m_i) - u(x,t,m)}{\Delta m_i} & \text{for } i = 1 \dots N \\ A \dfrac{\partial u}{\partial m_i} = -\dfrac{\partial A}{\partial m_i} u. & \text{for } i = 1 \dots N \end{cases} \tag{2.11}$$

The second option is numerically more accurate and requires the solution of $N \times N_s$ direct problems for a gradient calculation. As can be seen, $\dfrac{\partial u}{\partial m}$ has the same form of the forward problem in (2.2) with a modified source term (Pratt et al., 1998). Physically, $\dfrac{\partial u(x,t)}{\partial m_i}$ represents a diffracted wavefield by a heterogeneity in the model at position i.

Alternatively, the gradient can be found using the *adjoint state method* that renders the gradient computation independent of the number of model parameters, and only requires the solution of two forward problems per source : $2 \times N_s$ (Lions, 1968; Plessix, 2006; Chavent, 2009). The gradient equations in the time and frequency domain for FWI using the adjoint state method are widely used (Plessix, 2006), and we only outline the procedure.

- We defined an inner product here as $\langle f(x,t), g(x,t) \rangle = \int_0^T \int_\Omega f^*(x,t) g(x,t) dt dx$, but the definition of inner product will vary depending if we are working in time or frequency domain.
- A real valued Lagrangian is defined which contains the misfit function, and the restriction on u , $(Au = s)$,

$$\begin{aligned} \mathcal{L}(u, u^\dagger, \lambda, \lambda^\dagger, m) &= \phi(u;m) + \frac{1}{2} \langle \lambda, A(m)u - s \rangle + \frac{1}{2} \langle A(m)u - s, \lambda \rangle \\ &= \frac{1}{2} \langle Pu - d, Pu - d \rangle + \frac{1}{2} \langle \lambda, A(m)u - s \rangle + \frac{1}{2} \langle A(m)u - s, \lambda \rangle , \end{aligned}$$

  where λ are adjoint state variables. To find the saddle point of $\mathcal{L}$, we take its derivative with respect to $\lambda^\dagger$, $u^\dagger$ and m.
- The derivative $\partial \mathcal{L} / \partial \lambda^\dagger = 0$ gives,

$$A(m)u(x,t) = s \qquad t \in [0,T], \tag{2.12}$$

  which simply corresponds to the equation of the forward problem that we included as a constraint.

– The derivative $\partial\mathcal{L}/\partial u^\dagger$ requires two integration by parts in the time domain to isolate the variable u, and transfer the time derivatives on the variable λ. Once this is done, $\partial\mathcal{L}/\partial u^\dagger = 0$ gives,

$$\begin{cases} A^\dagger\lambda(x,t') & = -P^\dagger(Pu(t') - d(t')) \qquad t' \in [0,T], \\ \lambda(x,T) & = 0 \\ \dfrac{\partial\lambda(x,t')}{\partial t}|_{t'=T} & = 0 \end{cases} \tag{2.13}$$

known as the adjoint equation because it gives the equation of motion for the adjoint variable λ. Notice that the *adjoint* wavefield λ is a *back-propagated wavefield*, where the source term is the residual wavefield at the residual positions, and the modeling operator is not A but $A^\dagger$. That is, we have the *final condition for* $\lambda(x,t'=T)$*, and we will solve the equation back in time until* $\lambda(t'=0)$ *(hence the notion of adjoint).* This back-propagation takes the residuals at the receiver positions and retraces the trajectory back in time. By doing a change of variable

$$t' = T - t,$$

the above system of equations can be solved using an initial condition,

$$\begin{cases} A^\dagger\lambda(x,T-t) & = -P^\dagger(Pu(T-t) - d(T-t)) \qquad t \in [0,T], \\ \lambda(x,0) & = 0 \\ \dfrac{\partial\lambda(x,t')}{\partial t}|_{t=0} & = 0 \end{cases} \tag{2.14}$$

– Finally, the derivative with respect to m whenever the constraints are satisfied gives the gradient for one source,

$$\boxed{\nabla_m\phi(u;m) = \int_{t=0}^{T} \Re\left\{ u^\dagger(x,t)\left(\frac{\partial A}{\partial m}\right)^\dagger \lambda(x,t)\right\} dt.} \tag{2.15}$$

From equation (2.3), one obtains that $\dfrac{\partial A}{\partial m} = -\dfrac{\partial^2}{\partial t^2}$, leading to

$$\boxed{\nabla_m\phi(u;m) = -\int_{t=0}^{T} \Re\left\{ \frac{\partial^2 u}{\partial t^2}(x,t)\lambda(x,t)\right\} dt.} \tag{2.16}$$

The gradient in equation (2.15) or (2.15) is proportional to the cross-correlation of the direct wavefield $u(x,t)$ and the back-propagated wavefield $\lambda(x,t)$ weighted by the radiation term $\partial A/\partial m$. The real part ensures that the model parameters have a physical meaning.

The equivalent expression for the gradient using one source in the frequency domain is the Fourier transform of 2.15,

$$\boxed{\nabla_m\phi(u;m) = \sum_{\omega_i}^{N_f} Re\left\{ u^\dagger(x,\omega_i)\left(\frac{\partial A}{\partial m}\right)^\dagger \lambda(x,\omega_i)\right\},} \tag{2.17}$$

and a more detailed development of the gradient computation in the frequency domain is shown in Appendix 3.

Note that many of the steps in the gradient computation in the adjoint state method are simple because the wave equation operator (2.3) is self-adjoint ($A = A^\dagger$). This is due to the fact

that the model parameters that depends on the space variables $m(x)$ can all factored beside the time derivatives, and that there is only a time derivative with even exponent. When either of these ingredients are missing, the gradient computation may be more convolved. An example occurs when (2.3) is written as two first-order coupled partial differential equations. In this case, the operators are no longer self-adjoint. We show the gradient computation with the adjoint state method for a 3D isotropic viscoelastic media using the fist order stress-velocity formulation in Appendix 4, and can be found in (Castellanos et al., 2011). An extension to the anisotropic velocity-stress equations has been done by Brossier et al. (2013a).

1.2.a *Preconditioner*

As mentioned in Section 2.3, the descent direction can be improved by using a preconditioner. There is no general rule to choose a preconditioner, but the computational cost must not be too elevated compared to the computation of the gradient. For gradient descent algorithms the left preconditioned descent direction would have the form

$$\Delta m = \mathcal{P}\nabla_m \phi. \tag{2.18}$$

Comparing this descent direction (2.18) with the Newton descent direction (1.48) ($\Delta m = H^{-1}\nabla_m\phi$), one could think that a good preconditioner would be the inverse Hessian. However, in general the computational cost of the preconditioner must remain low, and the computing the inverse Hessian can dramatically increase the computational cost. This is because the Hessian needs the diffracted wavefields $\partial u/\partial m$, which would require solving additional forward problems, and making the preconditioner expensive to compute, as will be seen in the following section.

Therefore, a choice of preconditioner that has shown to be useful in FWI was introduced by Shin et al. (2001a) for depth migration. This preconditioner only uses the direct wavefields u (which we already computed to find the gradient), and the radiation patterns $\partial A/\partial m$ (for which we have analytical expressions). Shin et al. (2001a) thus computes what he calls a pseudo Hessian by approximating the diffracted wavefields by (Shin et al., 2001a)

$$\frac{\partial u}{\partial m} = -\frac{\partial A}{\partial m}u, \tag{2.19}$$

as if the modelling operator was the identity $A = \mathbb{I}$. Note that the right hand side $(\partial A/\partial m)u$ physically represents a virtual source located at the point in the space where $\partial A/\partial m$ are non zero. The resulting pseudo Hessian H^{ps} is,

$$H^{ps} = \left(\frac{\partial A}{\partial m}u\right)^{\dagger}\left(\frac{\partial A}{\partial m}u\right). \tag{2.20}$$

Therefore, the pseudo Hessian approximates the true Hessian (which consists of the correlation of diffracted wavefields) by a correlation of virtual sources. Since an approximation for the inverse of the Hessian is desired, and not an approximation of the Hessian itself, only the diagonal terms of (2.20) are taken into account to make it easy to invert. For numerical stability, a threshold value β is added to the diagonal terms of the pseudo Hessian before taking their reciprocal values to avoid division by very small numbers, leading us to

$$\mathcal{P}^n_{pp} = \operatorname{diag}\left(\frac{1}{H^{ps}_{pp} + \beta}\right), \tag{2.21}$$

where $\beta = \theta \max(H^{ps}_{ii})$, for $\theta \in [0, 1]$, and the index n is to stress the fact that the preconditioner changes in each iteration of the non linear inverse problem. This choice of preconditioner satisfies

the two general conditions of being easy to find and invert since we did not have to compute any additional wavefields, and being similar to the inverse of the original matrix, since we have taken the inverse of the diagonal of pseudo Hessian.

The same preconditioner is used with gradient algorithms, where the preconditioner is multiplied with the gradient to give the descent direction, which resembles a first approximation to a Gauss - Newton step. For the l-BFGS method, the preconditioner is used as an initial estimation of the Hessian $H_0^n = \mathcal{P}$ in each iteration, also helping the convergence of the optimization (Métivier et al., 2013b).

1.2.b *Physical interpretation of the gradient*

We refer to the imaging condition as the operator that goes from the data space to the model space, and determines which parameters in the model space must be modified in order to explain the measured data. The gradient (2.10) or (2.15) gives the imaging condition for FWI.

In discrete form, let the domain Ω consist of N points. The gradient is a real vector of N components, $g \in \mathbb{R}^N$. Using the gradient expression (2.15), we illustrated in the Introduction that the correlation of the direct and back-propagated wavefield will be non-zero on all the points along an ellipse, as that shown in Figure 2.1a.

(a)

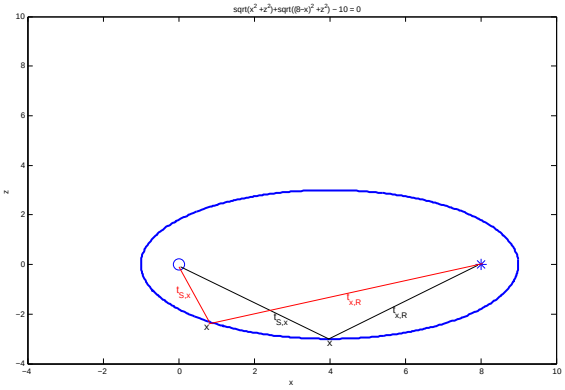


Figure 2.1: a) Example of the points satisfying the imaging condition $t_{S,x} + t_{x,R} = T$, for a reflection residual at time $T = 5s$, in a model with $v = 2m/s$, with a source at $x_S = 0$ and a receiver at $x_R = 8\ m$, resulting in an offset of $D = 8m$. The points form an ellipse with focal points at the source and receiver positions.

Now we will focus in giving a physical interpretation with the gradient equation (2.10),

$$g = \int_{t=0}^{T} \sum_{s=1}^{N_s} \sum_{r=1}^{N_r} \left(P \frac{\partial u_s(x,t)}{\partial m} \right)_r^{\dagger} \left((P u_s(x,t))_r - d_{s,r}(x,t) \right). \tag{2.22}$$

We recall that u is the wavefield propagating in the medium, generated by s as shown in Figure 2.2a. The diffracted wavefield $\dfrac{\partial u}{\partial m}$ has as a virtual source the term $-(\partial A/\partial m)\, u$, as shown in Figure 2.2b (Pratt et al., 1998). The term $\left(P \dfrac{\partial u}{\partial m} \right)_r$ is the diffracted at one receiver position r. The gradient g is the multiplication of the diffracted wavefield with the residual data $(Pu - d)$ at each receiver position.

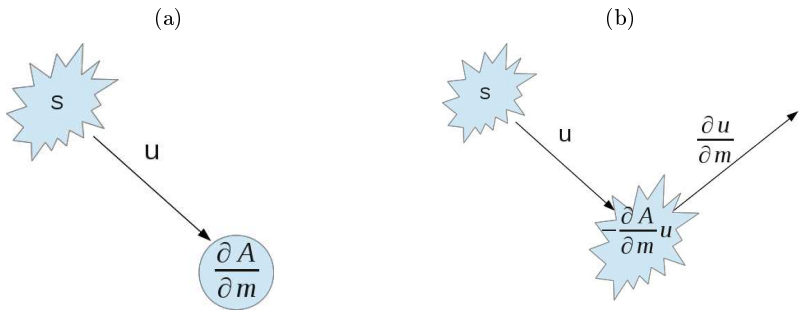


Figure 2.2: a) $Au = s$: Source s generates a wavefield u, that is propagated by the operator A b) $A\frac{\partial u}{\partial m} = -\frac{\partial A}{\partial m}u$: The source $-(\partial A/\partial m)\, u$ generates a diffracted wavefield $\partial u/\partial m$ that is propagated by the operator A.

The i_{th} component of the gradient will be non-zero if the diffracted wavefield by hetereogenity m_i arrives at the receiver position at the same time at as the recorded time for the residual. This is illustrated in Figure 2.3. Consider a true velocity model that consists of an homogeneous background m_0 and three scattering points. There is one source denoted by a star, and a line of receivers at the surface at the positions of the dashed line. The source s generates a wavefield that travels through the model m and that will be diffracted by each of the heterogeneities. In Figure 2.3a the partial derivative wavefield $\partial u/\partial m$ (diffracted by the green scatterer) at the receiver positions is plotted. In Figure 2.3c, the data residuals recorded at the receiver positions are shown. Three arrivals are seen due to the three diffraction points. Recall that the gradient of equation (2.10) is the correlation $\partial u/\partial m$ and the residuals. The sum over time of the correlation the partial derivative wavefields (Figure 2.3a and the data residuals (Figure 2.3 c) is going to generate a non-zero gradient value at point m_i, because we can see that the two arrivals coincide perfectly. A similar illustration of the physical meaning of the gradient can be found in (Pratt et al., 1998).

Consider we wish to reconstruct the true velocity as that shown in Figure 2.4a consisting of an homogeneous background if $1500m/s$, and two circular heterogeneities of $3500m/s$. Eighty four sources are placed around the imaging domain with a spacing of $100m$, and 184 receivers with a spacing of $50m$. A delta function is used as a source, and we do the modeling and the inversion in the frequency domain for one frequency $f = 4Hz$. The initial velocity in Figure 2.4b is the true velocity model after a Gaussian smoothing. The normalized gradient (imaging condition) in Figure 2.4c shows that the velocity should be increased at the position of the hetereogenities (denoted by the yellow dotted lines). At the end of the inversion, the two circular perturbations can be reconstructed. In Figure 2.5 the final velocity models obtained with different optimization methods and the convergence curves as a function of iterations and direct problems are shown. In this case, all of the optimization methods converge towards similar solutions. The converge curves are shown in Figure 2.6a, and the computational cost is plotted in Figure 2.6b. The optimization method $l-$BFGS provides the lowest computational cost. However, we will see below in the example shown in Figure 2.13, that due to the presence of local minima different optimization methods may converge towards different models.

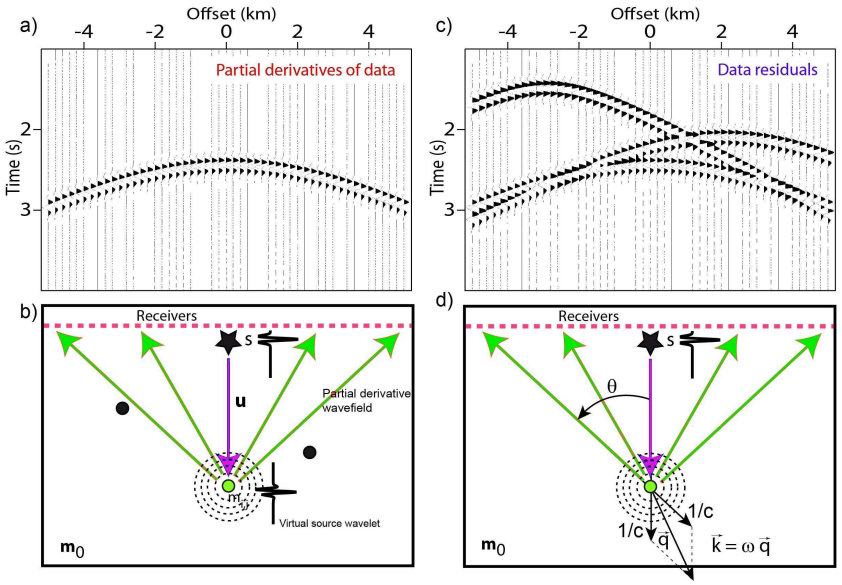


Figure 2.3: Figure taken from (Operto et al., 2013). The true velocity model consists of an homogeneous background m_0 and three scattering points. There is one source denoted by a star, and receivers at the positions of the dashed line. b) The source s generates a wavefield that travels through the model m and is going to be diffracted by each of the heterogeneities. a) The partial derivative wavefield $\partial u/\partial m$ (diffracted by the green scatterer) at the receiver positions. c) The data residuals recorded at the receiver positions. Three arrivals are seen due to the three diffraction points. The sum over time of the correlation the partial derivative wavefields (figure a) and the data residuals (figure c) is going to generate a non-zero gradient value at point m_i. d) The role of the acquisition can be analyzed, considering the angle between the source and the receiver. More details on resolution can be found in Section 1.5 or in the Introduction.

1.3 The Hessian

The computation of the Newton direction of descent (2.8) requires the evaluation of the second derivative of the misfit function known as the Hessian,

$$H = \nabla_m^2 \phi = \left(P \frac{\partial u}{\partial m}\right)^\dagger \left(P \frac{\partial u}{\partial m}\right) + \left(\frac{\partial^2 u}{\partial m^2}\right)^\dagger P^\dagger \left(Pu - d\right). \tag{2.23}$$

Once more, because of the model complexity, there is no analytical solution for the wavefield u. Therefore the wavefield u, diffracted wavefield $\partial u/\partial m$ and the double diffracted wavefield $\partial^2 u/\partial m^2$ must be found numerically. The only term we have not calculated up to now is the double diffracted wavefield. This can be evaluated using by differentiating (53) two times,

$$A \frac{\partial^2 u_s}{\partial m_p m_n} = -\frac{\partial A}{\partial m_n}\frac{\partial u_s}{\partial m_p} - \frac{\partial^2 A}{\partial m_p m_n} u_s - \frac{\partial A}{\partial m_p}\frac{\partial u_s}{\partial m_n}. \tag{2.24}$$

Due to the symmetry of the Hessian, there are only $N(N+1)/2$ different terms that need to be computed. The Hessian computation per source will then require the solution of N direct problems per source to find the diffracted wavefields, plus $N(N+1)/2$ to evaluate (2.24). In addition, one must not forget the additional two direct problems to find u and λ.

Another way is to compute the complete Hessian on the gradient through the adjoint state method. Métivier et al. (2013b) has developed the adjoint state equations to find the Hessian in

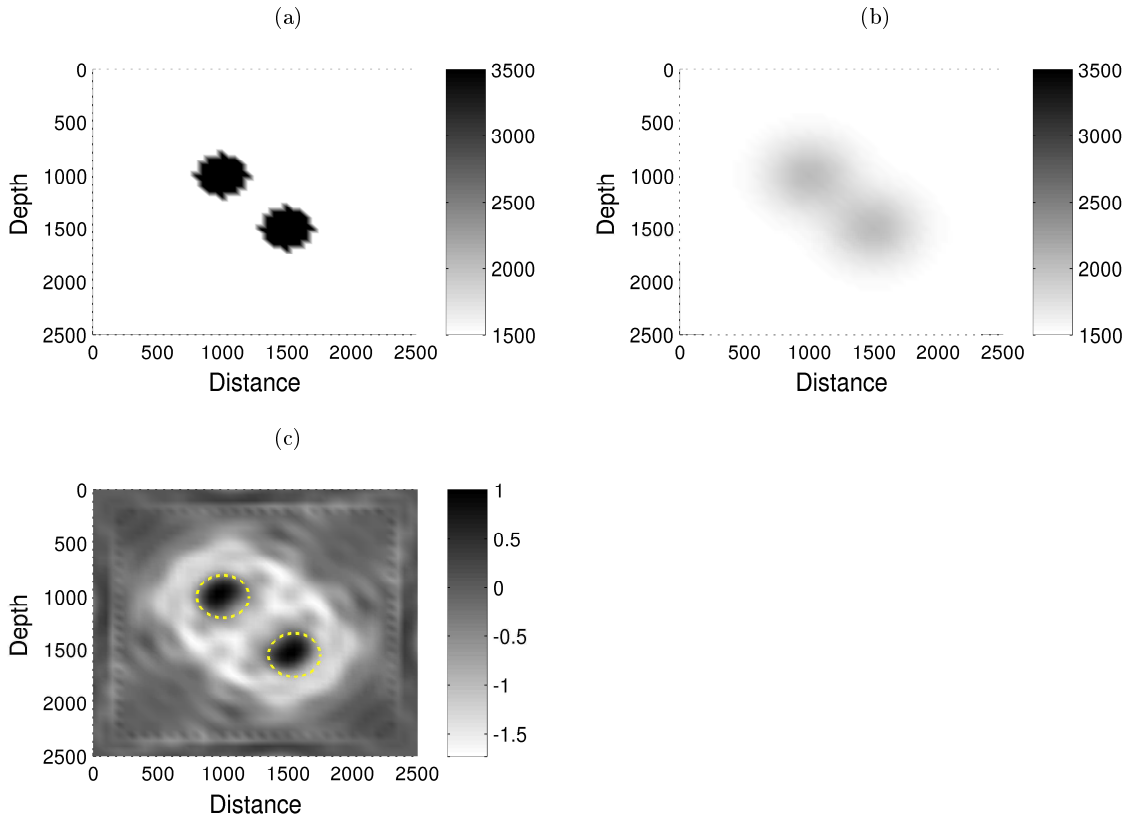


Figure 2.4: 84 sources are placed around the imaging domain with a spacing of $100m$, and 184 receivers with a spacing of $50m$. a) True velocity model V_{true} b) Initial velocity model V_0. c) First gradient, frequency $f = 4hz$.

the context of FWI, and we only outline the main key elements. We shall start by defining the Lagrangian as a real functional,

$$\begin{aligned}\mathcal{L}(u,u^\dagger,\lambda,\lambda^\dagger,\mu_1,\mu_1^\dagger,\mu_2,\mu_2^\dagger,m) &= \frac{1}{2}\left\langle u^\dagger\left(\frac{\partial A}{\partial m}\right)^\dagger\lambda, v\right\rangle + \frac{1}{2}\left\langle v, u^\dagger\left(\frac{\partial A}{\partial m}\right)^\dagger\lambda\right\rangle \\ &+ \frac{1}{2}\left\langle \mu_1, A^\dagger\lambda + P^\dagger(Pu-d)\right\rangle + \frac{1}{2}\left\langle A^\dagger\lambda + P^\dagger(Pu-d), \mu_1\right\rangle \\ &+ \frac{1}{2}\left\langle \mu_2, Au-s\right\rangle + \frac{1}{2}\left\langle Au-s, \mu_2\right\rangle,\end{aligned} \tag{2.25}$$

where μ_1 and μ_2 are adjoint state variables. An arbitrary real vector v is introduced, of the same dimensions as the gradient ($g = u^\dagger\left(\frac{\partial A}{\partial m}\right)^\dagger\lambda$). Therefore, the first and second terms in the Lagrangian represent the inner product of the gradient and v, resulting in a real scalar.

The Lagrangian is chosen this way because when the restrictions for u and λ are satisfied,

$$\begin{aligned}\mathcal{L}(\bar{u},\bar{u}^\dagger,\bar{\lambda},\bar{\lambda}^\dagger,\mu_1,\mu_1^\dagger,\mu_2,\mu_2^\dagger,m) &= \left\langle \bar{u}^\dagger\left(\frac{\partial A}{\partial m}\right)^\dagger\bar{\lambda}, v\right\rangle + \left\langle v, \bar{u}^\dagger\left(\frac{\partial A}{\partial m}\right)^\dagger\bar{\lambda}\right\rangle &(2.26)\\ &= \left\langle 2\Re\left\{\bar{u}^\dagger\left(\frac{\partial A}{\partial m}\right)^\dagger\bar{\lambda}\right\}, v\right\rangle, &(2.27)\end{aligned}$$

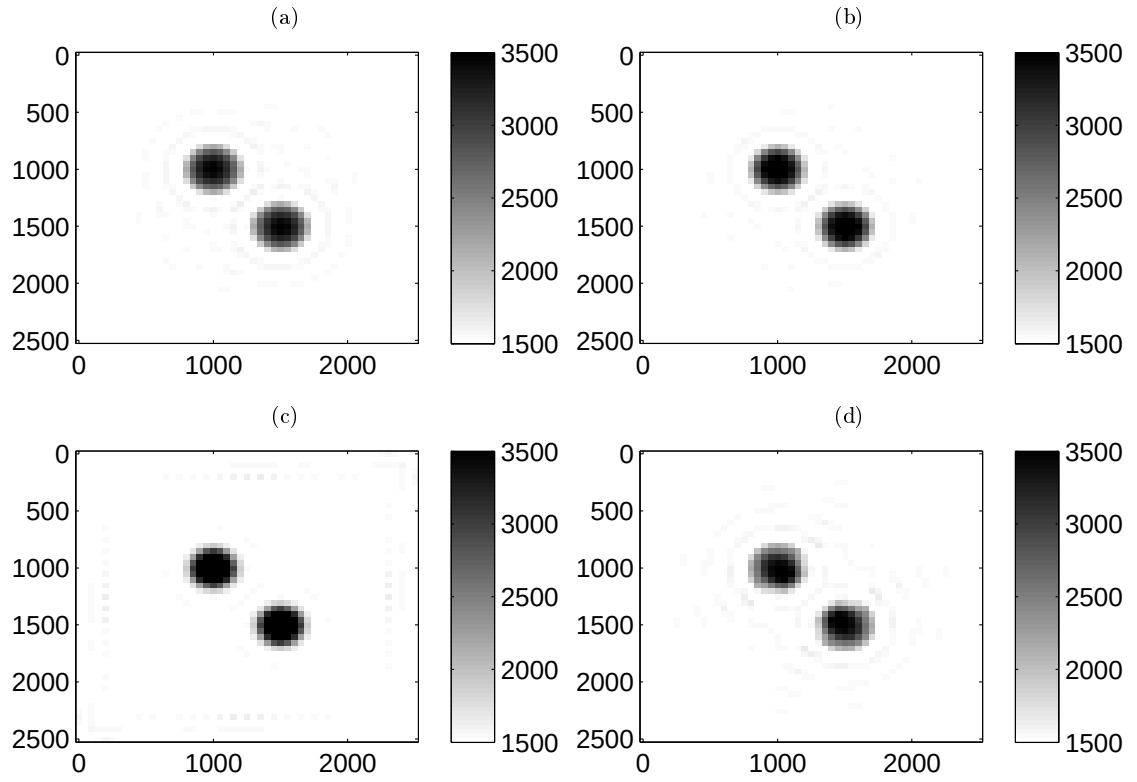


Figure 2.5: 84 sources are placed around the imaging domain with a spacing of $100m$, and 184 receivers with a spacing of $50m$. Frequency $4Hz$. Final velocity model using different optimization methods. a) S.D b) l-BFGS c) G.N d) F.N

and, therefore

$$\nabla_m \mathcal{L}(\bar{u}, \bar{u}^\dagger, \bar{\lambda}, \bar{\lambda}^\dagger, \mu_1, \mu_1^\dagger, \mu_2, \mu_2^\dagger, m) = \nabla_m (g \cdot v) = \langle H, v \rangle \,, \tag{2.28}$$

giving us a way to find H from g. By developing the relations that must be satisfied at the saddle point of $\mathcal{L}$ we find,

$$A\mu_1^j = \frac{\partial A}{\partial m_j} u \tag{2.29}$$

$$A^\dagger \mu_2^j = -P^\dagger P \mu_1^j - \frac{\partial A}{\partial m_j}^\dagger \lambda \tag{2.30}$$

$$H_{ij} = \Re \left\{ u^\dagger \left(\frac{\partial^2 A}{\partial m_i m_j} \right)^\dagger \lambda + \lambda^\dagger \frac{\partial A}{\partial m_i} \mu_1^j + u^\dagger \left(\frac{\partial A}{\partial m_i} \right)^\dagger \mu_2^j \right\}, \tag{2.31}$$

for $i, j = [1, N]$. Equations (2.29) and (2.30) correspond to the forward problems that need to be solved to find each component of μ_1 and μ_2. Finally (2.31) indicates how to constitute an element of the Hessian matrix. Through this adjoint approach, the number of direct problems needed to compute the full Hessian (173) is N for μ_1, N for μ_2. This amounts to a total of $2N$ **direct problems per source**. This represents a lower cost than computing the Hessian with equation 2.24, which was of the order of N^2. However, since N is the number of model model parameters, computing the complete Hessian via the adjoint state method as shown here still represents a great computational cost, specially compared to only 2 forward problems per source needed for the gradient.

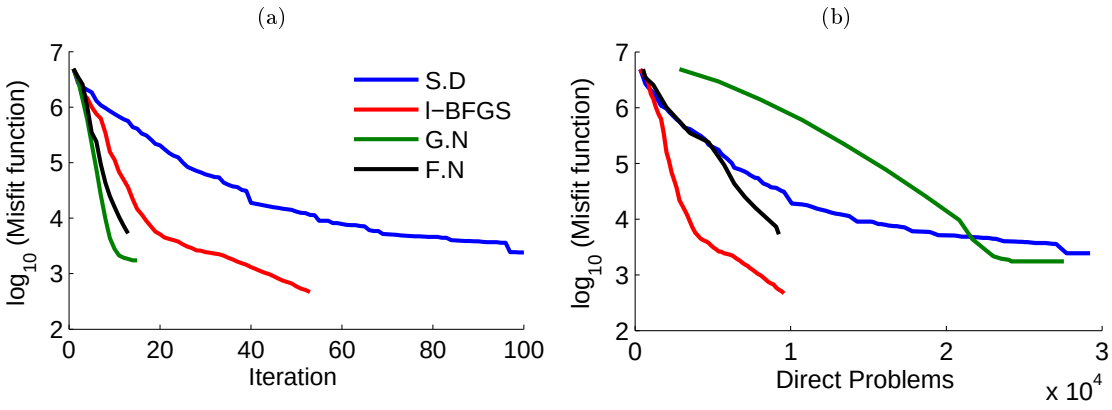


Figure 2.6: Frequency $4Hz$. Convergence curves for the inversions shown in Figures 2.5a-d. a) Reduction of the misfit function versus iterations b) Reduction of the misfit function versus direct problems.

Once the complete Hessian has been computed either with equation (2.31), we still have to go one step further to find the model update $\Delta m_n = -(H)^{-1}\nabla_m \phi_n$. Once we have H, there are basically three options to find Δm_n,

- Compute the inverse of H and

$$\Delta m_n = -(H)^{-1}\nabla_m \phi_n. \tag{2.32}$$

  Because the Hessian is usually an ill-conditioned matrix with many zero eigenvalues, computing its inverse is often discouraged.
- Perform a LU factorization of the Hessian $H = H_L H_U$, and solving the linear system through forward and backward substitution,

$$H_L H_U \Delta m_n = -\nabla_m \phi_n. \tag{2.33}$$

- Let $v = \Delta m_n$. Solve the linear system

$$H_n v = -g_n \tag{2.34}$$

  with a Krylov iterative method. Krylov methods are considered to be efficient for solving sparse linear systems (if an adequate preconditioner is used) (Saad, 2003). The solution is approximated iteratively

$$v_{k+1} = v_k + f(r_k) \text{ for } k = 0, \ldots, L \tag{2.35}$$

  where $f(\cdot)$ is some function that the depends on the specific choice of Krylov, r are the residuals

$$r_k = g - H v_k, \tag{2.36}$$

  and $v_0 = 0$. After L iterations of the iterative solver,

$$\Delta m_n = v_L \tag{2.37}$$

As can be seen, for iterative methods in (2.36), we do not need the full Hessian matrix to be computed and stored, but rather we only require the capability to perform matrix vector

products as Hv. For this purpose, the equation of the Hessian found via the adjoint state method is multiplied by a vector a vector v, which results in

$$A\tilde{\mu}_1 = \frac{\partial A}{\partial m}\tilde{u}v = \left(\sum_{j=1}^{N} v_j \frac{\partial A}{\partial m_j}\right)\tilde{u} \tag{2.38}$$

$$A^\dagger\tilde{\mu}_2 = = -P^\dagger P\tilde{\mu}_1 - \left(\sum_{j=1}^{N} v_j \frac{\partial A}{\partial m_j}^\dagger\right)\tilde{\lambda} \tag{2.39}$$

$$(\tilde{H}v)_i = \tilde{u}\left(\sum_{j=1}^{N}\frac{\partial^2 A^\dagger}{\partial m_i \partial m_j}v_j\right)^\dagger \tilde{\lambda} + \tilde{\lambda}^\dagger\frac{\partial A}{\partial m_i}\tilde{\mu_1} + \tilde{u}^\dagger\left(\frac{\partial A}{\partial m_i}\right)^\dagger\tilde{\mu}_2, \tag{2.40}$$

In this case, the Hessian vector product computation requires the solution of **two direct problems** : one for μ_1 and another one for μ_2, to compute the residual r_k, for each $k = 1, \dots L$. The variable μ_1 represents the diffracted wavefield ($\mu_1 = (\partial u/\partial m)$) where the virtual source is the diffraction pattern times the direct wavefield, weighted by the direction v. Observing equation (2.31), we see that the terms in the Hessian involve cross correlations between $(u, \lambda), (\lambda, \mu_1)$ and (u, μ_2), weighted by a matrix that represents the radiation pattern.

The Gauss-Newton approximation of the Hessian only takes into account first-order diffracted wavefields,

$$H_{GN} = \nabla_m^2\phi = \left(P\frac{\partial u}{\partial m}\right)^\dagger\left(P\frac{\partial u}{\partial m}\right). \tag{2.41}$$

With the adjoint state method, it is equivalent to setting $\lambda = 0$ in equations (2.38)-(2.40). The computational cost of the Gauss-Newton approximation using the adjoint state method is the same as that of the full Newton approximation.

In practice, to find the Hessian vector product, we solve the preconditioned linear system

$$\mathcal{P}H\Delta m = -\mathcal{P}\nabla_m\phi, \tag{2.42}$$

using the preconditioner specified in equation (1.2.a).

1.3.a *Physical interpretation*

The Hessian allows to correct artefacts in the imaging condition. The artefacts in the gradient will come mainly from the limited bandwidth of the source, the smearing artefacts of the imaging condition along an ellipse and double scattered energy that is mapped as single scattered energy. The first term in the Hessian in equation (3.12) represents the zero lag cross correlations of the partial derivative wavefields (diffracted at different points in the model) at the receiver positions. Let $\mathcal{D}$ be a neighborhood containing all the model parameters around m_{ij}. The size of this neighborhood will depend on the source spectrum. For all the points in this neighborhood, all the partial derivative wavefields $\partial u/\partial m$ will arrive at similar times to the receiver positions and thus all will correlate perfectly with the residuals (explained in Figure 2.3). This will cause a defocusing, because all the points around m_i will be imaged, giving rise to rather smooth models. As explained by Pratt et al. (1998), the structure of the GN approximation of the Hessian is similar to a convolutional or smoothing operator. The application of the inverse Hessian sharpens and refocuses the gradient, just as a spiking deconvolution in seismic data processing.

Even in the case of a perfect source with δ as a spectrum, there would still be artefacts in the gradient due to incomplete illumination. For one source-receiver pair, all the points along the ellipse that have the source and receiver as a foci, will be imaged. If there are enough sources and receivers in an adequate acquisition geometry, destructive interference will cancel all points except the target we wish to image. However, with a limited acquisition these artefacts may persist. The Figure 2.7 shows an example done by Pratt et al. (1998), where the gradient has the smile artefacts present. When the inverse GN approximation of the Hessian is applied, these artefacts disappear.

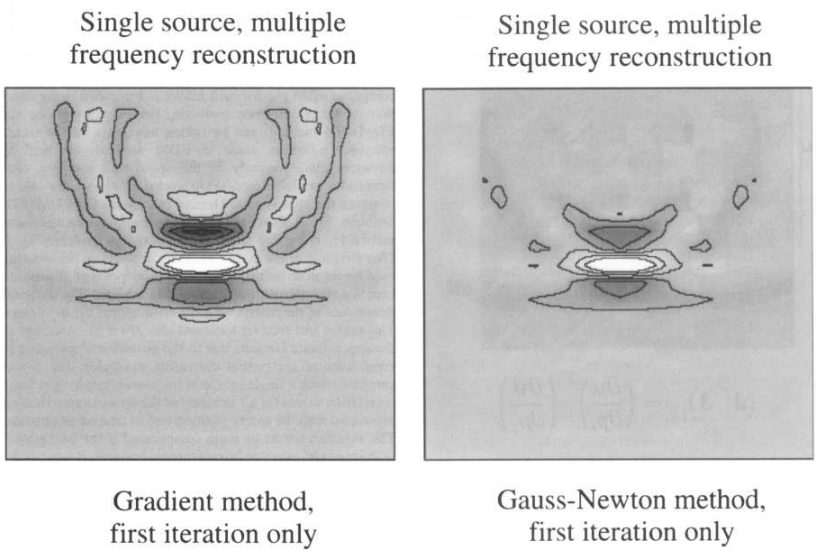


Figure 2.7: Figure from Pratt et al. (1998). With one source and receivers on the surface. On the left, gradient in the first iteration shows the smile artefacts. On the right, the Gauss-Newton descent direction has removed the smile artefacts.

The second term in the Hessian in equation (3.12) is less important if either the residuals are small, or the double diffracted wavefields are small (meaning we are in the linear regime). That is, the second order partial derivative wavefields are important when a change in a model parameter can affect the partial derivative wavefields. To illustrate the importance of the second term consider the example of the velocity model with two inclusions shown in Figure 2.4. We will use the same initial velocity model, but now we will invert for a higher frequency of $f = 6Hz$. In Figure 2.8 the gradient for the first iteration is shown. In Figure 2.8a we show the gradient, in Figure 2.8b the preconditioned gradient and in Figure 2.8c the descent direction given by the inverse of the Full Newton approximation times the Hessian. All three descent directions show that the velocity should be increased at the position of the heterogeneities but, contrary to the case of $4Hz$ there are some artefacts between the two spherical inclusions that arise because of the double diffracted wavefields that appear at this frequency [1]. The preconditioned gradient shows a better imaging condition than the gradient itself, particularly at the source and receiver positions. The descent direction given by the Newton method is slightly better because the artefacts between the two spherical inclusions are weaker. However, we will see that as iterations proceed the effect becomes more pronounced.

Consider the starting model shown in Figure 2.9a (which corresponds to the model given at

1. Diffracted wavefields will be present when the distance between the inclusions will be comparable to the wavelength.

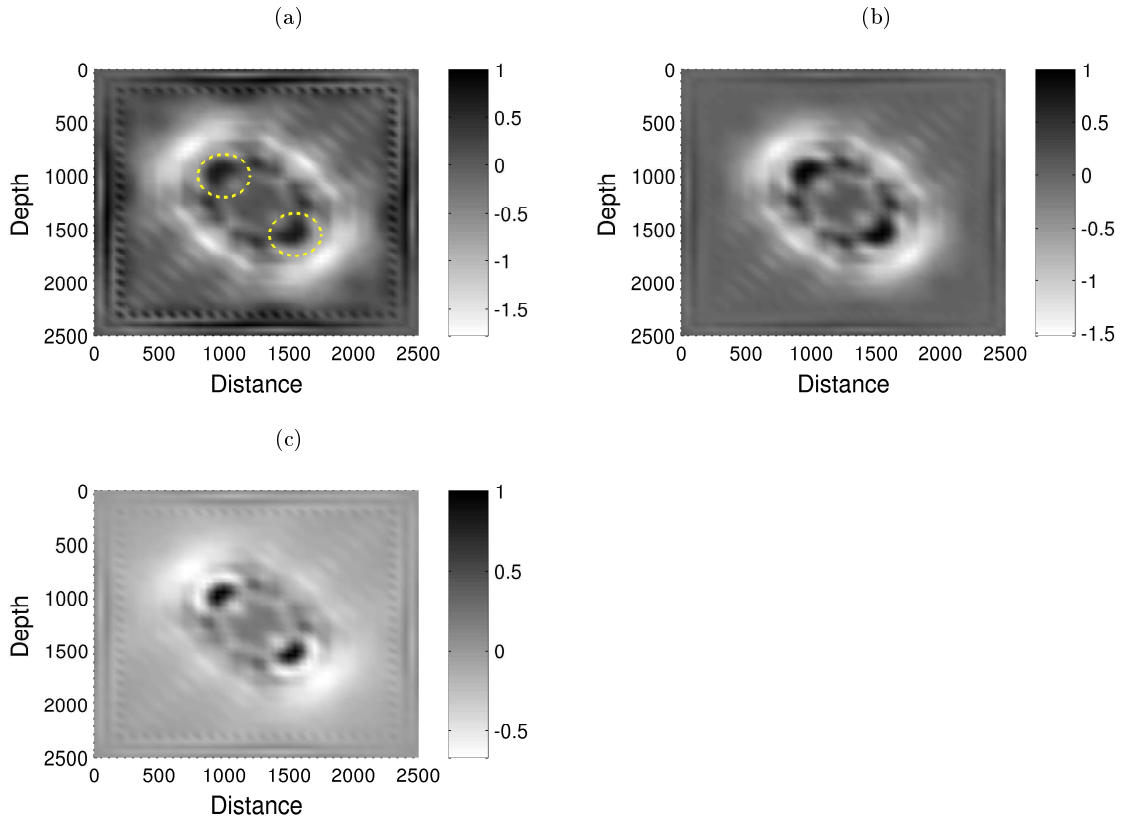


Figure 2.8: Frequency $6Hz$. True velocity model shown in Figure 2.4 a) Gradient b) Preconditioned gradient c) Descent direction with $H^{-1}g$

iteration 6 of the Full Newton inversion). The gradient and preconditioned gradient are shown in Figure 2.9b and Figure 2.9c. The imaging condition now requires that the velocity be increased at points outside the spherical perturbations. That is, the strong double scattered wavefields are being treated as single scattered events at points symmetrically opposite to the true velocity perturbations. The Newton descent direction corrects these artefacts in the imaging condition, as shown in Figure 2.10 a.

For this small test case case, we have computed the full Hessian matrix thus allowing the comparison of different methods to solve the Newton equations. Generally, one would not compute the full Hessian and equation a linear conjugate gradient would be employed to solve the linear system, as in (2.34). The resulting descent direction is shown in Figure 2.10a. Alternatively, we can find directly the inverse of the full Hessian and apply it to the gradient, as explained in 2.32. However, the Hessian may have singluar values (specially in the absence of complete illumination) and therefore finding directly the inverse is not generally possible. To find the inverse we thus find a the SVD decomposition of the Hessian, $H = USV'$. The inverse will be $H^{-1} = V(1/S)U'$, where U and V are the matrices of eigenvectors and S is a diagonal matrix of eigenvalues. We put to zero all the entries in $1/S$ where the eigenvalues are below a threshold (10^{-8} in this case). The structure of the Hessian, its inverse and the eigenvalue distribution are shown in Figure 2.11. The resulting descent direction is shown in 2.10 b. The two descent directions in Figure 2.10 are not the same, but both have corrected the artefacts present in the imaging condition.

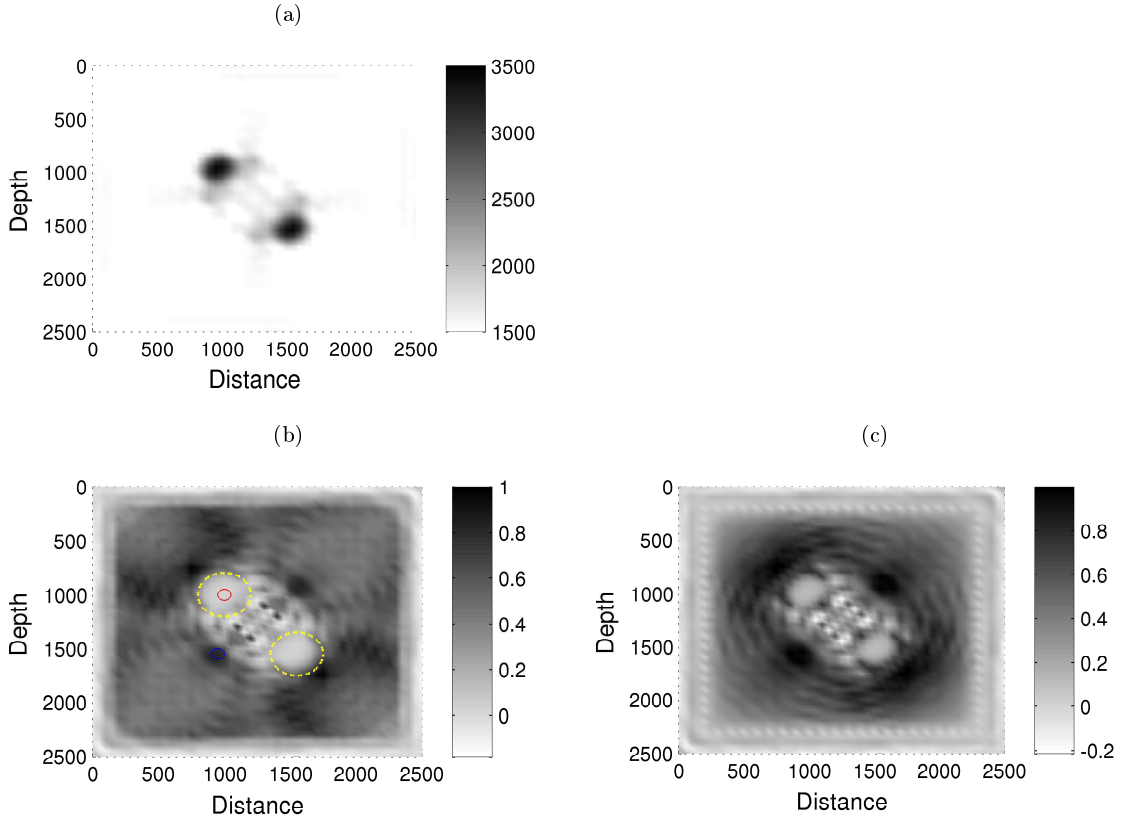


Figure 2.9: Frequency $6Hz$. a)Initial velocity model (at step n of the optimization) b) Gradient c) Preconditioned gradient

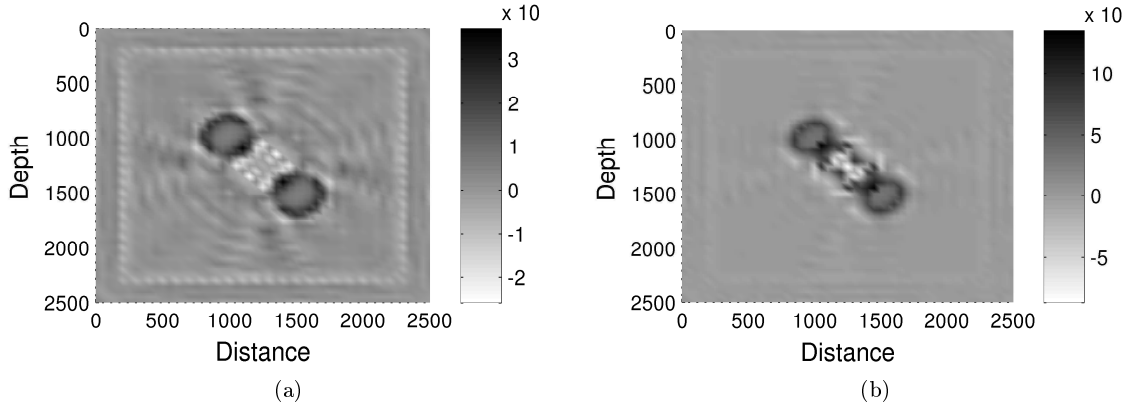


Figure 2.10: Frequency $6Hz$. a) Solution of the linear system $H\Delta m = -g$ with conjugate gradient. Seven iterations of conjugate gradient are performed before quitting. b) Inverse of the Hessian times the gradient. The inverse of the Hessian is found using an SVD and thresholding small eigenvalues. c)

To better understand the action of the inverse Hessian on the gradient let us consider two points in the model, circled in red and blue in Figure 2.9b. The point in the model circled in blue corresponds to the position where an artefact is created with the gradient method, and the point in the model circled in red is a point inside the circular heterogeneity. The first column in Figure 2.12 shows, for the model parameter circled in blue, one row of the Hessian, one row of the inverse Hessian, and the multiplication of the inverse Hessian times the gradient. The second column in Figure 2.12 shows, for the model parameter circled in red, one row of the Hessian, one row of the inverse Hessian, and the multiplication of the inverse Hessian times the gradient. Both columns are shown on the same scale to make a comparison between both diffracted wavefields. Figures 2.12 a) and 2.12b show one row of the Hessian plotted in the model dimensions. High values of the partial derivatives in the Hessian indicate strong correlations of the diffracted wavefields. High partial derivatives could also indicate strong correlations between the double-diffracted wavefields and the residuals (second-order term in the Hessian). Comparing 2.12a and 2.12b we see that at the point $(i,j) = (19,31)$ the correlation of the double diffracted wavefields with the residuals are much stronger than at the point $(i,j) = (20,20)$. One row of the inverse Hessian is shown in 2.12c and 2.12 d. The inverse Hessian row corresponding to the point $(i,j) = (20,20)$ in Figure 2.12c has a larger amplitude than the inverse Hessian row corresponding to the point $(i,j) = (19,31)$ in Figure 2.12d. Finally, in Figures 2.12e and 2.12f one row of the inverse Hessian is multiplied by the gradient, for each element. The sum of this product will be a scalar that determines the value of the descent direction for the position i,j. For the two points in the model considered we obtain, $H^{-1}_{19,31}\, g = 0.1 \cdot 10^6$ and $H^{-1}_{20,20}\, g = 3.56 \cdot 10^6$. Therefore, the amplitude of artefact at position $(19,31)$ will be decreased and the amplitude of the model parameter at $(20,20)$ will be increased.

The final velocity models when inverting for $f = 6Hz$ using four different optimization algorithms with the corresponding convergence curves are shown in Figures 2.13. The strong double diffractions create artefacts in the final velocity models using S.D, l-BFGS and G.N. The full Newton Hessian approximation mitigates these artefacts and correctly images the two inclusions. If we add a Tikhonov l_2 norm regularization term, the performance of the inversion can be highly improved, as shown in 2.14. The regularization terms helps to find a better descent direction and all the final velocity models have superior quality. However, Gauss Newton has the poorest performance in this case. If we push our exercise further and invert for a higher frequency of $f = 7Hz$, the results are not satisfactory, as can be appreciated in Figure 2.15. The final velocity models are all inaccurate, probably due to cycle skipping, although the full Newton method provides the best image.

1.4 Solving FWI numerically in the frequency or time domain

Full waveform inversion can be done using the acoustic or elastic wave equation to solve the forward problem, in one, two or three dimensions, $D = 1,2,3$. FWI was originally formulated in the time domain (Lailly, 1984; Tarantola, 1986; Mora, 1987) and a posteriori the frequency domain inversion was introduced (Pratt et al., 1998). The time or frequency FWI are equivalent when all the frequencies are taken into account. In two dimensions, the frequency domain approach is more efficient numerically both in the direct problem and in the inverse problem.

- Spatial discretization
  The discretization of the spatial domain Ω can be done through finite difference, finite elements or finite volumes. We focus on the use of finite differences. The simplest discretization

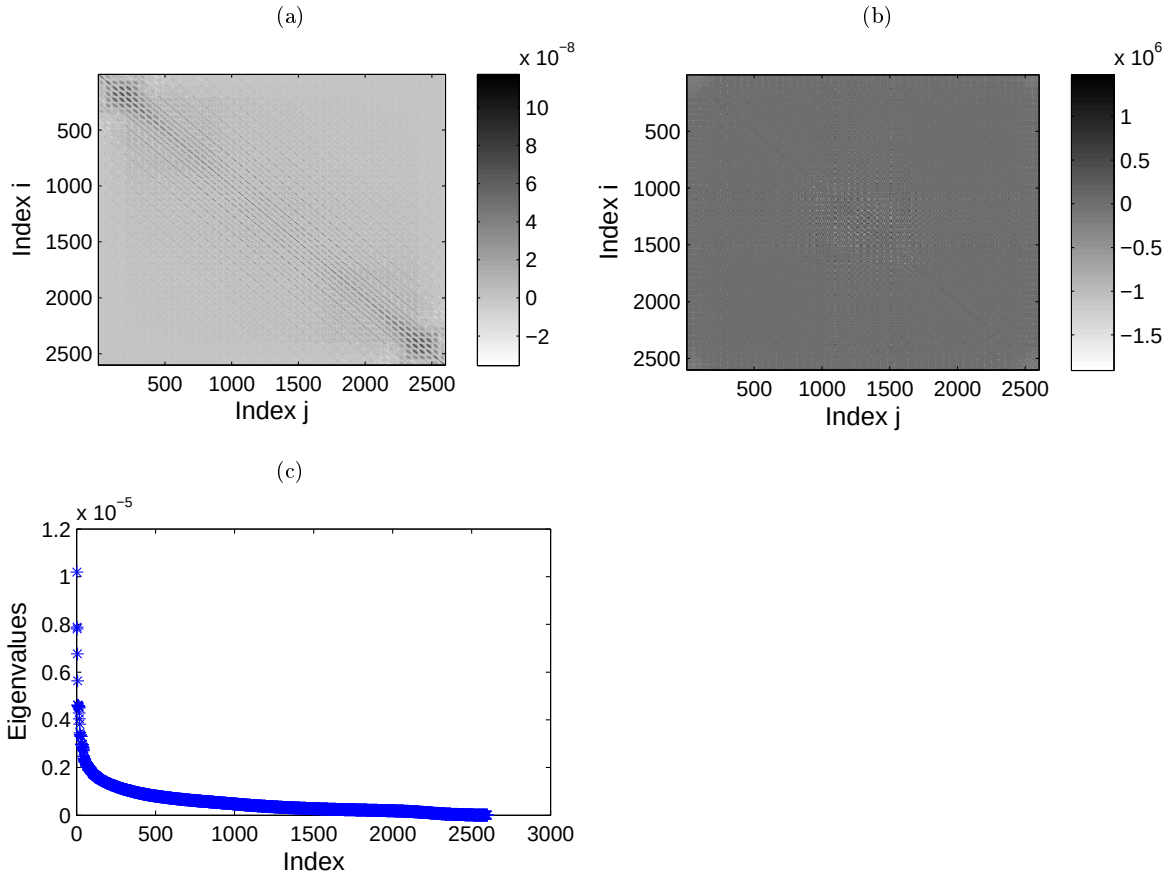


Figure 2.11: Frequency $6Hz$. a) Hessian b) Inverse Hessian c) Eigenvalues of the Hessian matrix.

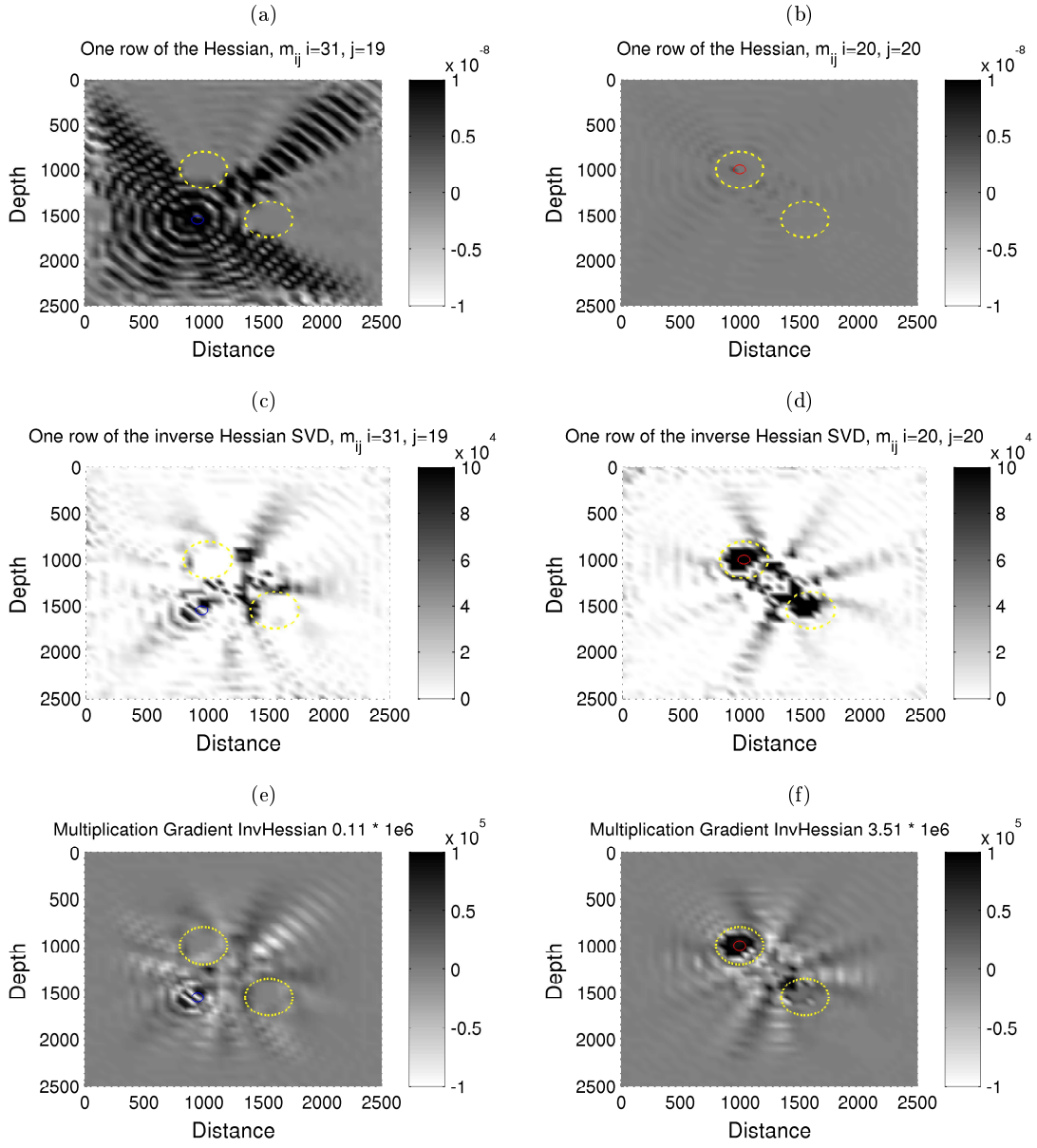


Figure 2.12: Frequency $6Hz$. Effects of the Hessian on two points of the model, circled in blue and red in Figure 2.9b. Figure a) and b) show one row of the Hessian plotted in the model dimensions for the model parameter circled in blue and red, respectively. High amplitudes mean strong correlations between diffracted wavefields and/or strong correlations of the double diffracted wavefields with the residuals. Comparing a) and b) we see that at the point $(i,j) = (19,31)$ the correlation of the double diffracted wavefields with the residuals are much stronger than at the point $(i,j) = (20,20)$. One row of the inverse Hessian is shown in c) and d). The inverse Hessian row corresponding to the point $(i,j) = (20,20)$ in c) has a larger amplitude than the inverse Hessian row corresponding to the point $(i,j) = (19,31)$ in d). Finally, in Figures 2.12 e) and 2.12 f) one row of the inverse Hessian is multiplied by the gradient, for each element. The sum of this product will be a scalar that determines the value of the descent direction for the position i,j. For the two points in the model considered we obtain, $H^{-1}_{19,31}\ g = 0.1 \cdot 10^6$ and $H^{-1}_{20,20}\ g = 3.56 \cdot 10^6$. Therefore, the amplitude of artefact at position $(19,31)$ will be decreased and the amplitude of the model parameter at $(20,20)$ will be increased.

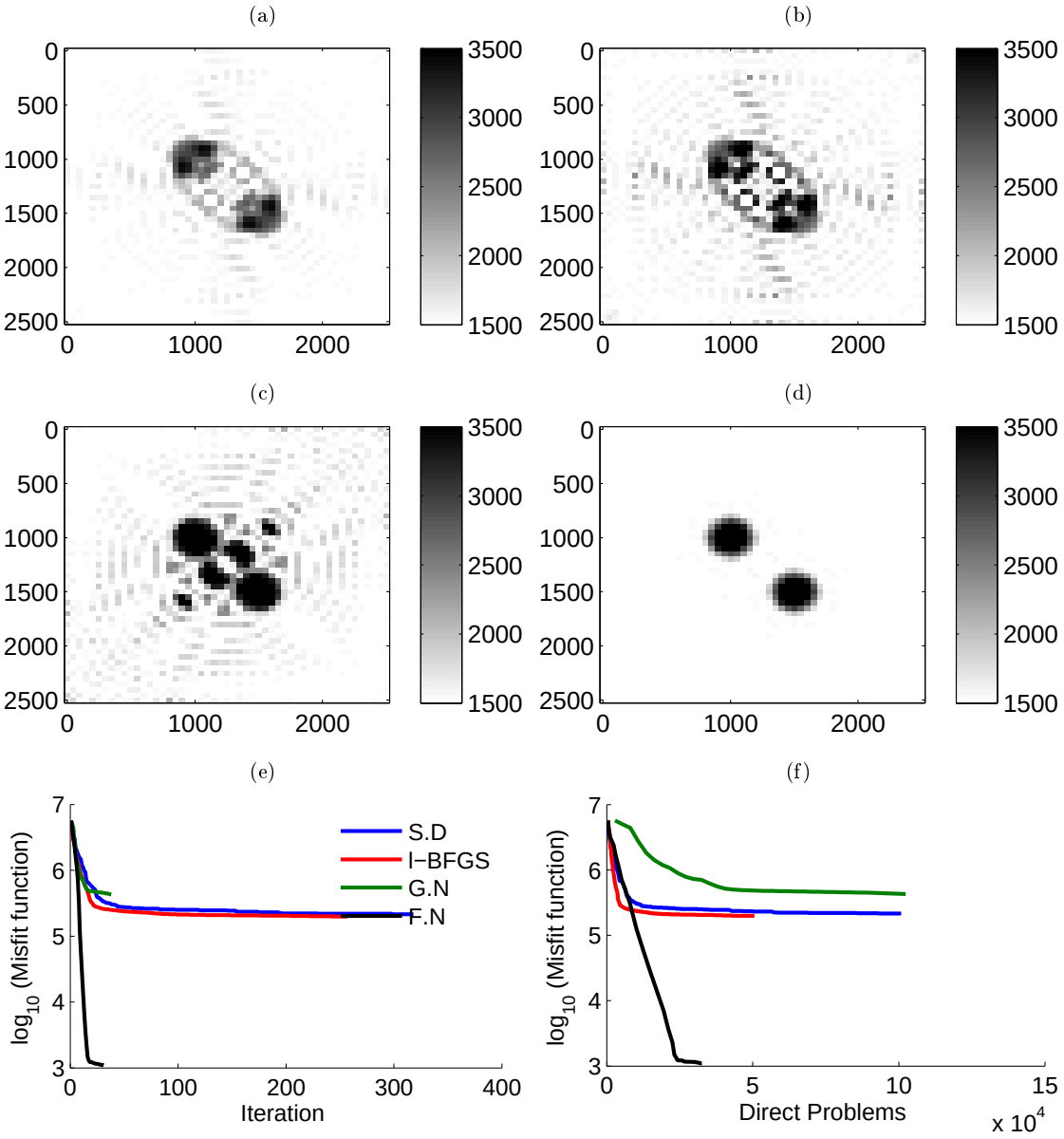


Figure 2.13: Frequency $6Hz$. Final velocity model for different optimization algorithms a) S. D b) l-BFGS. c) G. N d) F.N d)Reduction of the misfit function versus iterations e) Reduction of the misfit function versus direct problems. The strong double diffractions create artefacts in the final velocity models using S.D, l-BFGS and G.N. The full Newton Hessian approximation corrects these artefacts and correctly images the two inclusions. The true velocity model and the initial velocity model are shown in Figure 2.4

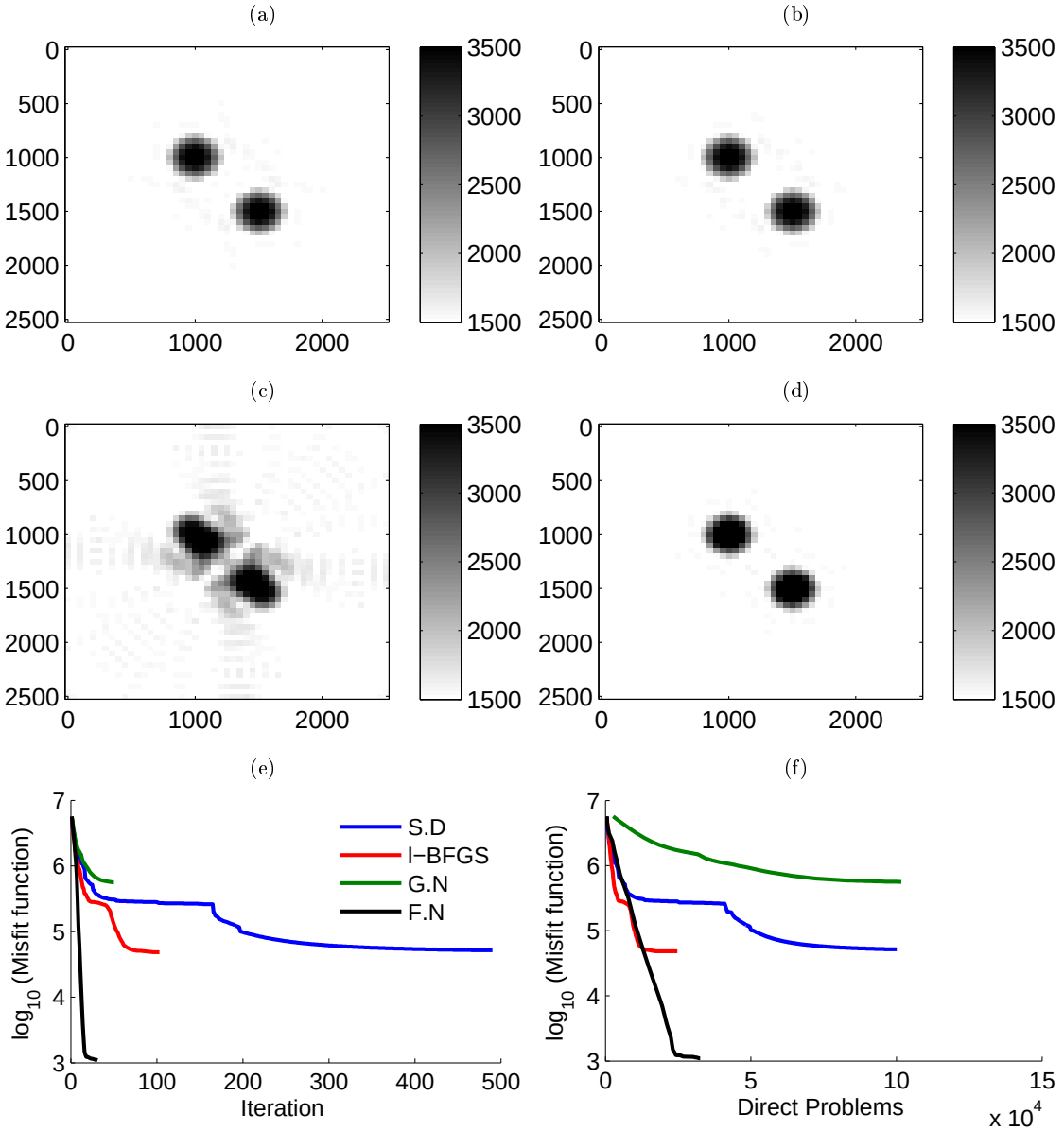


Figure 2.14: Frequency $6Hz$, with regularization. Final velocity models with a) S.D, $\lambda = 1e^{-2}$ b) l-BFGS, $\lambda = 1e^{-2}$ c) G.N, $\lambda = 1e^{-4}$, $l - CG = 15$ d) F.N , $\lambda = 0$, $l - CG = 15$ e) Reduction of the misfit function versus iterations f) Reduction of the misfit function versus direct problems.

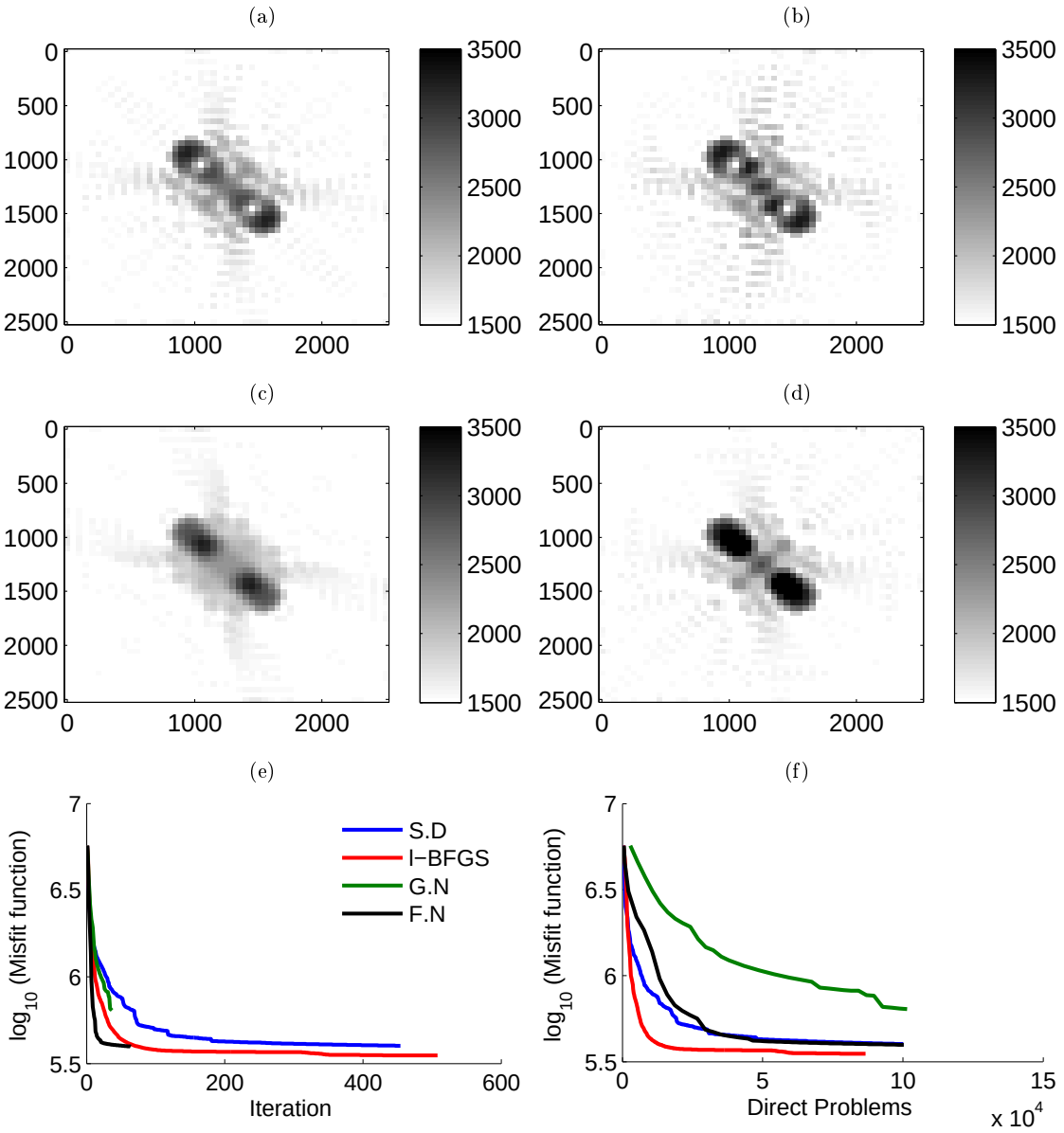


Figure 2.15: Frequency $7Hz$, with regularization. Final velocity models with a) S.D $\lambda = 1e^{-3}$ b) l-BFGS $\lambda = 1e^{-3}$ c) G.N $\lambda = 1e^{-3}$ d) F.N $\lambda = 1e^{-3}$ e)Reduction of the misfit function versus iterations f) Reduction of the misfit function versus direct problems.

of the Laplacian consits in using a second order centered scheme is,

$$\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2} \approx \frac{u(x-\Delta x,y)-2u(x,y)+u(x+\Delta x,y)}{\Delta x^2} + \frac{u(x,y-\Delta y)-2u(x,y)+u(x,y+\Delta y)}{\Delta y^2}+o(\Delta x^3,\Delta y^3). \tag{2.43}$$

If the 2D domain contains $N = N_x \times N_y$ points, the discretization of the second derivative (2.3) can be written in matrix form

$$\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2} = Au(x_i,y_j) \tag{2.44}$$

where A is an $N \times N$ matrix , and we have written $u(x_i, y_j)$ to denote that the wavefield has been discretized. The wave field u will be a vector of dimensions $N \times 1$. The matrix A consists mainly of three bands (because we are in two dimensions, the first band corresponds to neighbors above a point, the central band correspond to neighbors on the left and right, and the last band corresponds to neighbors below). The width of the bands depends on the discretization order of the spatial derivatives. As the order increases, so does the width of the bands. Other discretizations are possible. For example, if instead of using the second order wave equation, a system of two first order velocity-stress coupled partial differential equation, it is possible to implement staggered grids (Virieux, 1986).

– **Time Domain forward problem**

In the time domain, the second order time derivative must also be discretized, using either implicit and explicit discretization schemes. Implicit time schemes require the solution of linear systems, but have better stability properties. Using a second order explicit centered scheme to discretize the second order time derivative gives,

$$\frac{\partial^2 u(x,t)}{\partial u(x,t)^2} \approx \frac{u(x,t+\Delta t)-2u(x,t)+u(x,t-\Delta t)}{\Delta t^2}. \tag{2.45}$$

Therefore, solving the forward problem (2.3) amounts to solving

$$\boxed{u(x,z,t+\Delta t)=u(x,z,t-\Delta t)+2u(x,z,t)+\Delta t^2 Au(x,z,t)+\frac{\Delta t^2}{\Delta x^2}s(x,z,t).} \tag{2.46}$$

Alternatively, the second order wave equation can be written as two coupled first order stress-velocity equations, which requires only the information of one previous time step.

☛ Each source function $s(x, z, t)$ requires the solution of system (2.3) with a time marching algorithm.

– **Frequency domain forward problem**

To solve the problem in the frequency domain, we apply a Fourier transform to (2.3),

$$\left(\nabla^2+\frac{\omega^2}{v^2}\right)u(x,z)=s(x,z,\omega). \tag{2.47}$$

The spatial discretization of the operator $A = \left(\nabla^2+\frac{\omega^2}{v^2}\right)$ can be written in a matrix form of $N \times N$ as was done in (2.44). Since there are no time derivatives, this reduces the solution of the forward problem to

$$\boxed{A(x,z,\omega)u(x,z,\omega)=s(x,z,\omega).} \tag{2.48}$$

The solution $u(x, z, \omega)$ can be found by solving the linear system either by approximating the inverse of A (for example via an SVD decomposition). However, A is a large ill conditioned matrix and finding the inverse via an SVD decomposition is infeasible. Therefore, only iterative methods or direct matrix decomposition methods based on Gauss eliminations are used.

☛ An example of a direct method is the LU factorization of the A matrix. Once the factorization is performed, $u(x, z, \omega)$ can be found through forward and backward substitutions. A new LU factorization must be done for each frequency ω, but it is valid for all sources.

– **Time versus frequency domain in the forward problem**

In the time domain, solving the forward problem the number requires solving the wave equation (2.46) N_s times, where N_s is the number of sources. To solve the direct problem in 2D the frequency domain, the number of times the wave equation (2.48) must be solved is also equal to the number of sources, but for each frequency the same LU factorization is valid for all sources. The greatest effort is in performing the LU factorization, because afterwards it only requires simple forward-backward substitutions. Since generally the number of frequencies that are taken into account in the inversion is considerably less than the number of sources, solving the 2D forward problem in the frequency domain is faster (Pratt and Worthington, 1990; Pratt, 1990).

In 3D the situation in analogous ($N = N_x \times N_y \times N_z$), but the difficulty arises in terms of memory limitations to store the L and U matrices. There are some techniques, such as the nested dissection, that limit the filling of the L and U matrices, and can reduce the memory requirements by one order (George and Liu, 1981). In acoustic 3D approximations, the size of the matrices may still accept the use of LU solvers (Operto et al., 2007) but in general elastic applications, iterative solvers must be used. (Plessix, 2009). A summary of the approximate memory requirements and computational costs are synthesized in Table 2.1.

– **Time versus frequency domain in the inverse problem**

To solve the inverse or optimization problem, we need to make use of the gradient and eventually also the Hessian. In the time domain, the computation of the gradient for each source is given in equation (2.15). This would require storing in memory the wavefields $u(x, t)$ and $\lambda(x, t)$ on all the domain, for all time steps, for each source at a time. The number of time steps can be arbitrarily large and this makes the computation of the gradient in this way practically infeasible. What is usually done is that as the back-propagated wavefield is computed for a time $\lambda(x, t_i)$, the direct wavefield $u(x, t_i)$ is *recomputed* for each t_i and the correlation for t_i is performed, and added to the gradient found in the previous time steps. To make the re-computation of the direct wavefield more efficient, the value of the direct wavefield is stored in memory for some specific set of time values $\{t_{c_1}, t_{c_2}, t_{c_N}\}$. The re-computation of the direct wavefield is then done from the closest previous stored value and used as an initial condition (Symes, 2007). Another way to recompute the wavefield $u(x, t)$ in the absence of attenuation is to store the wavefield u at all times, but only on the boundaries. From the values on the boundaries, by time-reversal the wavefield can be computed everywhere on the domain. The computation of the Hessian vector products in the time domain also suffer from the same problems and thus the direct wavefields must be again recomputed to find (2.38).

Table 2.1: Approximate memory and computational complexity in the *forward problem* for iterative solvers and the LU factorization, for a matrix A of dimensions $N \times N$. In 2D $N = N_z \times N_x$ and in 3D $N = N_z \times N_y \times N_x$.

	Iterative Solvers	
	2D (for N sources)	3D (for N^2 sources)
Memory complexity	$O(N^2)$	$O(N^3)$
Computational complexity	$O(N^3)$	$O(N^6)$
	LU Direct Solver	
	2D	3D
Memory complexity	$O(N^2 log_2 N)$	$O(N^4)$
Computational complexity	$O(N^3)$	$O(N^6)$
	Time Marching	
	2D	3D
Memory complexity	$O(N^2)$	$O(N^3)$
Computational complexity	$O(N^4)$	$O(N^6)$

Table 2.2: Approximate memory requirements and computational complexities in the forward problem for frequency and time algorithms (Nihei and Li, 2007; Operto et al., 2007; Plessix, 2007).

On the other hand, the computation of the gradient in the frequency domain with equation (2.17), requires storing the solution of the wavefield on all the domain Ω for all the number of frequencies N_f. However, the inversion in the frequency domain is usually performed solving sequential optimization problems, where each optimization only uses a few limited number of frequencies (Sirgue and Pratt, 2004). It is therefore feasible to store the direct and back-propagated wavefields on all the domain, for a limited number of frequencies, and no recomputations need to be done.

1.5 Image resolution analysis

The resolution analysis is directly related to the imaging condition and through the understanding of the concept of Fresnel zone.

Fresnel zone

In the schematic illustration of the gradient for one pair source-receiver in Figure 2.1, the ellipse has no width (as with ray theory). For waves with finite frequency, the ellipse will have a width that will provide some insight on the spatial resolution of the image. In the frequency domain, the imaging kernel will show the ellipses that satisfy the imaging condition for all the arrivals in time. An example of an imaging kernel for one frequency (sometimes known as sensitivity kernel) is represented in Figure 2.16. The dark and light fringes denote the sensitivity of the model parameters to one frequency. The central ellipse defines the first Fresnel zone. The first *Fresnel zone* is the region around the direct path such that the travel time for points in this region differs in no more than $T/2$. All arrivals of scattered paths within half a period of one arrival will generate constructive interference in the seismograms. In terms of wavelength, the Fresnel zone is the region around the reflector path such that the ray path differs by no more than $\lambda/2$. The first Fresnel zone defined for each offset D is important for two reasons. If the background model is not sufficiently accurate such that the difference in between the arrival of scattered modeled data and the recorded data does not exceeds more than half a period, the

resulting data residual will be back-propagated on the wrong Fresnel zone leading to erroneous update of the subsurface model. This is referred to as cycle-skipping. The width of the first Fresnel zone is $\sqrt{\lambda D}$ (Williamson, 1990).

The second reason why the Fresnel zone is important is because it limits the resolution of the final image. Two point scatterers that generate reflected waves whose arrival times is less than half a period, will not be distinguishable. For a surface acquisition, the smallest distance dx and dz between two point scatters for them to have an arrival time difference of at least $T/2$ is approximately [2] (Schuster, 2007),

$$dx \approx \lambda\sqrt{D^2+z^2}/(4\cdot D) \tag{2.49}$$

$$dz \approx \lambda/4, \tag{2.50}$$

where z is the depth of the scatterer and D is the offset. From these equations we can see that the vertical resolution dz of the image does not depend on the offset, but only on the smallest wavelength. The horizontal resolution dz is different for shallow or deep scatterers. For example, for deep scatterers ($z >> D$), long offsets D are needed to improve the horizontal resolution.

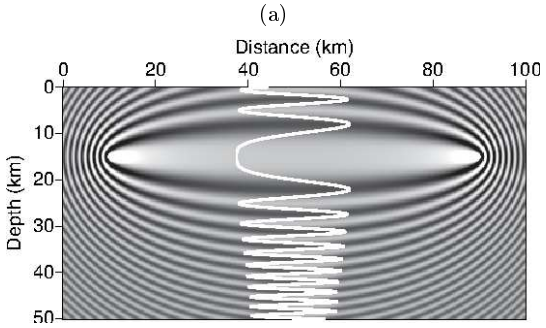


Figure 2.16: Imaging kernel (21) in the frequency domain for one source receive pair. The first central ellipse denotes the first Fresnel zone. The black and white fringes describe the iso-phase surfaces of *all* the scattered arrivals. (Each isochrone represents all the parameters in the model space that can explain a seismic phase arrival at time t). The width of the isochrones decreases in depth as the scattering angle decreases (Woodward, 1992).

Wavenumber reconstruction of the image

Another aspect of the quality of the image concerns the capability to reconstruct small and large details of the model. By applying a Fourier transform to the depth and horizontal distance variables (z, x), and working in the wavenumber domain (k_z, k_x) it is possible to analyze the resolution of the model in terms of coverage of the wavenumber spectrum (Devaney, 1984; Beylkin, 1987; Miller et al., 1987; Wu and Toksöz, 1987; Devaney and Zhang, 1991; Schuster, 1996; Chen and Schuster, 1999; Xu et al., 2001; Lambaré et al., 2003; Sirgue, 2006). To describe the transmission components over a wide range of offsets the low-wavenumber components of the velocity model must be accurate. To account for reflection amplitudes, the high wavenumber components of the velocity model must be correct (Mora, 1989).

A short summary of the analysis done in Sirgue and Pratt (2004), under far field assumptions and using the Born approximation, is presented below. Considering an inversion in the frequency

2. This resolution analysis is for poststack migration (Schuster, 2007).

domain, and using a delta a source function, the direct $u(x,\omega)$ and back-propagated wavefields $\lambda(x,\omega)$ in terms of the Green's functions are

$$u(x,\omega) = G(x,s) \tag{2.51}$$
$$\lambda(x,\omega) = G^{\dagger}(x,r)\left(Pu-d\right). \tag{2.52}$$

Plugging this back in the gradient expression (2.17),

$$\nabla_m \phi = \omega^2 \mathcal{R}e\left\{G^{\dagger}(x,s)G^{\dagger}(x,r)\left(Pu-d\right)\right\} \tag{2.53}$$

Under a plane wave approximation,

$$\nabla_m \phi(x) = \omega^2 \sum_{s=1}^{N_s}\sum_{r=1}^{N_r} \mathcal{R}e\left\{exp(-i\vec{k}\cdot(\hat{s}+\hat{r}))\left(Pu-d\right)\right\}, \tag{2.54}$$

where the explicit sum over the number of sources N_s and the number of receivers N_r has been included. The unit vector $\hat{s}$ points from the source to the diffracting point x and $\hat{r}$ is a unit vector from the receiver to the diffracting point. With (2.54) the gradient can be interpreted as an inverse Fourier where the coefficients are given by the residuals. Let us denote $\vec{k}_{S,x} = \vec{k}\cdot\hat{s}$ as the vector from the source to point x, and $\vec{k}_{R,x} = \vec{k}\cdot\hat{r}$ as the vector from the receiver to the point x. In Figure 2.17, the wave vectors are plotted, in the wave vector space. The angle between $\vec{k}_{S,x}$ and $\vec{k}_{R,x}$ at the diffracting point x is θ. The resulting wave vector at x is

$$\vec{k} = (k_x,k_z) = \frac{\omega}{\vec{v}} = \frac{2\pi f}{|v|} cos(\theta/2)\hat{n} \tag{2.55}$$
$$\hat{n} = (\vec{k}_{S,x}+\vec{k}_{R,x})/\sqrt{\|\vec{k}_{S,x}+\vec{k}_{R,x}\|^2}, \tag{2.56}$$

where $1/\vec{v}$ is known as the slowness vector.

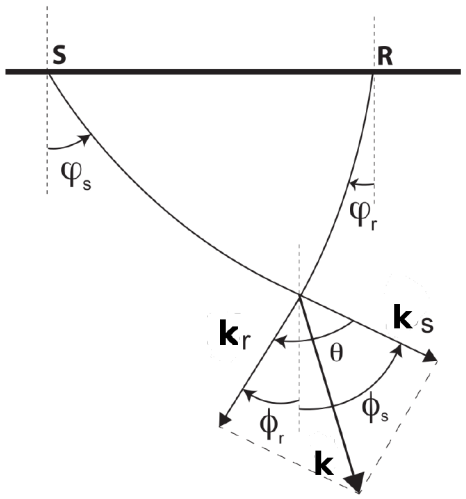


Figure 2.17: Figure adapted from Thierry et al. (1999); Sirgue and Pratt (2004). For a source receiver pair, the $k(\vec{x}) = \omega/v(\vec{x}) = \dfrac{\omega}{v}\cos(\theta/2)\hat{n}$. Where $\vec{k}_{S,x}$ and $\vec{k}_{R,x}$ are vectors from the source to the diffracting point and from the receiver to the diffracting point. $\hat{n} = \dfrac{\vec{k}_{S,x}+\vec{k}_{R,x}}{\sqrt{\|\vec{k}_{S,x}+\vec{k}_{R,x}\|^2}}$

In order to have a gradient that can reconstruct all wave numbers both in amplitudes and directions, different angles and different frequencies must be considered. The term f and the magnitude of v play a role in the magnitude of the reconstructed wave number, and the angle θ plays a role in the direction and magnitude. In Figure 2.18 we can see the effect of changing the angle θ on the wave number reconstruction, for a fixed frequency. The angle is increased from Figure 2.18a to Figure 2.18c, and as a consequence the reconstructed wave number decreases. In a frequency inversion, the frequencies must be chosen in order to have a full coverage of all the wave numbers, and ideally no gaps must be left. Sirgue and Pratt (2004) proposed a strategy to efficiently choose the frequencies to attain full wave number coverage, shown in Figure 2.21b. The idea consists in computing, for a certain frequency f_n, the range of wave numbers $[k^n_{min}, k^n_{max}]$ that can be reconstructed. The next frequency f_{n+1} will have a wave number range $[k^{n+1}_{min}, k^{n+1}_{max}]$ such that the lower bound k^{n+1}_{min} at least coincides with the upper bound of the previous frequency k^n_{max}.

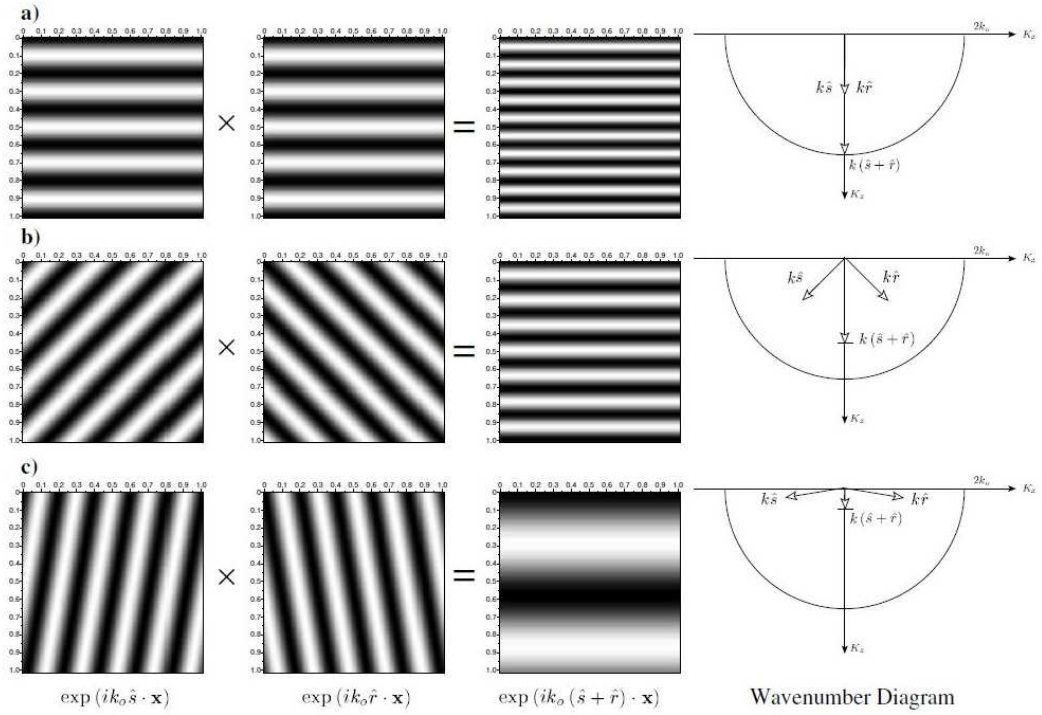


Figure 2.18: Figure from Sirgue (2003). Under the plane wave approximation, an illustration of the variation of the wave number contribution in the gradient, for different incidence-scattering angles. The first column represents the wave paths from the source to the scattering point $\exp(i\vec{k}\cdot\hat{s})$. The second column represents the wave paths from the receiver to the scattering point $\exp(i\vec{k}\cdot\hat{r})$. The last column is the multiplication of both paths $\exp(i\vec{k}\cdot(\hat{r}+\hat{s}))$, which is proportional to the gradient. a) The source are receiver are at the same position. The incidence and scattering angle is $0°$. The total wave number is high, which can be seen in the high resolution of the gradient. b) The incidence and scattering angle is $90°$. c) The incidence and scattering angle is $160°$. The lowest wave number contribution is given the high incidence and scattering angles. The high wave number contribution is given by small incidence and scattering angles.

The wavenumber resolution capacity k_x, k_z for a given source receiver-pair therefore depends on the imaging condition, the acquisition and position of the reflector (that determine θ), the

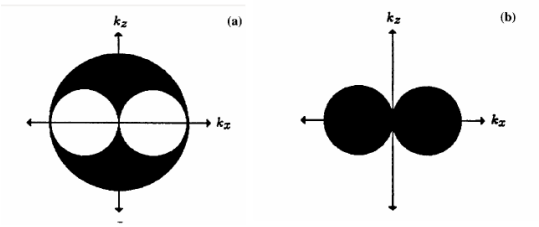


Figure 2.19: Figure from Mora (1989). a) Shows the part of the usual part of wave number spectrum in the image space that can be resolved with the infinite offset data. As it is clear, there is a gap in low number spectrum. b) Shows the part of the wave number spectrum in the image that can be recovered by introducing reflectivity information in the model, and expanding the imaging condition 20, in forward and back-scattered direct and back-propagated wavefields.

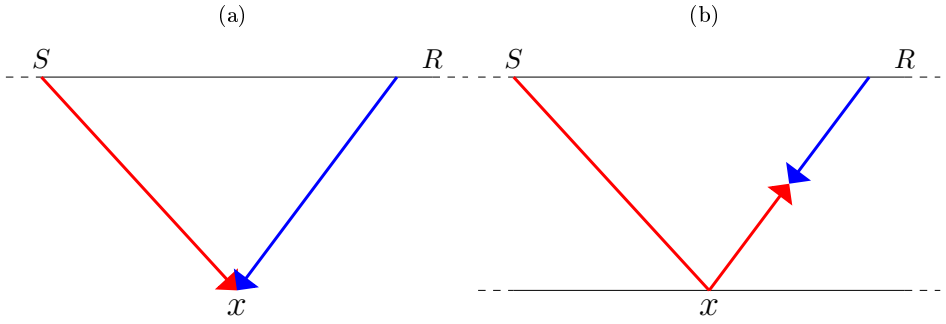


Figure 2.20: Imaging condition with and without a reflector in the background model. a) correlation of u and λ. Reconstructs high magnitude wavenumbers. b) Correlation of back-scattered u and λ. Reconstructs low magnitude wavenumbers.

magnitude of the propagation velocity of the waves v and the frequency f. A representation of the usual wave number illumination is shown in Figure 2.19a, using infinite offset data. There is a gap in low horizontal and vertical number spectrum. For example, note that for $k_z = 0$, there is no resolution in k_x. On the other hand, if $k_x = 0$, the whole k_z can be recovered. This is a consequence of the fact that we are using a surface acquisition and so, the resultant sum of the wavevector usually has k_z as its most important component. To try to recover the low wavenumber (tomographic information), Mora (1989) suggested to introduce reflectivity information in the model (Mora, 1989). The idea is that the reflector will create back-scattered energy. Therefore, the waves will not only coincide in a point x, as shown in Figure 2.20a, but the reflected direct wave will meet the back propagated wavefield from the receiver at other points, as illustrated in Figure 2.20b [3]. This configuration resembles a tomographic configuration where the reflector would correspond to the source and the receivers are on the opposite side. The resultant wavenumber has a low magnitude $|\vec{k}|$. By doing this, it is possible to recover the low wavenumber spectrum and fill the gap as shown in Figure 2.19b (Mora, 1989). In general, in the presence of strong reflectors, the direct and back-propagated wavefields can be separated in those travelling downward and upward, $u = u^+u^-$ and$\lambda = \lambda^+\lambda^-$, where $+$ indicates downward and $-$ indicates upward travelling. The correlation of u and λ will therefore provide four terms. The correlation of wavefields travelling in opposite directions will cover the low part of the wavenumber spectrum.

A 2D example of a layer velocity model in which this principle is applied is done by Wang

3. The opposite situation, in which the back-propagated wavefield is reflected and creates back-scattered energy that correlates with the direct wavefield is also possible

et al. (2013b) and shown in Figure 2.22. Since the sources and receivers are only placed on the surface and have a limited offset, the main information that is recorded are the arrivals of reflected waves when they detect the lower or higher border of the layer. Thus the imaging condition (2.17) provides mainly high frequency information. When a decomposition of the (direct and back-propagated) wavefields in down going and up-going is done, the four components of the gradient are shown in Figure 2.22. The high frequency components (correlation of wavefields travelling in the same direction) are shown in Figures 2.22a-b, which images the borders of the velocity layers. Without any modification to (2.17), the imaging would be more or less the sum of Figures2.22a-b. The deficiency of the imaging condition is revealed by noticing that, even though the interior of the velocity layer does not have the correct velocity value, the imaging condition in the interior of the layer is zero. However, when extracting the low frequency components of the imaging condition (correlation of wavefields travelling in opposite directions) this is no longer the case, as shown in Figures 2.22c-d. Although there are still some regions in the velocity layer for which the imaging condition is still zero, the low frequency components of the imaging condition represent an improvement with respect to the information provided by the high-frequency components. The extraction of the low frequency component is difficult because, as can be seen in the Figure 2.22, the amplitudes are much weaker compared to high frequency components. Interestingly, in RTM the same principle is used but not to enhance the low frequency correlations but rather filter them out. This is discussed in Appendix 1.

(a)

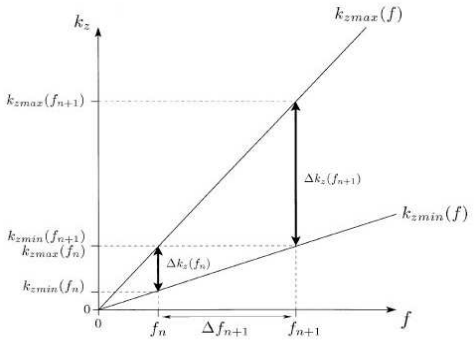


Figure 2.21: Figure from Sirgue and Pratt (2004). Illustration of how to choose the frequencies to attain a complete wave number coverage with a minimum number of frequencies. Each frequency has a range of wave numbers it can recover $[k_{min}, k_{max}]$. The frequencies should be chosen in such a way to not have holes in the wave number coverage.

Summary resolution limitations

Summarizing, one of the limitations when using the Born approximation to solve the direct problem is that a scale separation is imposed between m_0 and δm and only model perturbations can be found from the scattered wavefield. In surface acquisitions configurations, this means that only reflectivity perturbations can be reconstructed and in transmission acquisitions this means that only transmission perturbations can be reconstructed. With the capability to solve the wave equation numerically for any complex model, the Born approximation is no longer needed to solve the forward problem (18). There is therefore no separation between a model m_0 and a perturbation model δm, and by using all the waveform in the inversion, all features of the model (low and high wavenumbers) are susceptible of being reconstructed. Nonetheless, applications of FWI and resolution analysis show that there is a scale separation that remains because both

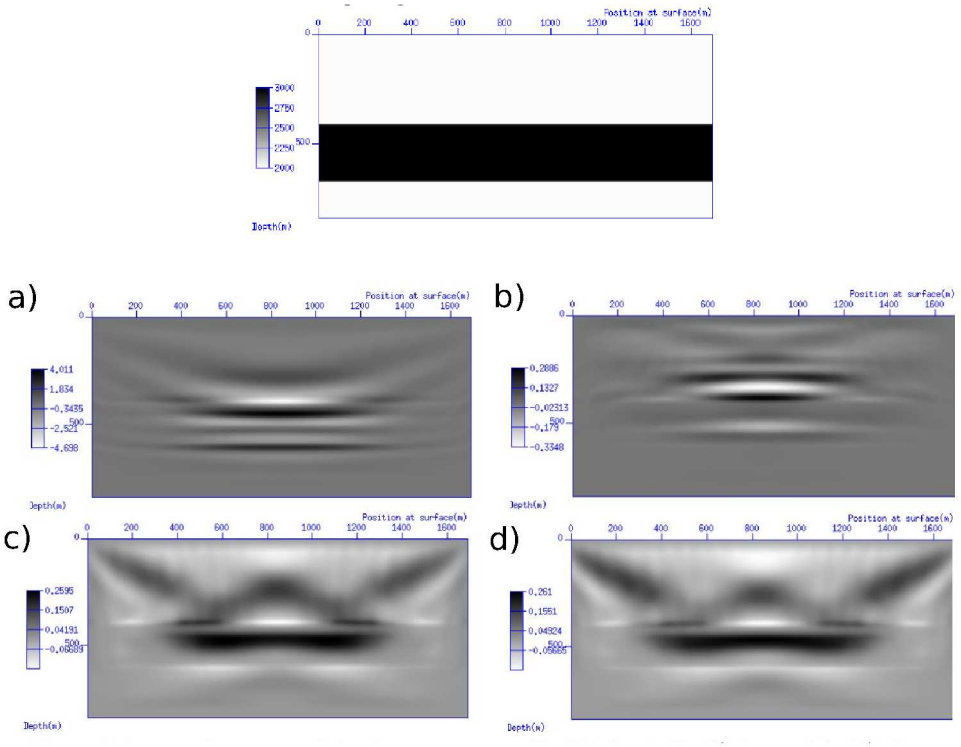


Figure 2.22: Figure from Wang et al. (2013b). For the 2D velocity layer model shown at the top, the high frequency components of the imaging condition are shown in figures a) and b). The low frequency components of the imaging condition are shown in Figures c) and d). The separation of the imaging condition is possible through a decomposition of the (direct and back-propagated) wavefields in up-going and and down-going wavefields (Mora, 1989).

transmission and reflection energy are necessary to effectively reconstruct all features of the model. The acquisition configuration and prior reflectivity information (Mora, 1989) are therefore crucial to exploit the full power of FWI and to have full coverage of the wavenumber spectrum and overcome the scale separation. A schematic representation of this scale separation is shown in Figure 2.23. Current efforts to overcome these difficulties are orientated towards increasing the offsets in surface acquisition geometries to measure transmitted (tomographic) energy. There are also attempts to create seismic sources with lower frequency content and better signal to noise ratio. There are also methodological efforts oriented in either modifying the imaging condition (Mora, 1989; Ma et al., 2010; Brossier et al., 2013b; Wang et al., 2013b), or changing the FWI work flow and, respecting the scale separation and updating iteratively the low and high wavenumbers (Almomin and Biondi, 2012; Biondi and Almomin, 2013).

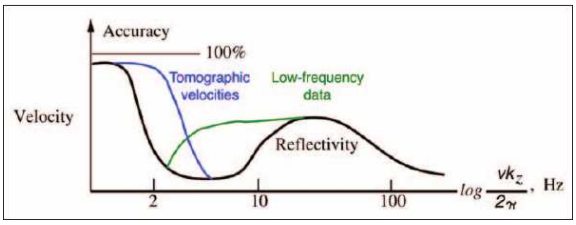


Figure 2.23: Figure from Biondi and Almomin (2013), adapted from Claerbout (1985). The black line schematically represents the scale separation in seismic imaging. The high frequencies correspond to the reflectivity information, and the low frequencies to the background model. The green line represents the ongoing attempts made to extract more low frequency information from the reflectivity data Biondi and Almomin (2013).

Chapter 3

Speed-up of FWI with source encoding

To perform 3D elastic full waveform inversion in time or frequency domain with the current computational resources, time marching or iterative solvers appear to be the most feasible solution. The number of direct problems are proportional to the number of sources, and with time marching or iterative solvers, each direct problem is computationally expensive. Instead of working with each source individually, source encoding techniques allow to work with linear combinations of the sources. The advantage is that the computational cost per iteration is greatly reduced. The counterpart that many more iterations need to be performed because the model update in each iteration is less precise, due to undesired interference effects between sources. In the end, there will be a computational gain (speed-up) if the number of direct problems solved to attain a predefined value of misfit function is less with source encoding than using the sources individually in the standard way.

Although source encoding techniques have been widely used (Neelamani et al., 2008; Krebs et al., 2009; Ben Hadj Ali et al., 2011; Schuster et al., 2011; van Leeuwen et al., 2011; Huang and Schuster, 2012), they have been mostly combined with gradient optimization algorithms, resembling stochastic gradient algorithms (Robbins and Monro, 1951; Spall, 2003). With the purpose of reducing even more the computational cost and improving the computational efficiency, we combine quasi-Newton and Newton optimization methods with source encoding techniques. However, the computational cost per iteration of the truncated Newton methods (Gauss-Newton and Full Newton) is higher per iteration, because the Newton descent direction requires the solution of an additional linear system. In the case of $l-$BFGS, we periodically restart the Hessian approximation. Therefore, a priori, it is not clear whether the computational savings can be further improved with second order optimization methods.

We compare the convergence and the computational efficiency of four optimization methods (non-linear conjugate gradient, $l-$BFGS, Gauss Newton (GN) and Full Newton (FN)) when they are implemented in efficient frequency-domain FWI with and without random source encoding. A stopping criterion of iterations is carefully designed allowing for a fair comparison of the opti-

mization methods and a fair assessment of the speed-up provided by the random source encoding method. These work flows are first applied on a realistic synthetic experiment inspired by the geology of the Gulf of Mexico both for noise-free data and noisy data. Then, we assess the benefit provided by random source encoding and second-order optimization methods when applied on a *2D* real ocean-bottom-cable (OBC) dataset recorded from the Valhall oil field. Even though the maximum advantage of source encoding can be seen with time marching or iterative solvers, we work with direct solvers. However, the conclusions we draw are based on the number of direct problems solved and therefore are directly applicable to iterative solvers in the frequency domain.

We found that, in an ideal noise-free data scenario with frequency groups that determine an approximately convex misfit function, truncated Newton methods remain more computationally expensive than the quasi-Newton method l-BFGS. As noise is added to the synthetic data and more aggressive regularization is used, the action of the Hessian becomes less effective and the convergence rate of the Newton-based methods is thus degraded. This contributes to level down the convergence rate of Newton-based methods relative to steepest-descent method. While all of the optimization methods generate subsurface models of similar accuracy for the synthetic example, application on real data from the Valhall field shows that the truncated Newton methods attain a lower misfit function value than the other optimization methods, hence suggesting a more robust behavior to noise and other source of errors such as incomplete wave physics. A speed-up of nearly one order of magnitude was reached for the selected stopping criterion of iterations. The accuracy of the subsurface models that was achieved for this stopping criterion of iteration was validated against published previous works, a sonic log and reverse time migration. Finally, we derive some formulas for the estimation of the variance of the encoded gradient, that can aid towards source encoding strategies. Parts of the contents of this chapter can be found in (Castellanos et al., 2013).

1 Introduction

FWI has been shown to be quite successful in *2D* applications under the acoustic (Plessix et al., 2012) and elastic (Brossier et al., 2009a) approximation. More recently, *3D* FWI has also been made possible thanks to the increasing development of high performance computing (Plessix, 2009; Plessix and Perkins, 2010; Sirgue et al., 2010; Vigh et al., 2013). However, FWI remains computationally intensive, particularly in three dimensions, when high frequencies are injected in the inversion, or complex wave propagation (visco-elastic) is accounted for. Hence, there is a natural interest to reduce this computational burden.

The full wave form inversion algorithm alternates the solution of a direct and an inverse problem. The direct problem consists in solving the wave equation, and the inverse problem consists in solving an optimization problem. Solving the discrete direct problem (in the time or frequency domain) amounts to solving the linear system

$$A(x,m)u(x,m) = s(x), \tag{3.1}$$

where $A(x,m)$ is the wave equation operator, $s(x)$ the source term and $u(x,m)$ the wave field solution. In FWI, the direct problem is often a computational bottleneck because the number of times the wave equation must be solved is proportional to the number of sources. For example, at each iteration, the computation of the gradient requires the solution of 2 N_s direct problems, and this may increase if a more precise descent direction is taken and due to the line search evaluations.

One possible solution to reduce the number of right hand sides in equation (3.1), known as source blending, is to fire several sources simultaneously or with some time delays at the acquisition level when the data is being collected, (Beasley, 2008; Berkhout, 2008). The measured data therefore has the wavefield response of the individual sources and the interference effects with each other. Source blending reduces the time of collection data in the field and also the number of sources that are processed when solving the direct problem. Another popular speed-up technique, called *source encoding*, consists in performing a traditional seismic acquisition experiment using only one source at time, for N_s sources. When the information is being processed, a limited number of super sources $\tilde{s}$ are created by performing a linear combination of individual sources that are weighted by random coefficients ($\tilde{s} = \sum_{i=1}^{N_s} \alpha_i s_i$). This technique does not save time in the acquisition of the data, but is more flexible at the processing level. It was first proposed for migration by Romero et al. (2000) using random phase encoding. After the pioneering work of Romero et al. (2000), random source encoding has been widely used in time-domain and frequency-domain FWI (Neelamani et al., 2008; Krebs et al., 2009; Ben Hadj Ali et al., 2011; Schuster et al., 2011; van Leeuwen et al., 2011; Huang and Schuster, 2012). As a result of the linearity in the wave equation between s and u, the encoded wavefield $\tilde{u}$ is a superposition of the individual wavefield solutions ($\tilde{u} = \sum_{i=1}^{N_s} \alpha_i u_i$). The same assembling operation is applied on the observed data, and on the adjoint wavefield in the gradient computation. The random coefficients must satisfy $\mathbb{E}[\alpha_i^* \alpha_j] = \delta_{i,j}$.

Figure 3.1 shows schematically how the encoded data are generated. Let N_s denote the number of sources, N_r the number of receivers and we will assume that all the sources share the same number of receivers. Under this assumption, it is possible to represent the recorded data in a matrix form, like the gray matrix in Figure 3.1. Each column represents the recorded data generated by one source, at all the receiver positions. The encoded recorded data $\tilde{d}$ is created by performing a linear combination of the data generated by all the sources, with weights given by a random encoding vector. For example, by taking the recorded data at the first receiver position generated by all sources (first row in blue of the data matrix), and multiplying it by a random encoding vector (vector in yellow), we obtain the entry of the encoded data at the first receiver. Doing the multiplication of the whole data matrix with the encoding vector, results in the encoded data at all receiver positions. Note that this approach is only valid when all sources share the same receivers, so that we can can assemble the data matrix. In general, K different super sources can be assembled by using K different realizations of the random encoding vectors.

Ideally, using source encoding will provide a computational gain because instead of solving the wave equation N_s times to solve the forward problem (3.1), we only need to solve the wave equation K times (where usually $K \ll N_s$). However, the computational gain is in practice much less because the descent direction with source encoding is less accurate than the descent direction provided by the full set of sources because it will be contaminated with artefacts. Consequently, with source encoding more iterations will be required to attain the same level of accuracy in the final model as that obtained without source encoding. For a fixed computational time and using source encoding, many low cost iterations with low accurate updates can be performed. Using the full set of sources for the same computational time, fewer iterations with more accurate updates can be done. The trade off that source encoding provides lies in the idea that each update does not need to be as accurate as possible and, on the other hand, it is more important to perform iterations that will allow the model to quickly evolve.

The model update provided in each iteration using source encoding methods is less accurate

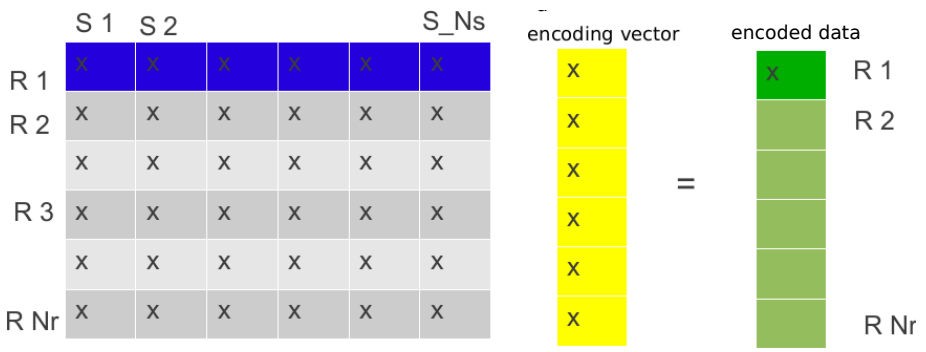


Figure 3.1: Let N_s denote the number of sources, and N_r the number of receivers. If all the sources share the same number of receivers, it is possible to represent the recorded data in a matrix form (the gray matrix). Each column represents the recorded data generated by one source, at all the receiver positions. The encoded data (in green) is generated by performing a linear combination of the data generated by all the sources, where the weights are given by a random encoding vector (in yellow).

due to the so-called cross talk noise which arises from the interference between the assembled sources. The gradient of the misfit function is the sum over all sources of the zero lag correlation between the direct and back-propagated wave field

$$g(x) = -\omega^2 \sum_{s=1}^{N_s} \mathcal{R}e \left\{ u_s^\dagger(x,\omega) \lambda_s(x,\omega) \right\}. \tag{3.2}$$

When the sources are encoded into *one super source*, the gradient becomes,

$$\tilde{g}(x) = -\omega^2 \mathcal{R}e \left\{ \tilde{u}^\dagger(x,\omega) \tilde{\lambda}(x,\omega) \right\}. \tag{3.3}$$

When the correlation between the encoded wave fields $\tilde{u}^\dagger(x,\omega)$ and $\tilde{\lambda}(x,\omega)$ is performed in (3.3), implicitly there are terms that correlate the direct wavefield from a source u_i, with the back-propagated wavefield of the same source λ_i, weighted by a factor $\alpha_i^* \alpha_i$. These correlations are, up to a constant, equal to the terms appearing in 3.2. In addition, there are terms that correlate the wavefield from a source u_i, with the back-propagated wavefield due to another source λ_j, weighted by a factor $\alpha_i^* \alpha_j$, where $i \neq j$. These correlations $\alpha_i^* \alpha_j u_i^\dagger \lambda_j$ do not appear in (3.2) and have no physical meaning in terms of the imaging condition. They represent artefacts in the gradient, that consequently also appear in the model. However, since it was imposed $\mathbb{E}[\alpha_i^* \alpha_j] = 0$, the average of these terms with many iterations will tend to zero.

Most of the literature has used gradient descent algorithms (steepest descent or conjugate gradient) in the optimization problem with source encoding, resembling stochastic gradient algorithms (Robbins and Monro, 1951; Spall, 2003). We wanted to explore if these artefacts could be attenuated in a fewer number of iterations using information about the Hessian. That is, we wanted to address the question weather second order methods could improve the convergence rates with source encoding, which had not been done in the context of FWI. In other application areas such as machine learning, this exploration is also relatively recent (Schraudolph et al., 2007). The lack of convergence proofs of second-order stochastic methods (Bottou and Le Cun, 2005), and the lack of efficient formulations to compute the Newton descent directions (Métivier et al., 2013b, 2014), are some of the reasons. We used the quasi-Newton method l-BFGS, and two truncated Newton algorithms, (Gauss Newton and Full Newton) (Métivier et al., 2013b). We performed synthetic numerical tests in the BP salt model inspired by the geology of the Gulf

of Mexico. The BP salt model is a challenging imaging model because there are high contrasting velocity structures (the salt domes) that generate very strong reflections and, shadow the weaker transmitted waves. We performed the synthetic numerical tests with and without adding Gaussian independent noise to the data.

The real data set application is done with the OBC data from the Valhall oilfield in the North Sea near Norway (Figure 3.2a). Valhall has been in production since 1982 and is projected to continue in production for at least another twenty years. The drilling conditions are difficult because the rock has a high porosity and a low permeability, resulting weak rocks that collapse when drilling. In addition there are gas clouds in the overburden, that until 1997 could not be imaged. The difficult drilling and imaging conditions have pushed advances in seismic acquisition, processing and imaging. In 2003 permanent ocean bottom cables (OBC) were installed (Figure 3.2b). The acquisition was wide-azimuth, wide-offset and allowed to create better images, including an image of the gas cloud. About $50,000$ sources and $2,414$ receivers were used. A 3D frequency domain inversion of this dataset was performed by (Sirgue et al., 2009), which demonstrated the high resolution power of FWI. Figure 3.3a shows a 2D profile of the reflection tomography velocity model, and Figure 3.3b shows the final FWI model using frequencies from $3.5Hz-7Hz$. The low velocity region between $1500m-2500m$ corresponds to a gas cloud. The base reservoir level at $\approx 2600m$ is characterized by a large velocity contrast. As can be seen from these results, the resolution in the velocity model is considerably improved by FWI. In the deep regions, for example, some reflectors that were not present in the tomography modelled have appeared. The difference becomes more evident when the velocity models at selected depths are compared. In Figure 3.4, the velocity maps provided by reflection tomography and FWI are compared at a depth of 150 and $1050m$. The channels appearing at 150 m below the surface can be imaged with high resolution. In Figures 3.4b and 3.4d, the gas cloud is imaged with a higher lateral resolution.

We will use one line of the Valhall dataset and perform a 2D frequency anisotropic mono-parameter inversion (Gholami et al., 2013b,a; Prieux et al., 2013a,b), initially without source encoding, to compare the convergence and final solutions provided by different optimization algorithms. Following, we apply source encoding using one super source and recover velocity models with a comparable quality.

The chapter is organized as follows. We first give a brief overview in Section 2 of several strategies that are available to speed up FWI, and we explain why we focus on source encoding. Following, we review the basic principles of FWI in section 3. We recall the role of the Hessian in FWI in section 3.2 and we describe how to efficiently perform matrix-free Hessian-vector products in the truncated Newton methods using the a second-order adjoint-state approach. We describe the preconditioner that we used in section 3.3, as it plays an important role in the convergence of the optimization methods. The interfacing of source encoding with FWI is described in section 4. In the Appendix 6, we rewrite the misfit function of the FWI with a compact Dirac notation to recast the source encoding as a simple change of basis in a vector space. The applications on a synthetic and real case studies are presented in section 5. We perform the comparison on two levels. On a first basis we compare the convergence rates and costs among different optimization methods with and without source encoding, and with and without noise. On a second instance, for each optimization method we determine the potential computational savings that can be attained (speed-up) with source encoding, and how it is affected by random noise. For the synthetic example, all the optimization methods converge to a subsurface model of similar accuracy with and without source encoding. We show that the truncated Newton methods have a higher convergence rate than $l-$BFGS and CG, although $l-$BFGS is the fastest method. Although CG is the slowest method, it shows the highest speed-up when source encoding is used because its convergence is less affected by cross-talk noise than Newton-based methods. We also

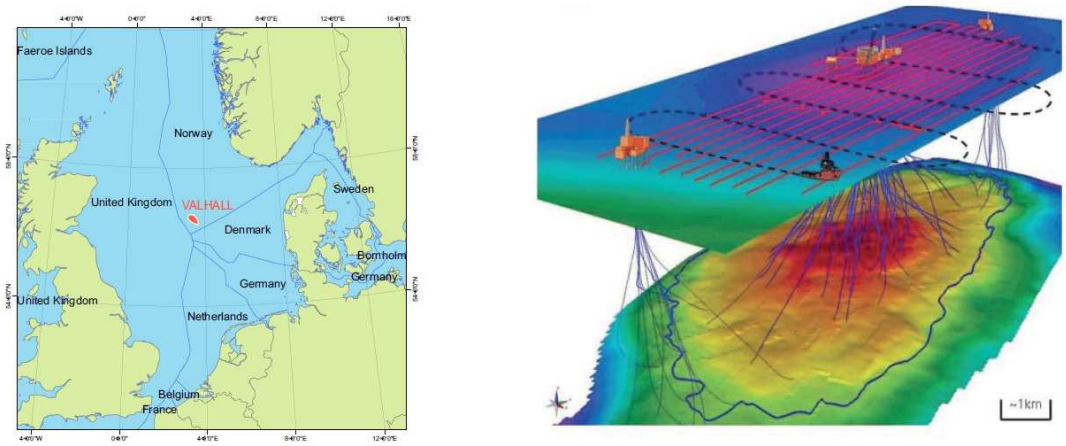


Figure 3.2: Figure from (van Gestel et al., 2008) a) Location of the Valhall oil field in the North Sea. b) General overview of the Valhall oil field. The blue lines correspond to the drilling wells. The red lines are the 4 components sensors. The topography of the sea bottom is imaged with a color scale.

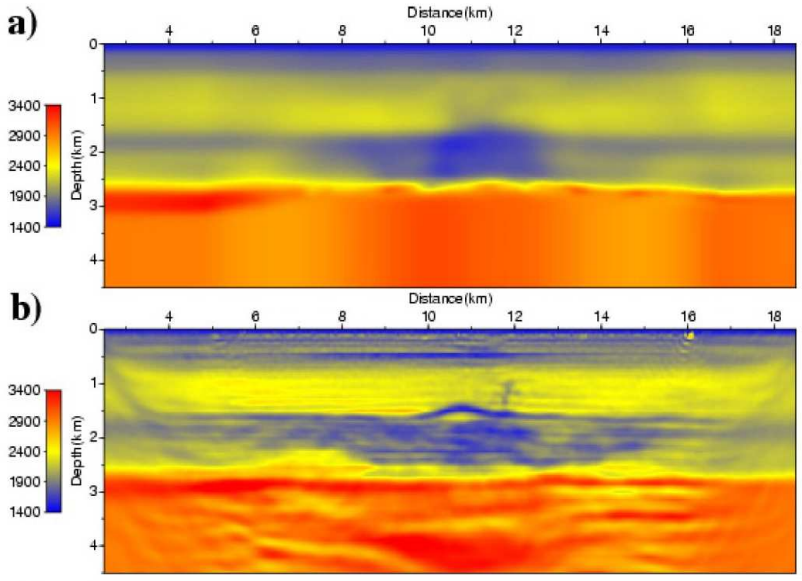


Figure 3.3: Figure from (Sirgue et al., 2009). a) A 2D vertical profile of the reflection tomography velocity model. b) A 2D profile of the final FWI model using frequencies from $3.5Hz - 7Hz$. From the gas cloud (in blue) there is a thin fault at a distance close to 12 km. This horizontal slice of this feature is shown below in Figure 3.4.

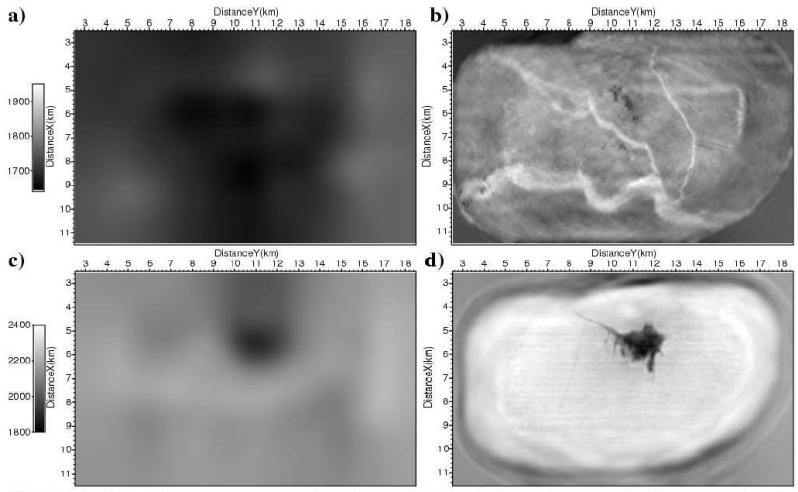


Figure 3.4: Figure from (Sirgue et al., 2009) 2D horizontal velocity profiles. a) At a depth of $150m$, using reflection tomography. b) At a depth of $150m$, using FWI. c) At a depth of $1050m$, using reflection tomography. b) At a depth of $1050m$, using FWI.

show that, when the misfit function is strongly non-convex, source encoding can help to guide the inversion toward an improved minimum of the misfit function, thanks to a broader exploration of the model space. The real data case study shows that the truncated Newton methods provide the most robust direction of descent leading to subsurface models of similar accuracy, regardless of the encoding. A speed-up of nearly one order of magnitude was attained thanks to a careful design of the stopping criterion of iteration.

2 Speed-up FWI : state of the art

FWI has been shown to be quite successful in *2D* applications under the acoustic (Plessix et al., 2012) and elastic (Brossier et al., 2009a) approximation. More recently, *3D* FWI has also been possible thanks to the increasing development of high performance computing (Plessix, 2009; Plessix and Perkins, 2010; Sirgue et al., 2010; Vigh et al., 2013). However, FWI remains computationally intensive, particularly in three dimensions, when high frequencies are injected in the inversion, or complex wave propagation (visco-elastic) is accounted for. Hence, there is a natural interest to reduce this computational burden.

In the simplest scenario, computational savings can be achieved by sub-sampling the data, hence reducing the volume of information injected in the inversion. An illustrative sub-sampling approach is the so-called efficient frequency-domain FWI (Sirgue and Pratt, 2004), where only a few discrete frequency components are processed in a hierarchical manner. Note that the sub-sampling is successful if the redundancy contained in the original data set is carefully decimated. In efficient frequency-domain FWI, the redundancy in the wavenumber coverage is reduced by coarsening the frequency interval, while avoiding wraparound in the space domain (Sirgue and Pratt, 2004). This frequency sub-sampling can be combined with another one performed in the source-receiver space (Gao et al., 2010; Habashy et al., 2011). In the source-receiver space, monochromatic data are stacked with judicious weighting coefficients determined by a singular value decomposition (SVD) of the source-receiver data matrix. This method not only reduces the data volume, but also improves the signal to noise ratio because small eigenvalues, which may correspond to noise, are not taken into account in the inversion. The computational gain

depends on the accuracy of the truncated SVD, which in turn, also depends on the redundancy of the original data set for each frequency. The setback of the sub-sampling methods is that we may not always have a redundant original data set, and not using all information at the same time may be detrimental.

Another possibility for compressing the data volume consists in creating a limited number of super sources by linear combinations of individual shots that are weighted by random coefficients, thanks to the linearity of the solution of the wave equation with respect to the source terms. This approach was first proposed for migration by Romero et al. (2000) using random phase encoding. More recently Godwin and Sava (2013) compared this random encoding method with deterministic methods to choose the coefficients and assemble the sources. In the deterministic methods, the weighting coefficients can be inferred from discrete Fourier transform (DFT) (Nihei and Li, 2007; Dai et al., 2013), discrete Hartley transform (Strang and Nguyen, 1996), which resembles a real valued Fourier transform, or appropriate time delays as when plane-wave migration is performed (Vigh and Starr, 2008; Dai and Schuster, 2013). The motivation behind the projection of the dataset on a Fourier or Hartley basis results because the basis vectors of these transformations are orthonormal. Thus, in the limiting case where the number of super sources is equal to the number of sources, each super source is encoded with a basis vector and the source encoded inverse problem is equivalent to the non encoded one. However, when the number of super sources is less that the number of sources, as should be the case, the encoding matrix has non-zero off diagonal terms. These off diagonal elements create undesirable cross-correlations between different sources, referred to as crosstalk noise. The illustrative and insightful comparison in Godwin and Sava (2013) shows that the DFT and Hartley transform perform better as the number of super sources tends towards the original number of sources, as expected. The plane wave migration produces the poorest results for any number of super sources. The random phase encoding produces the migrated image with the lowest error when a few number of super sources is used. It thus appears that the maximum computational gain can be obtained through the use of random encodings. This is the reason why we focus on this method in this study.

After the pioneering work of Romero et al. (2000), random source encoding has been widely used in time-domain and frequency-domain FWI (Neelamani et al., 2008; Krebs et al., 2009; Ben Hadj Ali et al., 2011; Schuster et al., 2011; van Leeuwen et al., 2011; Huang and Schuster, 2012). Random codes include phase shifts, time shits, convolution with random variables, among others. As is the case for migration, the setback of assembling the sources in super sources is the crosstalk generated by the undesired correlations of the incident wavefield emitted by one individual shot with the back-propagated wavefield emitted by the data residuals associated with the other shots involved in the super-shot setting during the gradient building. Regeneration of the random encoding at each non-linear iteration of the FWI is a possible approach to make the sum incoherent as the number of iterations increase. The sum over iterations to make the average term of the crosstalk noise tend to zero is an averaging technique that, as any Monte-Carlo method, has a slow convergence rate of the order $O(1/\sqrt{I})$ where I is the number of iterations (Nemirovskići and ëIěUdin, 1983). In contrast, deterministic methods using all the sources may have better convergence rates as, for example, $O(1/I)$ (Boyd and Vandenberghe, 2004). Therefore, using a few super sources allows one to dramatically reduce the computational cost of each non-linear iteration of FWI, but the low convergence rate of averaging techniques may impact the overall computational gain. Another setback is that the random encoding strategy is only valid for common receiver acquisitions. Some attempts have been made to handle marine streamer acquisitions by creating, for example, subsets of data that share the same receivers, or using correlation-based misfit function, which are less sensitive to amplitude effects (Baumstein et al., 2011; Choi and Alkhalifah, 2012).

Alternatively a hybrid technique has been developed to improve the convergence rate of stochastic methods (Friedlander and Schmidt, 2012). This technique does not use super sources but rather it is based on a batching strategy where an increasing number of randomly chosen

independent sources are used as the number of iterations grows. Since there is no assembling of super sources, no crosstalk noise is generated, and fixed-spread acquisition geometries are no longer required. The inversion resembles a stochastic optimization problem at the beginning of the inversion when a few number of sources are employed. Although the overall convergence rate of this hybrid method is theoretically superior (Friedlander and Schmidt, 2012), van Leeuwen and Herrmann (2012) showed numerically that, for FWI applications, the convergence rate of this method coupled with a l-BFGS minimization is similar to the convergence rate of the random encoding technique combined with a steepest descent algorithm. This may result because the hybrid method does not use all the information from all sources at the beginning of the minimization process. We therefore do not follow a hybrid approach to improve the convergence rate in the present study. Instead, we interface random source encoding methods with second-order optimization algorithms, which are expected to have higher convergence rates.

Theoretically, stochastic optimization is only proven to converge using a steepest-descent optimization algorithm (Robbins and Monro, 1951; Spall, 2003), under certain restrictions on the step-length computed at each iteration. However, in standard FWI, l-BFGS has shown to improve the convergence (Brossier et al., 2009a). This quasi-Newton method approximates the inverse of the Hessian by performing successive rank-2 updates of an initial estimation from the gradients and the models of the previous l iterations (Byrd et al., 1995). The recursive update of the Hessian over iterations in the $l-$BFGS method can be affected by the source encoding method when the random codes are regenerated at each FWI iteration because a regeneration of the code changes the misfit function and hence its gradient. Therefore, we first study how the l-BFGS method can be coupled with source encoding strategy based on random encodings, and if a computational gain can be expected from this optimization method.

Truncated Newton optimization methods such as Gauss-Newton (GN) or the full Newton (FN) (Métivier et al., 2013b, 2014) are second-order optimization methods that can be considered in the random source encoding framework. From a source encoding point of view, the advantage of these methods relative to l-BFGS is to account for the action of the Hessian only from quantities available at the current iteration. Therefore, the regeneration of the random codes at each non-linear iteration of the inversion is no longer an issue, as it is for the l-BFGS method. The drawback is that the (Gauss-)Newton approaches require additional seismic modelings per iteration. Therefore, we need to assess whether the higher computational cost of one non-linear iteration of the truncated Newton methods can be balanced by an improved convergence rate provided by a more accurate estimation of the Hessian.

3 Method

3.1 Full waveform inversion problem

Let us define a space $\Omega \subset \mathbb{R}^2$ as a subsurface medium with spatially varying model parameters $m(x)$ which may be, for example, the density $\rho(x)$ and the P-wave velocity $v_p(x)$ in the acoustic approximation. In the frequency domain, the wavefield $u(x, m, \omega)$ satisfies the wave equation,

$$A(x, m, \omega)u(x, m, \omega) = s(x, \omega), \tag{3.4}$$

where $A(x, m, \omega)$ is the forward modeling operator, which in the acoustic approximation is

$$A(x, m, \omega) = -\frac{1}{\rho(x)v_p(x)^2}\omega^2 - \nabla \cdot \left(\frac{1}{\rho(x)}\nabla u(x, m)\right), \text{ on } \Omega, \tag{3.5}$$

and the source function is denoted by $s(x, \omega)$. We impose a free surface boundary condition at the surface and absorbing boundary conditions on the other boundaries to simulate an infinite

half-space.

The inverse problem thus consists in finding the model m that iteratively minimizes the misfit function ϕ that measures the distance from the observed data d to the simulated wavefield u (e.g. Tarantola, 1984a). The wavefield $u(x)$ is defined on Ω ($u(x) : \Omega \rightarrow \mathbb{R}$) and $d(x)$ on Ω_r ($d(x) : \Omega_r \rightarrow \mathbb{R}$) , where Ω_r denotes the receiver space. We use as the misfit function the $_2$ norm of the difference between the modeled and the recorded wavefields,

$$\min_m \phi(u;m) = \min_m \frac{1}{2} \sum_{i=1}^{N_s} \|Pu_i(x,m) - d_i(x)\|_2^2, \tag{3.6}$$

where we emphasize that the misfit ϕ depends explicitly only on u and implicitly on m, and N_s is the number of sources. Since $\Omega \neq \Omega_r$, we use a projection operator P from the whole space Ω to the receiver space, $P : \Omega \rightarrow \Omega_r$. For sake of compactness we shall consider one frequency, although we shall implement multi-frequencies inversion in our applications for which we add an external sum over frequencies. The minimization of ϕ is an iterative process, where the initial model is updated around an initial model m^n along a direction of descent Δm.

$$m^{n+1} = m^n + \alpha^n \Delta m^n, \tag{3.7}$$

where n is the iteration index, α is a step length (i.e., the quantity of descent), and the descent direction Δm is given by the optimization algorithm of choice. The step length satisfies the strong Wolfe conditions (Nocedal and Wright, 2006). For steepest-descent (SD) algorithms, the descent direction is opposite to the gradient, $\Delta m = -\mathcal{P}^n \nabla_m \phi(u;m)$, where $\mathcal{P}$ is a preconditioner as explained in Section 3.3. The adjoint-state method allows to compute the gradient solving only two direct problems per source, using the following expression (Lions, 1968; Plessix, 2006; Chavent, 2009)

$$\nabla_m \phi(u;m) = \sum_{i=1}^{N_s} \Re \left\{ u_i^\dagger \left(\frac{\partial A}{\partial m} \right)^\dagger \lambda_i \right\} = \sum_{i=1}^{N_s} \Re \left\{ \lambda_i^\dagger \left(\frac{\partial A}{\partial m} \right) u_i \right\}, \tag{3.8}$$

where the back-propagated wavefield λ satisfies the adjoint-state equation

$$A^\dagger \lambda_i = -P^\dagger (Pu_i - d_i). \tag{3.9}$$

For each iteration n, the gradient (3.8) requires the solution of one direct problem for $u(x,m)$ and one for $\lambda(x,m)$, for each source: the gradient computation in each iteration requires to perform $2 \times N_s$ direct problems.

3.2 Second-order optimization methods

In convex optimization, SD algorithms have the lowest convergence rate (linear when α is calculated exactly), but are the easiest to implement. Other first order optimization methods that have better convergence properties include, for example, conjugate gradient (CG) and its variants. Nonetheless, to improve even further the convergence rate to the minimum, it may be beneficial to include information about the curvature of the misfit function around the critical points, known as the Hessian H. Assuming the misfit function close to the starting and current point is quadratic, a Taylor expansion up to second order of the misfit function gives rise to the so-called Newton equations for the model update Δm,

$$H \Delta m = -\nabla \phi. \tag{3.10}$$

If the inverse of the Hessian exists, the model update is equal to

$$\Delta m = -H^{-1} \nabla \phi. \tag{3.11}$$

For FWI, the expression of the Hessian is given by (Pratt et al., 1998)

$$H \quad = \quad \nabla_m^2 \phi = \left(P \frac{\partial u}{\partial m} \right)^\dagger \left(P \frac{\partial u}{\partial m} \right) + \left(\frac{\partial^2 u}{\partial m^2} \right)^\dagger P^\dagger \left(P u - d \right). \tag{3.12}$$

The first-order term of the Hessian, first term in equation 3.12, is built by the sum over the source-receiver pairs of the zero-lag correlation of the restriction at the receiver positions of the partial derivative of the modeled wavefields with respect to distinct model parameters. Multiplying the gradient by the inverse of this operator gives a model perturbation with correct physical units by removing wave-propagation effects such as geometrical spreading and deconvolving the gradient from limited bandwidth effects. The second-order term of the Hessian in equation 3.12, aims to correct the gradient for artefacts associated with double-scattering effects. Recall that the artefacts are due to the approximation made in the gradient building where only single-scattering events are considered for model updating (Pratt et al., 1998).

Despite its importance, involving the Hessian in the optimization process is often computationally too expensive. The BFGS method handles this difficulty by storing in memory all previous gradient computations, and approximates the inverse Hessian by performing successive rank-2 updates of an initial estimation (Byrd et al., 1995). The limited memory BFGS (l-BFGS) proceeds in the same way, but only keeps in memory the previous l computations of the gradient and of the solution (Nocedal and Wright, 2006). Specifically, the l-BFGS algorithm is quite efficient, as it finds directly the model update using (3.11), through a matrix-free recursive algorithm that computes the multiplication of the l-BFGS inverse Hessian by the gradient (Algorithm 9.1 in Nocedal and Wright (2006)). For the specific case of FWI, l-BFGS has shown very good results (Brossier et al., 2009a) and requires no additional solutions of direct problems.

Alternative second-order optimization methods are the truncated Newton algorithms where the Newton system, equation (3.10), is solved approximately with an iterative method such as the linear conjugate gradient method. The reader is referred to Métivier et al. (2014, 2013b) for a detailed description of the truncated Newton methods that are used in the present study. Only the governing idea is shortly described here.

We do not need the explicit building of the Hessian matrix to solve equation 3.10 with an iterative solver, but rather we only require the capability to perform matrix-vector products as $H\,v$. We introduce the following functional h_v (v is an arbitrary real vector):

$$h_v(m) = \langle \nabla_m \phi, v \rangle \,. \tag{3.13}$$

The gradient of $h_v(m)$ is the Hessian matrix-vector product we want to compute:

$$\nabla h_v(m) = H(m)\, v. \tag{3.14}$$

We compute this gradient with the adjoint-state method with a Lagrangian formalism (Plessix, 2006). Let us define the Lagrangian,

$$\mathcal{L}\left(m, \lambda_i, u_i, \mu_{1_i}, \mu_{2_i}\right)_{i\in[1;N_s]} = \sum_{i=1}^{N_s} \langle \nabla_m \phi(u;m), v \rangle + \sum_{i=1}^{N_s} \Re \left\langle \mu_1, A^\dagger \lambda_i - P^\dagger (P u_i - d_i) \right\rangle + \Re \sum_{i=1}^{N_s} \langle \mu_2, A u_i - s_i \rangle \,, \tag{3.15}$$

where μ_1 and μ_2 are adjoint-state variables. The adjoint-state equations are obtained by substituting $\nabla_m \phi(u;m)$ with its explicit expression in equation 3.8 and deriving with respect to the state variables u and λ. This gives the following relations:

$$A\,\mu_1 \quad = \quad \frac{\partial A}{\partial m} u v \tag{3.16}$$

$$A^\dagger \mu_2 \quad = \quad -P^\dagger P \mu_1 - \frac{\partial A}{\partial m}^\dagger \lambda v. \tag{3.17}$$

The Hessian vector product $H\,v$ is given by

$$H\,v = u^{\dagger}\left(\frac{\partial^2 A}{\partial m \partial m}\right)^{\dagger}\lambda v + \lambda^{\dagger}\frac{\partial A}{\partial m}\mu_1 + u^{\dagger}\left(\frac{\partial A}{\partial m}\right)^{\dagger}\mu_2. \tag{3.18}$$

The computation of $H\,v$ requires to perform two direct problems, one for μ_1 and one for μ_2, per iteration in the solution of the linear system (3.10). In the truncated Newton methods, an approximation known as Gauss-Newton (GN) can be carried out to take into account only first-order scattering term in the Hessian, that is, only the first term of the Hessian in equation (3.12). This amounts to removing all the terms that involve the data residuals and hence the wavefield λ, in the expression of the second adjoint-state equation (3.18). This approximation has an identical computational effort as the full Newton algorithm.

The maximum number of iterations that are executed to solve the linear system (3.10) is restricted by three conditions, which is why they are called truncated Newton methods. To guarantee that the direction taken is towards a minimum and not a maximum, the second derivative must be positive. Hence, it must be verified that the Hessian is positive definite which means

$$v^{\dagger}\,H\,v > 0. \tag{3.19}$$

When this inequality is not satisfied for a given vector v, the resolution of the linear system 3.10 is stopped. In addition, the Eisenstat and Walker criteria (Métivier et al., 2013b) that evaluates the quality of the local quadratic approximation is also employed as a stopping criteria. Alternatively, a maximum number of iterations may be simply imposed by the user. The early stopping of the iteration limits the accuracy of the Hessian estimation at the benefit of the computational cost. This is highly advised otherwise too many iterations may be performed for estimating the solution of the linear system, causing a dramatic increase of the number of direct problems that need to be executed in each iteration of the non-linear inverse problem, while not providing a significant improvement in the model update.

The total Hessian vector product will be the sum

$$Hv = \sum_{s=1}^{N_s}(Hv)_s, \tag{3.20}$$

using in equation (3.18) the wavefields u_s and λ_s produced by each source s. Therefore, for each iteration n, the descent direction given by the FN and GN truncated Newton methods requires the solution of one direct problem for $u(x,m)$ and one for $\lambda(x,m)$, for each source to compute the gradient. For each iteration n, the resolution of the linear system (3.10) requires the solution of one direct problem for μ_1 and one for μ_2, per iteration of the conjugate gradient solver. The total number of direct problems for truncated Newton methods per iteration is therefore $(2 + 2 \times N_{CG}) \times N_s$, where N_{CG} stands for the number of iteration performed by the linear conjugate gradient solver. In practice, truncated Newton methods not only demand a higher computational cost compared to quasi-Newton methods or gradient methods, but they also have higher memory requirements. This arises because to solve equations and 3.16 and 3.16, all the wavefields u_i and λ_i must be stored in memory for $i \in [1, N_s]$. In *2D* applications this is still feasible, but in *3D* study cases, alternative strategies should be investigated.

3.3 Preconditioner

In order to minimize the number of iterations performed by the linear conjugate gradient algorithm and hence to reduce the cost of the truncated Newton methods, we apply a left preconditioner to the Hessian matrix leading to the preconditioned Newton system:

$$\mathcal{P}^{-1}H\Delta m = -\mathcal{P}^{-1}\nabla\phi. \tag{3.21}$$

In the present study, we follow Métivier et al. (2014) and use as preconditioner the diagonal elements of the so-called pseudo-Hessian matrix that was introduced by Shin et al. (2001a) for depth migration. The pseudo-Hessian matrix is formed by the zero-lag correlation of the so-called virtual sources, $\left(\frac{\partial A}{\partial m}u\right)$ (Pratt et al., 1998), while the Gauss-Newton Hessian is formed by the zero-lag correlation of the partial derivative wavefields at the receiver positions, (3.12). The advantage of the pseudo-Hessian matrix relative to the Gauss-Newton Hessian matrix is that its expression does not depend on the receiver positions. Therefore, its diagonal elements can be computed at the same cost than the gradient. Although the wave-paths from the receiver to the model parameter are not taken into account in the pseudo-Hessian, the diagonal elements of the pseudo-Hessian provide a suitable scaling of the gradient that make the deep perturbations well balanced relative to the shallower ones. A damping coefficient β, defined as a fraction of the maximum diagonal coefficient, is added to the diagonal elements of the pseudo-Hessian to prevent instabilities resulting from division by very small numbers (Ravaut et al., 2004).

The same preconditioner is used for the SD algorithm, where the preconditioner is multiplied with the gradient to give the descent direction, which resembles a first approximation to a GN step. For the l-BFGS method, the preconditioner is used as an initial estimation of the Hessian $H_0^t = \mathcal{P}$ in each iteration, also helping the convergence of the optimization (Métivier et al., 2013b).

4 Source encoding

As the solution of the direct problem is commonly the most intensive computational part in FWI due to numerous right-hand sides in the direct problem (3.5), several authors (Krebs et al., 2009; Schuster et al., 2011; Ben Hadj Ali et al., 2011; van Leeuwen and Herrmann, 2012) have explored the possibility of creating a linear combination of the sources into one (or several) super sources $\tilde{s_k}$, defined as

$$\tilde{s}_k = \sum_{i=1}^{N_s} \alpha_i^k s_i, \tag{3.22}$$

where k labels one super source, $k \in [1, K]$. The quantities $\alpha_i^k \in \mathbb{C}$ are random complex scalar coefficients that satisfy (see Appendix 6),

$$\mathbb{E}\left[\alpha_i^* \alpha_j\right] = \delta_{i,j}. \tag{3.23}$$

As a consequence of the linearity of solution of the direct problem u with respect to the source s, and the solution of the adjoint problem λ and its source, it follows that the wavefields can be expressed as,

$$\tilde{u}_k = \sum_{i=1}^{N_s} \alpha_i^k u_i \tag{3.24}$$

$$\tilde{\lambda}_k = \sum_{i=1}^{N_s} \alpha_i^k \lambda_i. \tag{3.25}$$

These are the new encoded wavefields. Following exactly the same procedure as that without encoding, the misfit function and the gradient are

$$\tilde{\phi}(\tilde{u};m) = \frac{1}{2}\sum_{k=1}^{K}\|P\tilde{u}_k(x,m)-\tilde{d}_k(x)\|_2^2 \tag{3.26}$$

$$\nabla_m\tilde{\phi}(\tilde{u};m) = \sum_{k=1}^{K}\Re\left\{\tilde{u}_k^{\dagger}\left(\frac{\partial A}{\partial m}\right)^{\dagger}\tilde{\lambda}_k\right\} \tag{3.27}$$

where $\tilde{d}_k(x)=\sum_{i=1}^{N_s}\alpha_i^k d_i$ denotes the encoded recorded data set corresponding to the super source k.

4.1 Optimization algorithms with source encoding

When encoding the sources, we wish to regenerate the codes as often as possible in order to have an average crosstalk term that tends to zero. For the first-order optimization methods, we degrade the preconditioned $nl-$CG algorithm to a preconditioned SD optimization to use only information of the current iteration, and regenerate the encodings γ_k in every iteration. For the case of l-BFGS, we keep the same encoding for intervals of l iterations with which an estimate Hessian is constructed. At the end of the l_{th} iteration we regenerate the random variables, delete the gradients and models stored in memory, and restart the Hessian from a pseudo Hessian approximation. This restart version of $l-$BFGS, referred to as l-BFGS$_r$ in the following, allows to approximate the Hessian using gradients computed with the same encodings, but also to regenerate the random variables throughout the inversion to attenuate the crosstalk terms. Since l-BFGS does not require any additional computations of direct problems per non-linear iteration, the potential computational gain using this method or SD with source encoding is $2\times N_s$ versus $2\times K$.

Regarding the truncated Newton methods, the wavefields μ_1 and μ_2 can be directly encoded, because of the linear relationship between these wavefields and their source, equations 3.16 and 3.18. Consequently, the equations to be solved for the Hessian using source encoding translate directly to

$$A\tilde{\mu}_1 = \frac{\partial A}{\partial m}\tilde{u}v = \left(\sum_{q=1}^{N}v_q\frac{\partial A}{\partial m_q}\right)\tilde{u} \tag{3.28}$$

$$A^{\dagger}\tilde{\mu}_2 = =-P^{\dagger}P\tilde{\mu}_1-\left(\sum_{q=1}^{N}v_q\frac{\partial A}{\partial m_q}^{\dagger}\right)\tilde{\lambda} \tag{3.29}$$

$$(\tilde{H}v)_p = \tilde{u}\left(\sum_{q=1}^{N}\frac{\partial^2 A^{\dagger}}{\partial m_p\partial m_q}v_q\right)^{\dagger}\tilde{\lambda}+\tilde{\lambda}^{\dagger}\frac{\partial A}{\partial m_p}\tilde{\mu}_1+\tilde{u}^{\dagger}\left(\frac{\partial A}{\partial m_p}\right)^{\dagger}\tilde{\mu}_2, \tag{3.30}$$

where N represents the number of discretization points if the domain Ω. The total Hessian is given by the sum over each of the contributions of the super sources,

$$(\tilde{H}v)_p=\sum_{k=1}^{K}(\tilde{H}v_p)_k. \tag{3.31}$$

The potential computational gain using truncated Newton methods with source encoding is $(2+2\times N_{CG})\times N_s$ versus $(2+2\times N_{CG})\times K$. In the Appendix 6 the formulas for source encoding

are rewritten in a Dirac notation that allow to interpret it as a change of basis and visualize the structure of the cross-talk terms, suggesting a link with trace estimators of random matrices (Roosta-Khorasani and Ascher, 2013).

5 Numerical Examples

We apply frequency-domain FWI on synthetic and real data sets to compare the behavior of different optimization algorithms when combined with source encoding techniques, with and without the influence of noise. The direct problem is solved with the *2D* second-order acoustic wave equation for pressure (3.5), which is discretized with a 9 point mixed-grid finite difference stencil on a regular Cartesian grid of N points. We perform the synthetic experiment with the isotropic wave equation (Hustedt et al., 2004), while we introduce anisotropic effects (for vertical transversely isotropic media) in the modeling during the inversion of the real data (Operto et al., 2009). The absorbing boundary conditions are implemented with perfectly matched layers (PML) (Bérenger, 1994). The linear system resulting from the discretization of the frequency-domain wave equation (3.4) is solved with the sparse direct solver MUMPS (Amestoy et al., 2000), which first performs a LU factorization of the wave-equation operator before computing the monochromatic solutions for multiple right-hand sides (i.e., sources) by substitution (Amestoy et al., 2000). Notice that the same LU factorization is valid for all the sources and only depends on the current model m and frequency ω. This is beneficial for the truncated Newton methods because the LU factors can be re-used to perform the additional direct problems that allow for the Hessian-vector products during the iterative resolution of the Newton linear system. We fix the density and the Thomsen's parameters (for the anisotropic case) and perform a mono-parameter inversion for the wavespeed v_p (the vertical wavespeed in the anisotropic case). Thus, in our case, $m = v_p$, leading to N unknowns.

If the receivers are located on a subset of grid points on which the pressure wavefield u is computed, the operator P is simply a restriction or subsampling operator of the wavefield u on the subspace defined by the receiver positions. If, on the other hand, the wavefield u and the observed data d are not on the same grid, the operator P is a projection of wavefield u on the grid of the receivers through an interpolation operator (Hicks, 2002).

The frequency-domain FWI is decomposed into successive inversions of frequency groups with a limited overlap. Each group is composed of a limited number of discrete frequencies that significantly reduces the intrinsic redundancy of the inverted data. The high-frequency content increases from one frequency group to the next, hence defining a multi-scale approach of FWI, which helps to mitigate the non-linearity of the FWI (e.g., Bunks et al., 1995; Sirgue and Pratt, 2004). We consider fixed-spread acquisitions for both the synthetic and real data case studies, for which source encoding methods are suitable.

We add a Tikhonov regularization term to the misfit function so as to deal with the ill-posedness of the FWI, which results from noise and incomplete illumination provided by surface acquisition.

$$\phi(u;m) = \frac{1}{2}\|\Delta d\|^2 + \frac{1}{2}\lambda_x\|\nabla_x m\|^2 + \frac{1}{2}\lambda_z\|\nabla_z m\|^2 = \frac{1}{2}\|\Delta d\|^2 + \frac{1}{2}\lambda_x\|W_x m\|^2 + \frac{1}{2}\lambda_z\|W_z m\|^2, \quad (3.32)$$

where $\Delta d = Pu - d$. The Tikhonov regularization augments the data-space misfit function with smoothing constraints in the horizontal $\left(\|\nabla_x m\|_2^2\right)$ and vertical $\left(\|\nabla_z m\|_2^2\right)$ directions, whose partial weights are given by two hyper-parameters λ_x and λ_z. It is recalled that the minimization

of the augmented misfit function gives the following descent direction:

$$\Re\left(J^\dagger P^\dagger P J + \frac{\partial^2 J}{\partial m^2}^\dagger P^\dagger\left(\Delta d ... \Delta d\right) + \lambda_x W_x + \lambda_z W_z\right)\Delta m = \Re\left(J^\dagger P^\dagger \Delta d + \lambda_x W_x^\dagger W_x m + \lambda_z W_z^\dagger W_z m\right), \tag{3.33}$$

where $J = \partial u/\partial m$. In the Hessian, $J^\dagger J$ is a smoothing operator. The operators W_x and W_z are roughness operators and hence damp the deconvolution action of the operator $J^\dagger J$ through the hyper-parameters λ_x and λ_z. A suitable trade-off should be found during Newton-based optimization between the need to preserve the deconvolution and the scaling actions of the Hessian for improved convergence rate and spatial resolution and the need to mitigate the effect of noise to follow a robust direction of descent and prevent convergence towards a local minimum. When source encoding is applied, the expected value of the misfit function is the same as that without encoding (213), once a sufficient number of iterations have been performed. However, this is not the case during the early stage of the inversion when the crosstalk term significantly increases the noise level in the data-space misfit function making the problem even more ill-posed. This implies that FWI with source encoding can require to increase the hyper-parameters λ_x and λ_z relative to those of FWI without source encoding in order to account for the increased level of noise during the early iterations.

A suitable stopping criterion of iterations should be defined, as a fair assessment of the speed-up provided by source encoding method when combined with different optimization methods is a key element of this study. The maximum number of non-linear iterations that are performed during each frequency-group inversion is controlled by the three following criteria:

- Criterion 1 : The relative reduction of the misfit function is below a predefined threshold ϵ_1,

$$\phi(m_k)/\phi(m_0) < \epsilon_1. \tag{3.34}$$

  This criterion, which will be the effective one for most of the experiments presented hereafter, allows us to stop each optimization method at a similar level of data misfit reduction. We checked, during the synthetic example, that the chosen value of ϵ_1 allows the FWI to converge towards subsurface models of similar accuracy, for any of the considered optimization methods used. This is a necessary condition for a fair assessment of the computational efficiency of each optimization method. The accuracy of the reconstructed models is quantitatively checked with the 2 norm of the relative model error:

$$Error_m = \frac{\sum_{i=1}^{N}(m_i^k - m_i^{true})^2}{\sum_{i=1}^{N} m_i^{true2}} \tag{3.35}$$

  Note that, even when source encoding is applied, we pay the price to compute periodically the deterministic misfit function using all the sources independently to test whether this stopping criterion of iterations is reached.
- Criterion 2 : The relative change of the misfit function with respect to the average of the misfit function from the previous 30 iterations (ϕ_{avg}) is below a threshold ϵ_2,

$$\left(\phi_{avg} - \phi(m_k)\right)/\phi_{avg} < \epsilon_2 \tag{3.36}$$

  We define this criterion to avoid unnecessary iterations when the previous criterion is not reached and the misfit function does not significantly decrease any more.
- Criterion 3 : A user predefined maximum number of direct problems is attained.

The optimization methods considered when using all the sources independently are the preconditioned non-linear conjugate gradient ($nl-$CG), preconditioned limited memory BFGS (l-BFGS), preconditioned truncated Gauss Newton (GN), and the preconditioned truncated full Newton approximation (FN). When using the encoded sources, the four optimization methods used are

preconditioned steepest descent (SD), preconditioned limited memory BFGS with periodic restart (l-BFGS$_r$), GN and FN. To encode the sources and the data, we created K super sources, using the following three different distributions for the random variables,

$$\gamma_i \in \{-1,1\} \quad \text{where} \quad p(\gamma_i = 1) = p(\gamma_i = -1) = 1/2 \tag{3.37}$$

$$\gamma_i \backsim \mathcal{N}(0,1) \tag{3.38}$$

$$\gamma_i \backsim \exp(-i\chi), \text{ where } \chi \backsim U[0,2\pi], \tag{3.39}$$

which satisfy the desired properties (212). Note that the first two encodings provide an amplitude encoding and the last distribution provides a phase encoding. For a fair comparison of the different optimization algorithms, we use the same value of the hyper-parameters (λ_x, λ_z) and threshold value in the preconditioner β for all optimization methods. When using source encoding, the same number of super sources is used for all optimization methods. Indeed, the hyper-parameters and threshold value in the preconditioner shared by the four optimization methods are adapted during each experiment to the problem at hand, in particular to the signal-to-noise ratio in the data.

The computational saving provided by source encoding is measured by the speed-up S given by

$$S = \left(1 - \frac{DP_s}{DP_d}\right) \cdot 100\%, \tag{3.40}$$

where DP_d denotes the number of direct problems when using the full set of sources independently and DP_s the number of direct problems with source encoding.

5.1 Synthetic example

The BP-2004 salt velocity model is a benchmark for seismic imaging whose key attribute is that it has a considerable difference in the velocity of P waves between the water ($\approx 1500m/s$) and the salt ($\approx 4899m/s$) (Figure 3.5a). This sharp contrast generates high-amplitude primary reflection arrivals from the salt as well as energetic multiples between the salt and the free surface. This makes the recovery the sub-salt structures difficult because the sharp velocity contrast on top and on bottom of the salt hampers the transmission of a significant amount of seismic energy below salt and the information in the seismograms that constrains these parts of the model can be overprinted by high-amplitude multiples. We consider a limited target of the BP-2004 velocity model of horizontal and vertical dimensions 6.2×4.2 km, respectively. The sources and receivers are deployed all along the surface at $25m$ in depth below the water level. There are 62 sources and 248 receivers with a horizontal spacing of 100 m and 25m, respectively. The initial velocity model is a smoothed version of the true model (Figure 3.5b). The velocity in the water layer is kept unchanged during inversion (we set the gradient to zero), since we do not want to update the velocity in this area. We use two frequency groups without overlap, with a frequency interval of 1 Hz: [1, 2, 3, 4] Hz, [5, 6, 7, 8, 9, 10] Hz. For all of the tests, we fix the maximum number of direct problems to 10^5.

5.1.a *Illustration of the problem : Convergence rates of stochastic versus deterministic algorithms*

The convergence rate of deterministic gradient-based algorithms depends on the analytical properties of the misfit function, such as degree of convexity and smoothness. For a simple gradient algorithm, when the misfit function is strongly convex and smooth, it can attain a linear convergence rate $O(1/I)$ and can be even improved with accelerated gradient algorithms to $O(1/I^2)$ (Boyd and Vandenberghe, 2004). However, the convergence rate of stochastic gradient algorithms is $O(1/\sqrt{I})$ and Nemirovskiĉi and ĕIĕUdin (1983) showed it can not be easily

improved. To the present, there are no general theoretical proofs for the convergence for convex second-order stochastic optimization methods (Bottou and Le Cun, 2005; Bottou and Bousquet, 2011). Under certain conditions on how the approximate Hessian converges towards the Hessian, a proof of convergence for second-order stochastic methods can be obtained (Bottou and Le Cun, 2005). In particular, l-BFGS does not satisfy the conditions on the Hessian and thus there is no convergence proof for this stochastic optimization algorithm (Schraudolph et al., 2007). Moreover, even in the case where the conditions on the convergence on the approximate Hessian are satisfied and second-order stochastic methods converge, the convergence rate is not improved and only the constants bounding the errors are decreased (Bottou and Bousquet, 2011). Nonetheless, Schraudolph et al. (2007) implemented a stochastic l-BFGS algorithm and showed that it outperforms standard stochastic gradient algorithms for convex functions for some large scale applications.

These results are well established in the machine-learning community which mainly treat convex misfit functions. However, it is not clear how these results extend for large-scale non-convex optimization problems. We perform a numerical test in the FWI context to understand the convergence properties of stochastic versus deterministic optimization algorithms. Using the BP -2004 salt model and without noise in the data, we solve the inverse problem using only the first frequency group ($1Hz - 4Hz$). We show the results with l-BFGS and l-BFGS$_r$ as optimization algorithms. The only criterion to stop the inversion is when it reaches the maximum number of direct problems (10^5), or when the line search fails. The misfit function versus the iteration number are plotted in Figure 3.6a. The deterministic algorithm (solid line) has a linear or sub-linear convergence. However, we can see that the convergence rate starts to decrease when $\log_{10}\phi < 4$. We systematically observed that, as the inversion with full set of sources starts to reach a minimum (we do not know how to differentiate between global or local), the convergence rates of all optimization methods deteriorate. The inversion with source encoding (dashed line) shows the well expected asymptotic convergence rate of $1/\sqrt{I}$. We thus numerically obtain results that agree with the theoretic estimates of Bottou and Bousquet (2011), suggesting that the Hessian information indeed does not change the asymptotic convergence rate. The computational costs shown in Figure 3.6b indicate the region where a speed-up is possible. Although the cost of stochastic methods is less per iteration, the difference in convergence rates creates an intersection between the two curves thus bounding the region of potential computational gain. We should suspect that the higher the convergence rate of the deterministic method, the smaller the room for computational gain because the convergence rate of stochastic methods is asymptotically $O(1/\sqrt{I})$.

The result in Figure 3.6b, indicates that the potential gain in computational cost is related to the iteration number where the optimization is terminated. If the inversion is stopped somewhere in the region highlighted by the box in Figure 3.6b, the stochastic methods will solve less direct problems than deterministic ones to attain a desired value of misfit, and thus a speed-up will be attained. Outside the bounded region, deterministic methods are more efficient. Therefore, in the numerical experiments that follow, we predefine stopping criteria that allow to attain computational gain, while leading to a sufficient accuracy of the subsurface model. We now proceed to compare the convergence rates using various optimization algorithms. We find that, even though asymptotically the convergence rate is always $O(1/\sqrt{I})$, in the early part of the inversion, stochastic second-order optimization algorithms do indeed outperform the standard stochastic gradient descent, as was shown numerically by Schraudolph et al. (2007) for the convex case.

5.1.b *Synthetic data without noise*

For noise-free data, we expect the misfit function to have less local minima, and therefore all methods should converge toward a similar solution after a sufficient number of iterations. We set the relative decrease of the misfit function ϵ_1 to be equal to 0.1% and 1% for the first and second frequency groups, respectively. The second criterion of iteration ϵ_2 was set to 1%, but was never triggered during this test. For all tests we use l-BFGS with a memory of $l = 5$. The maximum number of internal iterations for the truncated Newton methods were set to be $N_{CG} = 30$. The value of the damping parameter β in the preconditioner is set to 10^{-2} and the hyper-parameters are set to $\lambda_x = \lambda_z = 10^{-8}$. All these values are outlined in Table 3.1. For source encoding, we assemble the sources with random coefficients following a Gaussian distribution (3.38), and create three super sources ($K = 3$). We do not use a fewer number of super sources, because in the upcoming experiment, the inversion fails to converge with $K = 1$ when we add noise to the data. Therefore, we perform all our numerical experiments with three super sources to share the maximum amount of parameters with and without noise. We use the same values for β, λ_x and λ_z than those used with all the sources processed independently and l-BFGS is restarted every five iterations (Table 3.1).

Unless otherwise mentioned, the four optimization methods applied with this experimental set-up attain the same final value of the misfit function, whether source encoding is used or not.

Convergence rate

We first compare the convergence rate (measured by the number of iterations required to attain a given misfit function reduction) of the four optimization methods when all the sources are processed independently (i.e., without source encoding) (Figure 3.7(a-b)). As expected, the truncated Newton methods have the highest convergence rate, followed by l-BFGS, while the convergence rate of nl-CG is significantly slower. Among the truncated Newton methods, GN outperforms FN, probably because the GN approximation leads to a positive definite Hessian, while the additional second-order term in the FN Hessian approximation may not be. This may cause an earlier termination of the linear CG iterations during the resolution of the Newton system as the positivity condition is violated (3.19).

The convergence curves with source encoding (Figure 3.7(c-d)) shows a similar trend than those obtained when no source encoding is used (Figure 3.7(a-b)): the truncated Newton methods still have the highest convergence rate, followed by $l-$BFGS and SD. However, the difference in convergence rates between the different optimization methods is clearly less pronounced when using source encoding, in particular for the inversion of the second frequency group. We conclude from this statement that the Newton methods are more penalized than the steepest-descent method by the source encoding method and this suggests that the action of the Hessian in the Newton methods is hampered by the cross-talk noise injected in the gradient of the misfit function.

The accuracy of the final velocity models obtained for all methods is similar whether source encoding is used or not. Therefore, we only show the final FWI model obtained with GN and its error when all the sources are processed independently and when source encoding is used (Figure 3.8). We only show the results using the normal distribution (3.37) for the random variables, because there are no visible difference with the models obtained with the other distributions (3.38) and (3.39). The small body in the sedimentary cover (x,z)=(5.5km,1km) is well reconstructed as well as the contours of the salt body. The sub-salt structures are well identified in particular the small body at (x,z)=(1.7km,2.2km). The focusing of the deep reflectors could have been improved by allowing more iterations (i.e., by using a smaller value of ϵ_1). Qualitative comparison between Figure 3.8(a-b) and Figure 3.8(c-d) shows that the accuracy of the final FWI models inferred from the GN optimization method was not significantly

hampered by the source encoding and this is supported by the model errors outlined in Table 3.2.

Computational efficiency and speed-up

Now that we check that all the optimization methods converge to the same solution, we can compare on the one hand the computational efficiency of each optimization method and on the other hand the speed-up provided by the source encoding method for each of these methods. We compare the reduction of the misfit function as a function of the number of direct problems for the two frequency groups when all the sources are processed independently and when source encoding is used in Figure 3.9. This comparison is shown separately for each optimization method for sake of clarity. From top to bottom in Figure 3.9, comparison between the number of direct problems performed by each optimization method to reach the desired value of the misfit function informs us about the computational efficiency of each optimization method in an absolute sense. This comparison can be performed when all the sources are processed independently (Figure 3.9, solid lines) or when source encoding is used (Figure 3.9, dash lines). Second, in each panel of Figure 3.9, the difference between the number of direct problems with and without source encoding for a given value of the misfit function value represents the computational gain provided by the source encoding method for one optimization method. Normalization of this difference by the value of the misfit function obtained without encoding gives the speed-up (3.40) (Figure 3.10).

Although we have shown that the truncated Newton methods have the highest convergence rate, l-BFGS has the lowest computational cost, followed closely by GN, whether source encoding is used or not. This results because l-BFGS performs a more limited number of direct problem per non-linear iteration than truncated Newton methods. The $nl-$CG/SD method has the highest computational cost due to its poor convergence rate compared to the quasi-Newton and truncated Newton methods. Although $nl-$CG/SD has the highest computational cost, it shows the best speed-up (Figure 3.10(a-b) and Table 3.2)). This reflects the fact that, among all the optimization methods, it is the one whose convergence rate has been less affected by the cross-talk noise. This might result because SD is the only optimization method that does not account for the Hessian, whose action is affected by the cross-talk noise introduced by source encoding. Notice that, for the realization shown in Figure 3.9, the GN optimization methods does not show any speed-up for the second frequency group.

Independent of the optimization method, a general conclusion is that the speed-up decreases as the misfit function gets closer to the minimum, i.e., as the convergence rate of the optimization slows down to reach a plateau (Figure 3.10(a-b)). Therefore, a trade-off must be found between the speed-up provided by source encoding and the quality of the final FWI model.

5.1.c *Synthetic data with noise*

We now add 25 % of mean zero additive Gaussian noise to the data. We use $N_{CG} = 5$, $\beta = 10^{-2}$ and $\lambda_x = \lambda_z = 10^{-6}$, as described in Table 3.3. When compared to the noise-free experiment, we increase the value of they hyper-parameters and we reduce the maximum number of iterations performed by the linear CG algorithm during the resolution of the Newton system. This was required to make the truncated Newton methods to converge when source encoding is used. As the regularization damps the action of the Hessian (3.33), we can anticipate that this more aggressive regularization will penalize the convergence rate of the Newton methods. We decrease ϵ_1 to 25% and 70% for the frequency groups 1 and 2, respectively, as the relative reduction of the misfit function is expected to be much less compared to the previous case because of the noise in the data. Regarding condition (3.36), we stop the inversion if the average of the misfit function over the previous 30 iterations has not changed more than 10% ($\epsilon_2 = 0.1$). For this test, the ϵ_1 criterion was generally triggered during the inversion of the first frequency group, while the ϵ_2 criterion was generally triggered during the inversion of the second frequency group

before the misfit function reaches a value corresponding to ϵ_1. This statement reflects a slower convergence rate during the inversion of the second frequency group. All the tuning parameters are outlined in Table 3.1 and can be compared with those used for the experiment performed without noise. When we apply source encoding, we use three encoded sources ($K = 3$) and the same values for the free parameters and for the stopping criteria of non-linear iterations.

Convergence rate

A key difference with the noise free experiment is that the relative reduction of the misfit function is much smaller because the data noise level is higher. Once the noise level has been reached, the FWI may continue to significantly update the model without a perceptible decrease in the data misfit. This occurs because improvements in the model have a small weight in the data misfit compared to the high noise energy. As a consequence we do not immediately terminate the inversion when the misfit function is flat, but rather proceed to measure the relative change of the average value over a certain number of iterations, giving rise to stopping criterion 2 controlled with ϵ_2. Finding a suitable value of ϵ_2 that provides the best-trade-off between computational efficiency and quality of the subsurface model is not obvious because the risk is either to stop the iterations too early (before a sufficient accuracy of the subsurface model is reached) or too late (iterations do not lead to significant update of the subsurface model). The speed-up estimation is always sensitive to the stopping criteria but in this context it is even more critical.

This trend in the convergence curves is illustrated in Figure 3.11, which shows the slowly decreasing misfit functions as a function of the number of iterations for the four optimization methods with and without source encoding. These curves can be compared with those obtained when the data does not contain noise (Figure 3.7; notice that the horizontal and vertical scales in Figure 3.11 spans over a much narrow range of misfit function values than in Figure 3.7). As for the noise-free case, the truncated Newton methods reach the stopping criterion of iteration with a smaller number of iteration than $nl-$CG and to a lesser extent to $l-$BFGS when all the sources are processed independently (Figure 3.11(a-b)) and when source encoding is used (Figure 3.11(c-d)). However, the difference between the convergence speed of the different optimization methods is less pronounced than for the noise-free case and this leveling down of the convergence speed is still accentuated when source encoding is used. This leveling down of the performances results because the convergence of the truncated Newton methods now reaches a plateau before satisfying the stopping criterion of iterations when a significant amount of noise is added to the data, unlike in the case of noise-free data.

The final FWI velocity models inferred from the GN optimization with and without source encoding are similar and compare well with the subsurface models inferred from the noise-free experiment (compare Figures 3.8 and 3.12).

Computational efficiency and speed-up

The reduction of the misfit function as a function of the number of direct problems, when all the sources are processed independently and when source encoding is used, is shown in Figure 3.13 for each optimization method. These curves can be compared with those inferred from noise-free data (Figure 3.9). The speed-up of each optimization method as a function of the misfit-function value is synthesized in Figure 3.10(c-d) for noisy data.

The $nl-$CG/SD method still has the best speed-up. Moreover, the difference with the other optimization methods has even increased compared to the case of noise-free data (see also Table 3.2). This results because, as already mentioned, the convergence rate of the truncated Newton methods was affected more than the $nl-$CG/SD optimization by both the Gaussian noise and the cross-talk noise. The effect of noise on the speed-up is clearly illustrated by the comparison of the speed-up curves inferred from noise-free and noisy data (Figure 3.10). In the case of

noise-free data, the speed-up decreases slowly as the value of the misfit function decreases from right to left in the figure, while it decreases much more rapidly in the case of noisy data as the convergence curves start reaching a plateau. As the slope of the speed-up curves increases near the smallest values of the misfit function in Figure 3.10(c-d), differences between the speed-up of each optimization methods are emphasized.

Statistical stability

We perform 50 independent realizations of FWI with source encoding to test the statistical stability of source encoding, i.e., to test whether the final FWI velocity model changes dramatically from one realization to the next. The fifty convergence curves of the misfit function for the second frequency group are plotted in Figure 3.14. In general all realizations tend towards the same minimum, for all the optimization methods, although $l-$BFGS shows a less robust behavior (Figure 3.14b). To quantify this variability, we compute the variance of the final models for each optimization method,

$$\mathrm{Var}(m) = \frac{1}{MC}\sum_{j=1}^{MC}(m_j - \bar{m})^2 \tag{3.41}$$

$$\bar{m} = \frac{1}{MC}\sum_{j=1}^{MC} m_j, \tag{3.42}$$

where MC is the number of realizations (in this case $MC = 50$), m_j is the final velocity model of realization j. The variance shows that there is a maximum variability for the l-BFGS$_r$ method, but confirms that the inversion of noisy data with source encoding methods is statistically stable (Figure 3.15). The maximum of the variance is shown on top of the salt body near the end of the model where a more limited illumination is available. GN has the smallest variance (Figure 3.15).

5.1.d *Stochastic gradient*

Implementation of source encoding in a steepest descent method mimics a stochastic gradient algorithm (Robbins and Monro, 1951) that is widely used in many applications. In machine learning, for example, a very large set of training examples are available and instead of using all of them at once to compute the gradient, each training example is individually employed to find a descent direction. Although the descent direction inferred from one training example is not as accurate as the one inferred from all the training examples at once, it has been shown to be more efficient to do a sweep using one at a time. This way, many low cost inaccurate iterations are performed that converge to the global minimum. Strictly speaking, however, the series of the step lengths $\{\alpha_n\}$ must satisfy the following conditions for the stochastic gradient algorithm to converge

$$\sum_n \alpha_n = \infty \tag{3.43}$$

$$\sum_n \alpha_n^2 < \infty. \tag{3.44}$$

These conditions are imposed so that the step length does not decrease too quickly because it will be trapped in a premature local minimum, but it must eventually tend to zero otherwise the algorithm will oscillate around the global minimum and never spiral into it. A common choice that satisfies these conditions is $\alpha_n = (1/n)^a$ or $\alpha_n = \alpha_0\left(\frac{\tau}{\tau+n}\right)^a$, where η_0, a and τ are some

tuning parameters. This algorithm is only valid for steepest descent method, although some limited theoretical extensions of the stochastic gradient to second-order stochastic optimization have been proposed (Bottou and Le Cun, 2005) and implemented (Schraudolph et al., 2007; Roux and Fitzgibbon, 2010) in the machine learning community.

In source encoding, the randomness is not generated by choosing a different source at each iteration, which would be the analogous of the machine-learning setting, but rather the randomness is introduced by changing the weights in the linear combination that create the super sources. In addition, we do not satisfy the conditions imposed on the step length, equations 3.43 and 3.44, but instead we do a line search to satisfy the strong Wolfe conditions. Indeed, we tried to implement the stochastic gradient algorithm in an equivalent manner, choosing only one source randomly per iteration and/or, choosing a series of step lengths to satisfy conditions (3.43) and (3.44) in order to honor the stochastic gradient algorithm as closely as possible. However, neither of these approaches showed an acceptable performance. Moreover, since there are several ways to make the step length satisfy the required conditions, searching for the adequate tuning parameters (a, α_0, τ) is computationally expensive because it must be done in a trial and error fashion. Therefore, it is much faster to do a line search at each iteration, even if it leads us further from the classical stochastic gradient algorithm.

In addition to the benefits in computational cost, stochastic gradient has advantages inherent to stochastic optimization techniques, such as simulated annealing, that aid to find a global minimum of a misfit function that may possess several local minima. Bottou (1991) highlights the analogy between the temperature in simulated annealing and the step length (referred to as learning rate) in the stochastic gradient method. In the stochastic gradient method, we allow the misfit function to increase for when we change the encoding (Figure 3.14), allowing us to explore regions of the model space that would never be accessible with deterministic methods. This relaxation of the explored model space, which is illustrated by the small fluctuations of the convergence curves in Figure 3.14c, allows one to overcome small local minima and we confirmed this advantage through the following synthetic numerical test.

We apply FWI on the noise-free data without the multi-scale strategy by inverting nine frequencies ranging from $1Hz$ to $9Hz$ in one group, such as to render the misfit function highly non-convex with many local minima. Using the initial model shown in Figure 3.5b, the inversion with and without source encoding converge to equivalent final solutions (Figure 3.16(a-b)). When we degrade the initial model to that depicted in Figure 3.5c, the final FWI velocity model obtained when all the sources are processed independently is less accurate than the one inferred from the stochastic optimization (Figure 3.16(c-d)). Therefore, we conclude that, with source encoding, we may not only reduce the computational cost but we may also steer the solution towards a better local minimum thanks to a broader exploration of the model space.

5.1.e *Summary of the BP-2004 salt experiment*

We draw five main conclusions from the BP-2004 experiment: [1] the speed-up tends to decrease as the misfit function reaches a value close to the global minimum. Therefore, the best trade-off between computational efficiency and quality of the subsurface model must be found through a judicious stopping criterion of iterations (Figure 3.6). [2] When noise is added to the data, the speed-up with respect to the value of the misfit function decreases more rapidly, and the convergence of the optimization near the global minimum of the misfit function slows down (Figure 3.10). [3] The speed-up of the optimization methods that do not account for the Hessian in the optimization, i.e., $nl-$GG/SD method, is higher than the one achieved with Newton-based optimization. We interpret this result as the penalizing effect of noise on the action of the

Hessian. Therefore, the performance of all the optimization methods tend to be leveled down, as noise is added to the data and the action of the Hessian is damped. [4] Indeed, the speed-up is a relative measure in the sense that it gives the computational gain provided by source encoding for one optimization method. In absolute, the truncated Newton methods have the best convergence rate with and without source encoding and $l-$BFGS is the fastest method with and without source encoding. However, the robustness of $l-$BFGS to noise needs be assessed in more complex setting as we shall see with an application on real data presented in the following of this study. [5] The randomization underlying the source encoding method can help to steer the inversion toward the global minimum thanks to a broader exploration of the model space around some minima of the misfit function.

5.2 Real data example

To validate that source encoding techniques have a true interest in real data applications, we use a *2D* ocean-bottom-cable (OBC) data set from the Valhall oil field in the North Sea. This data set contains 320 sources that are located 5m beneath the water level, with a spacing of 50m, and are recorded by 210 hydrophone receivers on the sea bottom at 68m below the water level, also with a 50m spacing. The dimensions of the subsurface model are 16×4 km. Several studies have already been conducted using this data set (Prieux et al., 2011, 2013a; Gholami et al., 2013a). The subsurface model is mainly characterized by soft quaternary sediments below the sea level, low-velocity gas layers between 1.5 km and 2.5 km in depth above the reservoir which delineates a sharp positive velocity contrast. These structures are highlighted in a reverse time migrated image computed in a background subsurface model that will be used as initial model for FWI in the following of the present study (Figure 3.17). Anisotropy, which is significant and can reach a maximum value of 15 %, is taken into account in the seismic modeling performed during FWI. The initial models for the vertical wavespeed and the Thomsen parameters δ and ϵ were developed by reflection travel-time tomography (courtesy of BP) and are shown in Gholami et al. (2013a). The background density model is inferred from the initial vertical wave-speed model by Gardner's law and the quality factor is fixed at a constant value of 200. We perform a mono-parameter FWI for the vertical velocity v_{P0} keeping the Thomsen's parameters δ and ϵ, the density and the quality factor fixed. The relevance of the VTI parametrization $(v_{P0}, \delta, \epsilon)$ for mono-parameter FWI is discussed in Gholami et al. (2013b), Gholami et al. (2013a) and Operto et al. (2013).

Experimental set-up

We use four overlapping frequency groups, ranging from 3.5Hz to 6Hz: [3.5, 3.78, 4], [4, 4.3, 4.76], [4.76, 5, 5.25], [5.25, 5.6, 6] Hz. We didn't see significant reduction in the misfit function and no improvement in the velocity models for higher frequencies, and the data is too noisy for inversion at lower frequencies. For each frequency group, the stopping criterion of non-linear iterations is controlled by the relative reduction of the misfit function ($\epsilon_1 = 0.7$ for each frequency group) with and without source encoding. When each source is processed independently, we also set the maximum number of non-linear iterations to 20. For the truncated Newton methods, we use $N_{CG} = 3$ because the real data is considered to be noisy (Métivier et al., 2013b).
When all sources are used independently, the free parameters for all optimization methods are $\beta = 10^{-2}, \lambda_x = 10^{-3}$ and $\lambda_z = 2.5\,10^{-4}$. When source encoding is used, we succeeded in making the inversion to converge with one super source (K=1). The horizontal regularization weight λ_x is increased to 10^{-2} (Table 3.4). We use the spatial reciprocity of Green functions to process receivers as sources during FWI. Therefore, 210 sources are stacked to form a super source. This number of sources is significantly higher than the one used during the BP experiment (62) and this difference must be taken into account in the speed-up analysis.

Convergence of FWI

We first compare the convergence of the different optimization methods without source encoding for the four frequency groups (Figure 3.18(a-d)). For the first two frequency groups, the two truncated Newton methods converge to the same misfit function value, which suggests that the second-order term in the truncated FN is smaller than the regularization term and thus has no significant effect in the inversion. For the third and fourth frequency groups, however, the two methods follow different optimization paths and GN attains a lowest misfit function value for the fourth frequency group. For this particular choice of free parameters, l-BFGS does not decrease the misfit function as much as the other optimization methods. In particular, it fails to update the model during the second frequency group. The $nl-$CG method performs a misfit function reduction close to the one achieved by the truncated Newton methods during the inversion of the first two frequency groups, unlike for the third and fourth frequency groups for which the convergence rate of the $nl-$CG method is poorer and the minimum of the misfit function that was reached is higher. Overall for all frequency groups, l-BFGS performs the fewest number of iterations but with a poorer convergence level. We conclude from these results that the truncated Newton methods clearly provide the most robust direction of descent relative to quasi-Newton and conjugate gradient methods.

The final velocity models that were obtained with each optimization method at the end of the fourth frequency group are shown in Figure 3.19(a-d). Comparison between a sonic log at 9.5 km in distance and the corresponding profiles of each FWI model is shown in Figure 3.20(a-d). The FWI models obtained with the truncated Newton methods clearly provide the best trade-off between signal-to-noise ratio and resolution. For example, shallow artefacts near the end of the model inferred from the $nl-$CG method (Figure 3.19a) are not present in the truncated Newton models (Figure 3.19(c-d)). Moreover, the deep reflector below the reservoir level shown in Figure 3.17 is far less contrasted in the $l-$BFGS model (Figures 3.19b) than in the truncated-Newton models (Figures 3.19(c-d)), although the geometry of this reflector seems well reconstructed in the $l-$BFGS model. The weaker amplitudes of the velocity perturbations retrieved by $l-$BFGS may result from the more limited number of iterations performed by this optimization method.

When source encoding is used, a similar hierarchy among the different optimization methods is shown with superior results achieved by the truncated Newton methods both in terms of convergence rate and convergence level (Figure 3.18(e-h)). The quasi-Newton l-BFGS$_r$ method failed to converge during the first frequency group for the realization shown in Figure 3.18(e).

The final velocity models obtained with source encoding are shown in Figure 3.19(e-h). The FWI model obtained with $nl-$CG shows significant artefacts along the shallow reflector at around 0.6 km in depth. The deep reflector is also reconstructed with weak amplitudes in the $nl-$SD model (Figure 3.19e). The footprint of the cross-talk noise is clearly visible in the shallow part of the $l-$BFGS$_r$ model, while the deep reflector is better reconstructed relative to the one obtained without source encoding (compare Figure 3.19b and 3.19f). The truncated Newton methods show a more robust behavior with respect to source encoding in the sense that the velocity models inferred from these methods with and without source encoding are quite consistent (compare panels (c-d) and (g-h) in Figure 3.19).

Computational efficiency

The convergence curves as a function of the number of direct problems are shown for the four optimization methods when source encoding is used or not in Figure 3.21. The truncated Newton methods are around two times more expensive than $nl-$CG when all the sources are processed independently, while the cost of truncated Newton methods and SD is similar when source encoding is used (Table 3.5). This leads to a higher speed-up of the truncated Newton methods

relative to the $nl-$CG/SD method (around 96% against 92%). The speed-up is quite significant and represents almost one order of magnitude in terms of computational saving. However, this speed-up is highly sensitive to the choice of the stopping criterion of iteration ϵ_1 (Figure 3.6). If a value of 60% instead of 70% would have been chosen, almost no speed-up would have been shown because we would have let the inversion to perform many iterations without significant decrease of the misfit function whether source encoding is used or not.

Quality control: statistical stability

We perform 50 independent realizations of the FWI with source encoding to test the statistical stability of inversion. The convergence curves plotted as a function of the number of direct problems for the fourth frequency group confirm that the truncated Newton methods are the most robust relative to SD and $l-$BFGS$_r$ optimization methods (Figure 3.22). Accordingly, the velocity models built by the truncated Newton models have the smaller variance (Figure 3.23). The higher values are shown in the shallow part near the ends of the receiver cable where inversion has more degrees of freedom to perturb the subsurface model (Figure 3.23(c-d)). Significant values of the variance are also shown at the reservoir level at around 2.5 km in depth near the ends of the reflector segment that was imaged by FWI, still in relation with a more limited illumination. It is worth noting that the variance is almost zero in the bottom right and bottom left of the model where a strong deficit of illumination exists. This is consistent with the fact that the (damped) regularized Hessian prevents the updating of the subsurface model where the sensitivity of the inversion to the information contained in the data is below some given threshold. The variance in the SD model reaches the highest values in the shallow part, which is consistent with the shallow artefacts highlighted in Figure 3.19e. The variance of the $l-$BFGS$_r$ realizations reflects the imprint of the cross-talk noise in the shallow part of the model already highlighted in Figure 3.19f as well as shallow artefacts near the ends of the cable (Figure 3.19b). Moreover, we would like to confirm that, through our numerical tests, the conclusions regarding the convergence rates, costs and variance are independent of the distribution chosen for the random variables, showing a similar behavior for all of the distributions (3.37) and (3.39).

Quality control: reverse time migration

We apply anisotropic reverse time migration to the Valhall data using the FWI models inferred from the FN optimization method with and without source encoding as background models (Figure 3.24). The experimental set-up to perform reverse time migration is outlined in Prieux et al. (2011). The two migrated images (Figure 3.24) can be compared with the one computed in the initial model (Figure 3.17). We superimposed in transparency on each migrated image the background velocity model that was used to perform migration to check the consistency between the reflectors mapped by migration and the velocity variations built by FWI. The accuracy of the migrated images can be assessed by the flatness of the reflectors in the common image gathers (CIGs) (Figure 3.25). As it is highlighted in Prieux et al. (2011), it is quite challenging to improve the migrated images computed in the background model built by reflection travel-time tomography, because reflection travel-time tomography is designed to optimally focus reflection energy. However, FWI has improved the imaging of the reflectors in the shallow part (down to 600 m in depth) where reflection travel-time tomography can encounter difficulty to pick travel-times. This is highlighted by an improved focusing of the shallow reflectors between 0.4 and 0.6 km in depth in Figure 3.24. The highest resolution of the FWI velocity models relative to the travel-time tomography model is also highlighted by the closer correlation between the reflectors mapped by migration and the sharper velocity variations imaged by FWI. This improvement was also shown in some close-up of the CIGs centred on the shallow reflectors in (Prieux et al., 2011, their figure 10). Aside the shallow reflectors, other local improvements of

the migrated images inferred from the FWI background models are highlighted in Figure 3.24, gray and black arrows) and in the CIGs (Figure 3.25, shaded area). The most obvious one is that the base of the reservoir is more continuous in the FWI-based migrated images than in the tomography-based migrated image (compare Figures 3.17 and 3.24, gray arrows). The deep reflector below the reservoir is also more continuous in particular where this reflector has more significant dip (compare Figures 3.17 and 3.24, black arrows). We do not see any imprint of the cross-talk noise in the RTM image computed in the FWI model obtained with the source encoding method (Figure 3.24b). This RTM image generally shows more continuous and focused reflectors, in particular at the reservoir level, than the RTM image computed in the FWI model obtained without source encoding (compare the two panels in Figure 3.24). This probably results because we use a stronger horizontal regularization weight during FWI when source encoding is used. This statement reflects the trade-off between resolution and error and how the errors that are accumulated over the non-linear iterations of FWI models are mapped in the migrated image.

6 Estimation of the variance with source encoding, without noise.

In this section, we suppose that there is no noise in the data, so the only randomness appearing in the minimization of $\tilde{\phi}(\tilde{u};m)$ (equation 3.26) comes from the coefficients of the encoding $\alpha_i \in \mathbb{C}$. Recall that the choice of the coefficients for encoding the signals is arbitrary up to satisfying one condition,

$$\mathbb{E}[\alpha_i^* \alpha_j^*] = \delta_{i,j}. \tag{3.45}$$

In Appendix 6, the expected values of the misfit function and the gradient are expressed. The role of condition 3.45 is that, when it holds,

$$\mathbb{E}_\alpha \left[\nabla_m \tilde{\phi}\right] = \nabla_m \phi, \tag{3.46}$$

where just for this time we have used the notation $\mathbb{E}_\alpha$ to denote explicitly that the average is taken exclusively with respect to the randomness in the coefficients. This is an important property because the minimization problem with source encoding can be seen as a minimization procedure using a "noisy" gradient with no bias. This is exactly the case generally known as stochastic descent algorithms, and we are able to use the existing results.

Concentrating now only on gradient descent algorithms, it is then natural to wonder which distribution would give better results. We answer this question by considering the variance of the encoded gradient term (3.27). Analysis of the variance have been performed for noise blending experiments in the frequency domain (De Hoop et al., 2012). Intuitively, reducing the variance of the obtained gradients would give faster convergences, as we are able to control deviations from the average. In fact the control of variance terms is a key point in most convergence proofs for the gradient algorithm as those in Laruelle and Pagés (2012), and some works on using variance reduction techniques have been proposed in Wang et al. (2013a). To simplify, we will *suppose that the mixing coefficients are real*, $\alpha_i \in \mathbb{R}$.

Let us define D by

$$D(x,m,\omega) := \frac{\partial A}{\partial m}(x,m,\omega). \tag{3.47}$$

From equation (3.27), it follows that [1]

1. $V(x) = \mathbb{E}[x^2] - \mathbb{E}[x]^2$.

$$
\begin{aligned}
V_{\nabla\tilde{\phi}} &= \mathbb{E}_\alpha\left[(\nabla_m\tilde{\phi})(\nabla_m\tilde{\phi})^\dagger\right] - \left(\mathbb{E}_\alpha[\nabla_m\tilde{\phi}]\right)\left(\mathbb{E}_\alpha[\nabla_m\tilde{\phi}]\right)^\dagger \\
&= \sum_{i,j,i',j'}^{N_s} \mathbb{E}_\alpha\left[\alpha_i\alpha_{j'}\alpha_{i'}\alpha_j\right] \Re\left[(Du_i)^\dagger \lambda_j\right] \Re\left[\lambda_{j'}^\dagger (Du_{i'})\right] \\
&\quad - \sum_{i,j'}^{N_s} \Re\left[(Du_i)^\dagger \lambda_i\right] \Re\left[\lambda_j^\dagger (Du_j)\right] \\
&= \sum_{i,j\neq i,i',j'\neq i'}^{N_s} \mathbb{E}_\alpha\left[\alpha_i\alpha_{j'}\alpha_{i'}\alpha_j\right] \Re\left[(Du_i)^\dagger \lambda_j\right] \Re\left[\lambda_{j'}^\dagger (Du_{i'})\right] \\
&\quad + \sum_{i,i',j'\neq i'}^{N_s} \mathbb{E}_\alpha\left[\alpha_i^2\alpha_{i'}\alpha_{j'}\right] \Re\left[(Du_i)^\dagger \lambda_i\right] \Re\left[\lambda_{j'}^\dagger (Du_{i'})\right] \\
&\quad + \sum_{i,i',j'\neq i'}^{N_s} \mathbb{E}_\alpha\left[\alpha_i^2\alpha_{i'}\alpha_{j'}\right] \Re\left[(Du_{i'})^\dagger \lambda_{j'}\right] \Re\left[\lambda_i^\dagger (Du_i)\right] \\
&\quad + \sum_{i,j}^{N_s} \left[\mathbb{E}_\alpha[\alpha_i^2\alpha_j^2] - 1\right] \Re\left[(Du_i)^\dagger \lambda_i\right] \Re\left[\lambda_j^\dagger (Du_j)\right] \qquad (3.48)
\end{aligned}
$$

Let us divide our analysis in some different cases:

– **Case 1: The α_i are i.i.d. (Using all sources in the super source)**

Condition (3.45) and independence imply $\mathbb{E}_\alpha[\alpha_i] = 0$ for all $i = 1, \ldots, N_s$. Clearly,

$$\mathbb{E}_\alpha[\alpha_i^2\alpha_j^2] = 1 \text{ for } i \neq j.$$

Hence, for i, j, i', j' such that $i \neq j$ and $i' \neq j'$, we have

$$\mathbb{E}_\alpha[\alpha_i\alpha_j\alpha_{i'}\alpha_{j'}] = \begin{cases} 1 & \text{if } i = i' \text{ and } j = j' \\ 1 & \text{if } i = j' \text{ and } j = i' \\ 0 & \text{otherwise} \end{cases}.$$

On the other hand, if $i' \neq j'$, we prove using independence that

$$\mathbb{E}_\alpha[\alpha_i^2\alpha_{i'}\alpha_{j'}] = 0.$$

Therefore, equation (3.48) becomes

$$
\begin{aligned}
V_{iid} &= \sum_{i,j\neq i}^{N_s} \Re\left[(Du_i)^\dagger \lambda_j\right] \left(\Re\left[\lambda_j^\dagger (Du_i)\right] + \Re\left[\lambda_i^\dagger (Du_j)\right]\right) \\
&\quad + \sum_{i}^{N_s} \left[\mathbb{E}_\alpha[\alpha_i^4] - 1\right] \Re\left[(Du_i)^\dagger \lambda_i\right] \Re\left[\lambda_i^\dagger (Du_i)\right].
\end{aligned}
$$

Then, under the considered case, the smallest possible variance is obtained choosing the smallest possible value for $\mathbb{E}[\alpha_1^4]$. By Cauchy-Schwarz inequality, we know that

$$1 = \mathbb{E}_\alpha[\alpha_1^2] = \mathbb{E}_\alpha[1 \cdot \alpha_1^2] \leq \mathbb{E}_\alpha[\alpha_1^4]^{1/2},$$

so that the minimum attainable value compatible with the assumed conditions is $\mathbb{E}[\alpha_1^4] = 1$. This value, and the other conditions, can be attained when taking each α_i with values in

$\{-1, 1\}$ each with probability $1/2$. Hence, reorganizing the the expression the variance is simply

$$\boxed{V_{iid^*} = \sum_{i=1}^{N_s}\sum_{j>i}^{N_s}\left(\Re\left[(Du_i)^\dagger \lambda_j + (Du_j)^\dagger \lambda_i\right]\right)\left(\Re\left[\lambda_j^\dagger (Du_i) + \lambda_i^\dagger (Du_j)\right]\right).} \tag{3.49}$$

To motivate our following study test, let us comment on the optimal distribution under the i.i.d hypothesis: the effect of having, grossly, half the coefficients of one sign and half on the other, may be understood as a form of variance reduction technique. As a consequence, we end up explaining the variation of the random variable exclusively by the cross-talk between the forward wavefield corresponding to one source and the back propagating wavefield of a different source (and represented by the terms like $\lambda_j^\dagger Du_i$). If these coefficients are small, then we have managed to reduce the variance. Moreover, the obtained distribution would be quite effective to calculate.
The story is different if we suppose that the cross-talk terms are big. In this scenario, we would like to consider a different approach. This motivates our next study case, in which we avoid any cross-talk.

– **Case 2: Mutually exclusive case. (Using one source in the super source)**

Here, we draw *one source* i^* with uniform probability $1/N_s$ and fix

$$\alpha_j = \delta_{j,i^*}\sqrt{N_s}$$

.
We can easily verify that condition ec:chvarecalpha holds in this setup, as $\mathbb{E}_\alpha[\alpha_i^2] = \frac{1}{N_s}(\sqrt{N_s})^2 = 1$ and $\mathbb{E}_\alpha[\alpha_i\alpha_j] = 0$ if $i \neq j$.
Now, from equation (3.48),

$$\begin{aligned} V_{exclu} = &\sum_{i}^{N_s}\left[N_s - 1\right]\Re\left[(Du_i)^\dagger \lambda_i\right]\Re\left[\lambda_i^\dagger (Du_i)\right] \\ &- \sum_{i,j\neq i}^{N_s}\Re\left[(Du_i)^\dagger \lambda_i\right]\Re\left[\lambda_j^\dagger (Du_j)\right], \end{aligned}$$

and reorganizing one obtains,

$$\boxed{V_{exclu} = \sum_{i=1}^{N_s}\sum_{j>i}^{N_s}\left(\Re\left[(Du_i)^\dagger \lambda_i - (Du_j)^\dagger \lambda_j\right]\right)\left(\Re\left[\lambda_i^\dagger (Du_i) - \lambda_j^\dagger (Du_j)\right]\right).} \tag{3.50}$$

Note that V_{exclu} does not introduce any cross-talk noise. Instead, as we are taking only one active coefficient, the disparities in the gradient of ϕ are completely explained by the disparities of the gradients one would obtain when restricting the cost function to only one source. This is exactly what is reflected in V_{exclu}.
Intuitively, when close sub-gradients and high cross-talk are related. We could the think that the i.i.d. case and the exclusive cases are in some sense complementary. Finally we consider a mixed approach.

– **Case 3 : Combination of Case 1 and Case 2 (Using l sources in the super source)**.

The final condition we explore results from mixing the previous two approaches.

Assume at each minimization step we choose uniformly l coefficients to be chosen to be $\pm\sqrt{N_s/l}$, independently and with equal probability, and set the remaining ones to 0.
As before, let us check that condition (3.45) holds. Indeed,

$$\mathbb{E}_\alpha[\alpha_i^2] = \frac{N_s}{l}\mathbb{P}(\alpha_i \neq 0) = \frac{N_s}{l}\frac{\binom{N_s-1}{l-1}}{\binom{N_s}{l}} = 1$$

and if $i \neq j$,

$$\mathbb{E}_\alpha[\alpha_i\alpha_j] = \mathbb{E}_\alpha[\alpha_i\alpha_j|\alpha_i \neq 0, \alpha_j \neq 0]\mathbb{P}(\alpha_i \neq 0, \alpha_j \neq 0) = 0$$

With similar arguments, we find that

$$\mathbb{E}_\alpha[\alpha_i^4] = \frac{N_s^2}{l^2}\frac{\binom{N_s-1}{l-1}}{\binom{N_s}{l}} = \frac{N_s}{l},$$

$$\mathbb{E}_\alpha[\alpha_i^2\alpha_j^2] = \frac{N_s^2}{l^2}\frac{\binom{N_s-2}{l-2}}{\binom{N_s}{l}} = \frac{N_s(l-1)}{(N_s-1)l} \text{ for } i \neq j.$$

Hence, for i, j, i', j' such that $i \neq j$ and $i' \neq j'$, we have

$$\mathbb{E}_\alpha[\alpha_i\alpha_j\alpha_{i'}\alpha_{j'}] = \begin{cases} \frac{N_s(l-1)}{(N_s-1)l} & \text{if } i = i' \text{ and } j = j'; \text{or if } i = j' \text{ and } j = i' \\ 0 & \text{otherwise} \end{cases}.$$

And if $i' \neq j'$

$$\mathbb{E}_\alpha[\alpha_i^2\alpha_{i'}\alpha_{j'}] = 0.$$

Therefore, equation (3.48) becomes

$$\begin{aligned} V_{group} = &\sum_{i,j\neq i}^{N_s} \frac{N_s(l-1)}{(N_s-1)l}\Re\left[(Du_i)^\dagger \lambda_j\right]\left(\Re\left[\lambda_j^\dagger (Du_i)\right] + \Re\left[\lambda_i^\dagger (Du_j)\right]\right) \\ &+ \sum_{i}^{N_s}\left[\frac{N_s}{l} - 1\right]\Re\left[(Du_i)^\dagger \lambda_i\right]\Re\left[\lambda_i^\dagger (Du_i)\right] \\ &+ \sum_{i,j\neq i}^{N_s}\left[\frac{N_s(l-1)}{(N_s-1)l} - 1\right]\Re\left[(Du_i)^\dagger \lambda_i\right]\Re\left[\lambda_j^\dagger (Du_j)\right] \end{aligned}$$

Let

$$\boxed{\theta := \frac{N_s(l-1)}{(N_s-1)l}}$$

then, noting that $N_s/l - 1 = (N_s - 1)(1 - \theta)$ and rearranging the terms as we did for the previous cases, we find

$$\boxed{V_{group} = \theta V_{iid^*} + (1-\theta)V_{exclu}.} \tag{3.51}$$

– **Remarks on the analysis of the variance**

We seek the distribution that will minimize the variance of the expression of the encoded gradient. In a stochastic gradient descent algorithm, this would allow for a faster convergence. We assume we generate one super source. The above analysis of the variance depends on the values of λ_j and u_i at each iteration. However, some general observations can be retrieved,

- If the all the sources are used, the distribution $\{-1, 1\}$ each with probability $1/2$, minimizes the variance. The variance (3.49) is dominated by cross talk terms (correlations of the form $u_i\lambda_j$, $i \neq j$. Therefore, groups of sources that have a high correlation (for example sources illuminating the same parts of the medium) will have a high variance.
- If only one randomly chosen source is used, the variance (3.50) does not contain cross-talk terms, but is proportional to difference between gradients produced by different sources. If the gradients generated from different sources are very different (for example sources illuminating different parts of the medium) the variance will be very large.
- If l sources are randomly chosen in the super source, the resulting variance (3.51) is a linear combination of the first two cases. That is, a part of the variance will be due to the cross talk, and a part of the variance will be due to the differences in gradients.

The analysis of the variance suggests that, in absence of information about the gradients produced by each source, choosing all sources in the super source is perhaps the best strategy, as it is more likely that the variance (3.49) produced by the cross talk will be smaller than the variance (3.50) produced by the difference of gradients. However, in the knowledge of the similarity amongst gradients (because of the geometric disposition, for example) perhaps more clever strategies to create the super sources can be defined. We are still studying this.

7 Conclusion

We have applied *2D* efficient frequency-domain FWI on synthetic and real data when random source encoding is interfaced with different optimization methods in order to determine the best strategy to perform a fast FWI in a robust manner. In particular, we succeeded in combining random source encoding techniques with second-order optimization methods, despite stochastic optimization was only shown to converge for steepest descent algorithms respecting certain criteria for the step length. A careful design of several stopping criteria of iteration allows for a fair assessment of the computational efficiency of each optimization method and the speed-up provided by the source encoding method.

Without using source encoding we found that, in an ideal noise-free data scenario with frequency groups that determine an approximately convex misfit function, truncated Newton methods have the highest convergence rate, and thus require less iterations to attain a desired relative reduction of the misfit function. However, truncated Newton methods remain more computationally expensive than the quasi-Newton method l-BFGS as truncated Newton methods require additional direct problems per non-linear iteration. As noise is added to the synthetic data and more aggressive regularization is used, the action of the Hessian becomes less effective and the convergence rate of the Newton-based methods is thus degraded. This contributes to level down the convergence rate of Newton-based methods relative to steepest-descent method. All optimization methods when combined with source encoding were shown to be statistically stable, meaning that, when several independent inversions are carried out, each inversion converges to approximately the same model. However, the truncated Newton methods have a more robust behavior showing a smaller variance in the final solution. The continuous or discrete probability distribution of the random variables has no effect in the inversion, as long as the desired statistical properties are satisfied.

The gain in computational cost provided by source encoding is measured by the ratio of the number of direct problems that have to be solved with and without source encoding. During the early iterations of the inversion, the misfit function with source encoding decreases sufficiently fast to make the speed-up large. As the iterations proceed and the convergence slows down,

stochastic methods require more iterations than deterministic methods to decrease the error by the same amount and the speed-up decreases accordingly. Therefore, we showed that the speed-up strongly depends on the value of the misfit function for which iterations are stopped. To that end a suitable stopping criterion of iterations should be designed such that the best trade-off between computational efficiency and quality of the subsurface model is found.

Besides the computational gain that can be attained, we showed that source encoding techniques can also be used in the presence of highly non-convex misfit function. In this framework, the randomization of each search direction may allow for small local minima to be overcome. Indeed, stochastic optimization methods allow for the exploration of regions in the model space that are never accessible in the deterministic case because stochastic approaches allow the misfit function to increase. We showed a numerical test where we inverted nine frequencies in a single frequency group to render the misfit function highly non-convex, and we found that the final model is more accurate when we use random source encoding relative to the case where we use all the sources independently.

While all of the optimization methods generate subsurface models of similar accuracy for the synthetic example, application on real data from the Valhall field shows that the truncated Newton methods attain a lower misfit function value than the other optimization methods, hence suggesting a more robust behavior to noise and other source of errors such as incomplete wave physics. A speed-up of nearly one order of magnitude was reached for the selected stopping criterion of iterations. The accuracy of the subsurface models that was achieved for this stopping criterion of iteration was validated against published previous works, a sonic log and reverse time migration.

To further improve the convergence rates, it may be possible to design encoding strategies that take into account the variance analysis developed here. For example, start with a super source using all sources (case 1) and tend towards a super source that encodes only sources that have low cross-talk (for example because they are far away). Other hybrid strategies based on the behaviour of the misfit function may be implemented. It seems feasible to use a few number of super sources until the misfit function starts reaching a plateau. At this point, one could use all the sources independently and switch to a deterministic optimization problem, converging at a higher rate.

Finally, the conclusions that were inferred from the *2D* case studies need to be tested against applications of FWI in three dimensions. Having more sources may restrict more the model null space and may require less regularization. This should help to preserve the action of the Hessian in truncated Newton methods. However, the increased number of sources resulting from *3D* acquisitions will also generate crosstalk noise of higher amplitude. Therefore, there is a need to assess the benefit provided by the increased redundancy provided by *3D* acquisitions with respect to increased noise level created by cross-talk noise.

Currently, it is required to store in memory or on disk the direct and back-propagated wavefields for each source in the truncated Newton methods to build the sources of the adjoint-state equations during the computation of the Hessian vector product. This may not be feasible or too computationally expensive on a *3D* perspective, in particular if the inversion is performed in the time domain. However, this may be once more viable from an implementation point of view when source encoding is used with a few number of super sources.

Three-dimensional visco-elastic FWI is one of the main challenge of seismic imaging for the next decade. Taking into account the Hessian in multi-parameter FWI is crucial to manage the cross-talk between parameters of different nature. Owing that elastic seismic modeling is two to three orders of magnitude more expensive than acoustic modeling, the combination of random source encoding with truncated Newton methods should be of particular interest to manage both

the computational burden and the ill-posedness of *3D* visco-elastic FWI.

8 Tables

Table 3.1: BP- 2004 case study. Tuning parameters for optimization algorithms. The same parameters are used for all of the optimization methods. β: damping factor of the Hessian preconditioner. λ_x, λ_z: weighting factors applied to the Tikhonov regularization in the misfit function. ϵ_1, ϵ_2: stopping criteria of non-linear iterations (see text for more details). The number of memory models in l-BFGS is 5. The maximum number of CG iterations, N_{CG}, in truncated Newton methods is 30. The number of super-source K equals to 3 when source encoding is used. The maximum number of direct problems is 10^5. For this case study, the same tuning is used when source encoding is used or not. Note that we increase the weight of the Tikhonov regularization when noise is added to the data.

	No Noise				With Noise			
Source Encoding	β	$\lambda_x = \lambda_z$	ϵ_1	ϵ_2	β	$\lambda_x = \lambda_z$	ϵ_1	ϵ_2
No	10^{-2}	10^{-8}	$10^{-3}/10^{-2}$	0.01	10^{-2}	10^{-4}	0.25/0.7	0.1
Yes, $K = 3$	10^{-2}	10^{-8}	$10^{-3}/10^{-2}$	0.01	10^{-2}	10^{-4}	0.25/0.7	0.1

Table 3.2: BP- 2004 case study without noise. DP: number of direct problems. SE stands for source encoding. $S(\%)$: speed-up. $NLIt$: Number of non-linear iterations without SE. $NLIt_{SE}$: Number of non-linear iterations with SE. The values for each frequency group are shown separately and the last quantity is the total value for the whole inversion. (frequency group 1/ frequency group 2/ frequency group 1 + frequency group 2).

Optimization	DP without SE	DP with SE	$S(\%)$	m_{error} no SE
CG	54,684/100,068/154,752	22,476/33,582/56,058	59/66/64	0.0035
l-BFGS	15,128/16,492/31,620	9,510/7,746/17,256	37/53/45	0.0034
GN	15,996/21,328/37,324	7,200/20,070,27,270	55/6/27	0.0031
FN	44,268/39,556/83,824	20,706/19,632/40,338	53/50/52	0.0033
	$NLIt$	$NLIt_{SE}$	$NLIt/NLIt_{SE}(\%)$	m_{error} with SE
CG	283/547/830	1,880/2,790/4,670	85/80/82	0.0033
l-BFGS	114/126/240	1,280/1,030/2,310	91/88/89	0.0026
GN	29/23/52	310/1,070/1,380	90/98/96	0.0029
FN	66/50/116	1030/850/1,880	94/93/93	0.0029

Table 3.3: BP- 2004 with twenty five percent of noise in the data. The same nomenclature than for Table 3.2 is used.

Optimization Algo.	DP without SE	DP with SE	$S(\%)$	m_{error} no SE
CG	14,508/48856/ 63,364	5,412/19,425/24,864	62/60/61	0.0051
l-BFGS	8,184/18,600/26,784	6,27/5,967/12,252	23/68/54	0.0049
GN	8,308/53,940/62,248	6,144/44,592/50,736	26/17/18	0.0050
FN	13,640/43,400/57,040	12,066/28,782/40,848	11/33/28	0.0050
	It all sources	It Encoded Sources	Iteration Ratio (%)	m_{error} with SE
CG	96/288/384	460/1610/2070	79/82/81	0.0043
l-BFGS	58/133/191	870/810/1680	93/83/89	0.0039
GN	25/83/108	340/1930/2270	93/96/95	0.0047
FN	28/78/106	640/1210/1850	96/94/94	0.0043

Table 3.4: Valhall case study. Tuning parameters for optimization. $N_{CG} = 3$. The number of memory models in $l-$BFGS is 5. N_{nlit}^{max}: maximum number of non-linear iterations. See Table 3.1 for the nomenclature.

Source Encoding	β	λ_x	λ_z	ϵ_1	N_{nlit}^{max}
No	10^{-2}	10^{-3}	$2.5 \cdot 10^{-4}$	0.7	20
Yes, $K = 1$	10^{-2}	10^{-2}	$2.5 \cdot 10^{-4}$	0.7	-

Table 3.5: Valhall field data. Statistics of the FWI with and without source encoding. See Table 3.2 for the nomenclature.

Optimization Algo.	DP without SE	DP with SE	$S(\%)$
CG	45,360	3,464	92%
l-BFGS	36,120	1,714	95 %
GN	94,080	3,072	96 %
FN	94,080	3,572	96%

9 Figures

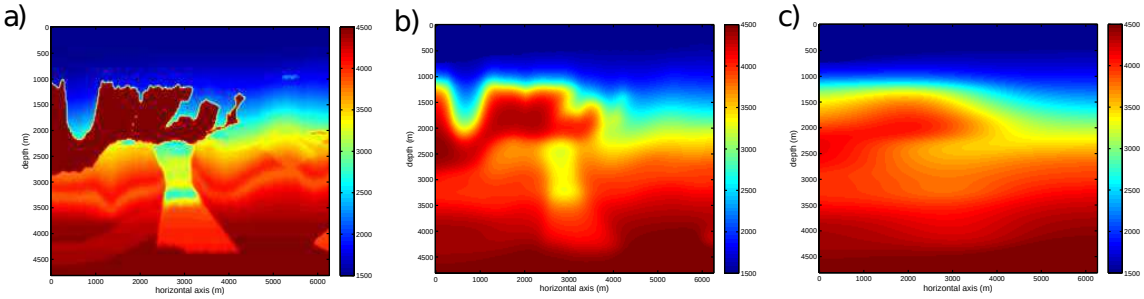


Figure 3.5: BP-2004 Salt v_p model. a) True model b) Initial model. c) Smoother initial model

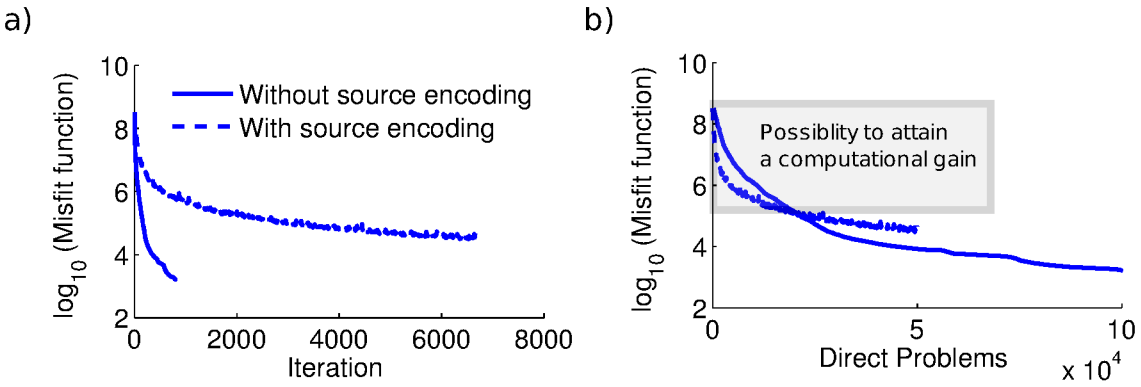


Figure 3.6: BP case study without noise - Comparison between the convergence rates (a) and the computational cost (b) of stochastic (dotted lines) versus deterministic (solid lines) algorithms. The optimization methods are l-BFGS and l-BFGS$_r$. FWI is performed for the first frequency group. The only stopping criterion is the maximum number of direct problems, which is set to 10^5. This criterion was not reached by the stochastic approach because of line search failure. (a) The convergence rate is higher for the deterministic method ($O(1/I^{\alpha})$, $\alpha = 1, 2$) than for the stochastic one ($O(1/\sqrt{I})$). (b) The computational cost (measured by the number of direct problems) for the stochastic methods is lower at the beginning of the inversion. However, the deterministic inversion will eventually catch up with the stochastic inversion because deterministic methods have higher convergence rates.

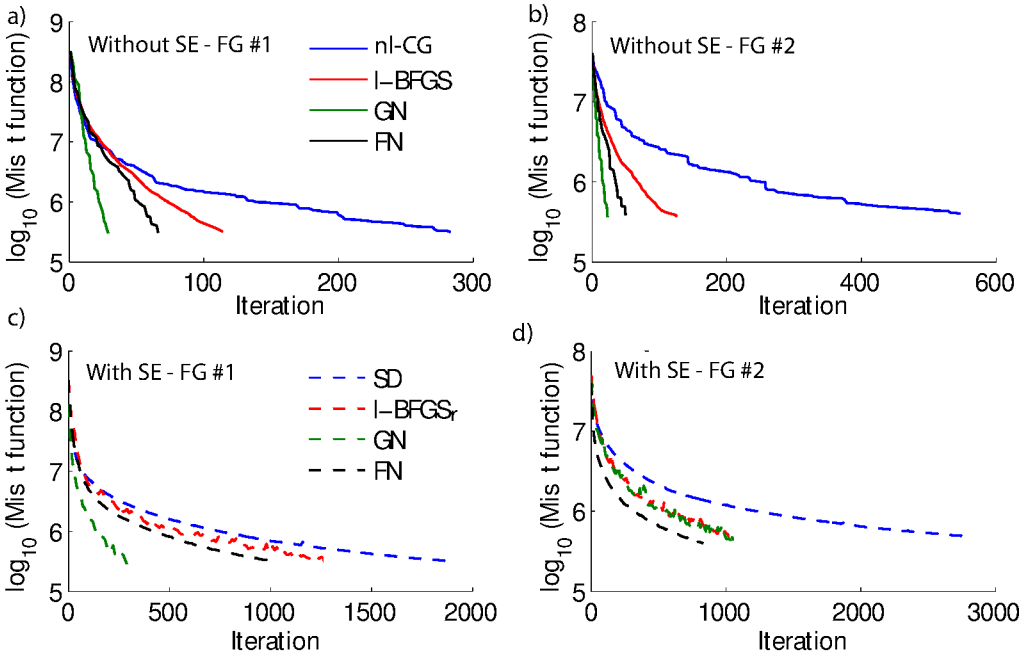


Figure 3.7: BP case study without noise - Convergence rate. Reduction of the misfit function as a function of the iteration number without (a-b) and with (c-d) source encoding. (a, c) First frequency group (FG #1). (b, d) Second frequency group (FG #2). The curves are shown for the four optimization methods. $nl-$CG/SD: blue lines. l-BFGS($_r$): red lines. GN: green lines. FN: black lines.

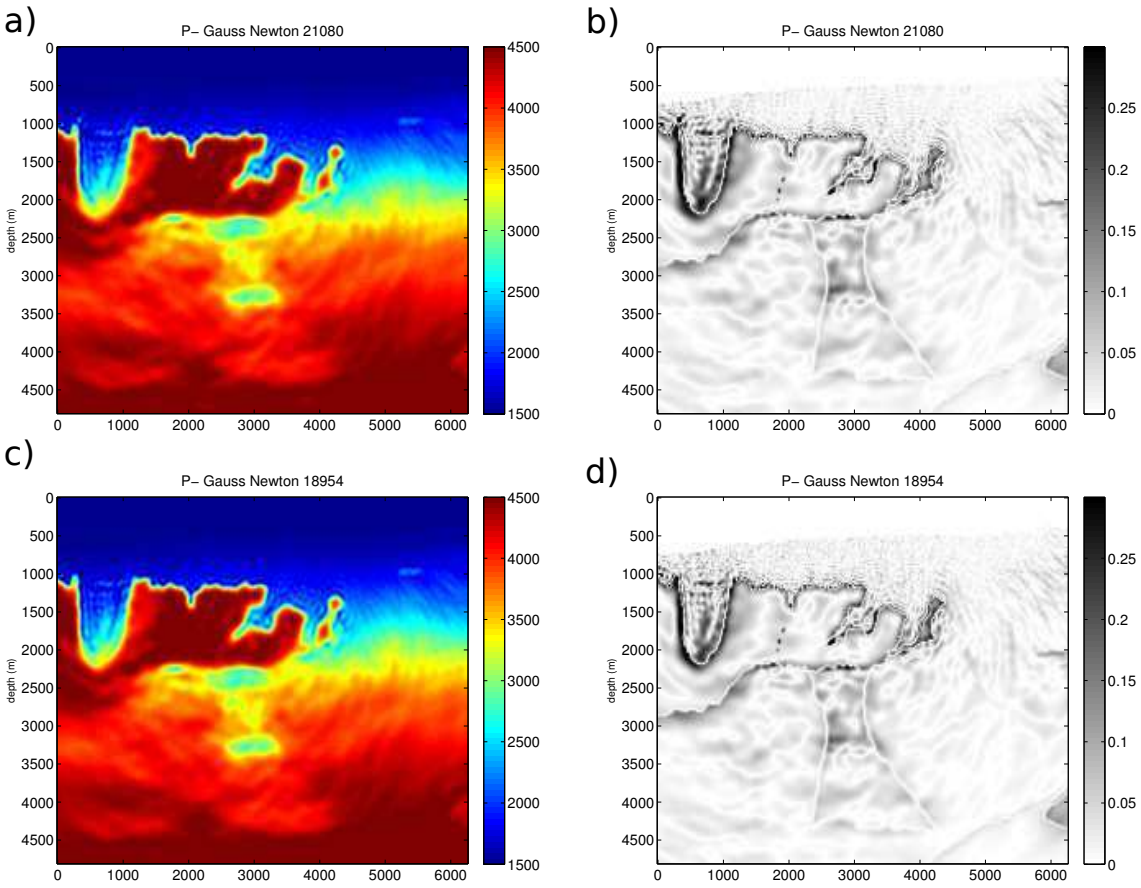


Figure 3.8: BP case study without noise. (a, c) Final FWI model for the GN optimization method without (a) and with (c) source encoding ($K = 3$). (b, d) Velocity model error (difference between the final FWI model and the true subsurface model).

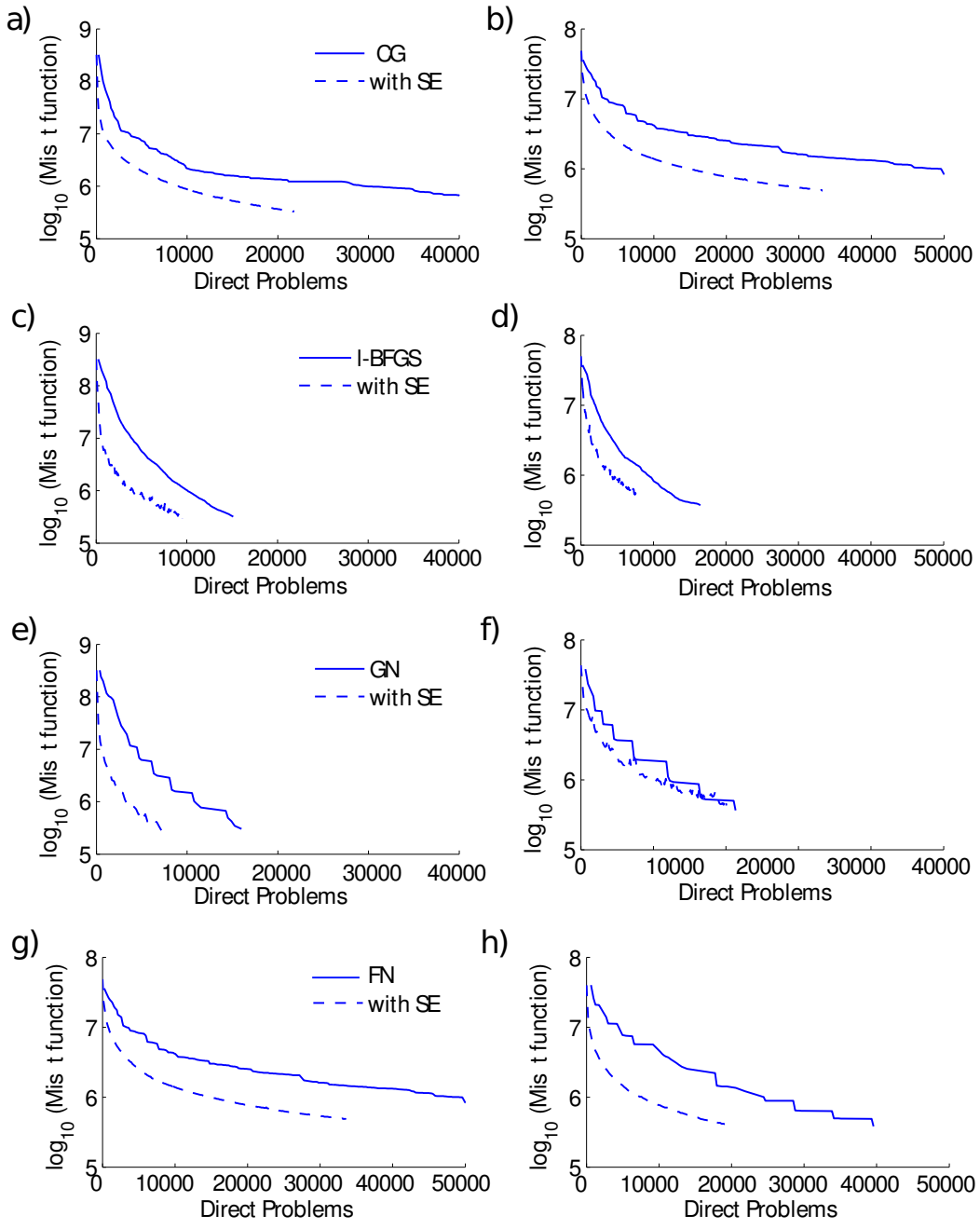


Figure 3.9: BP case study without noise. Assessment of the computational efficiency provided by source encoding - Reduction of the misfit function as a function of the direct problems, for the first (a,c,e,g) and second (b,d,f,h) frequency groups. (a-b) $nl-$CG optimization method. (c-d) $l-$BFGS optimization method. (e-f) GN optimization method. (g-h) FN optimization method. The computational gain is provided by the difference between the number of direct problems performed with (dash lines) and without (solid lines) source encoding for a given misfit function value.

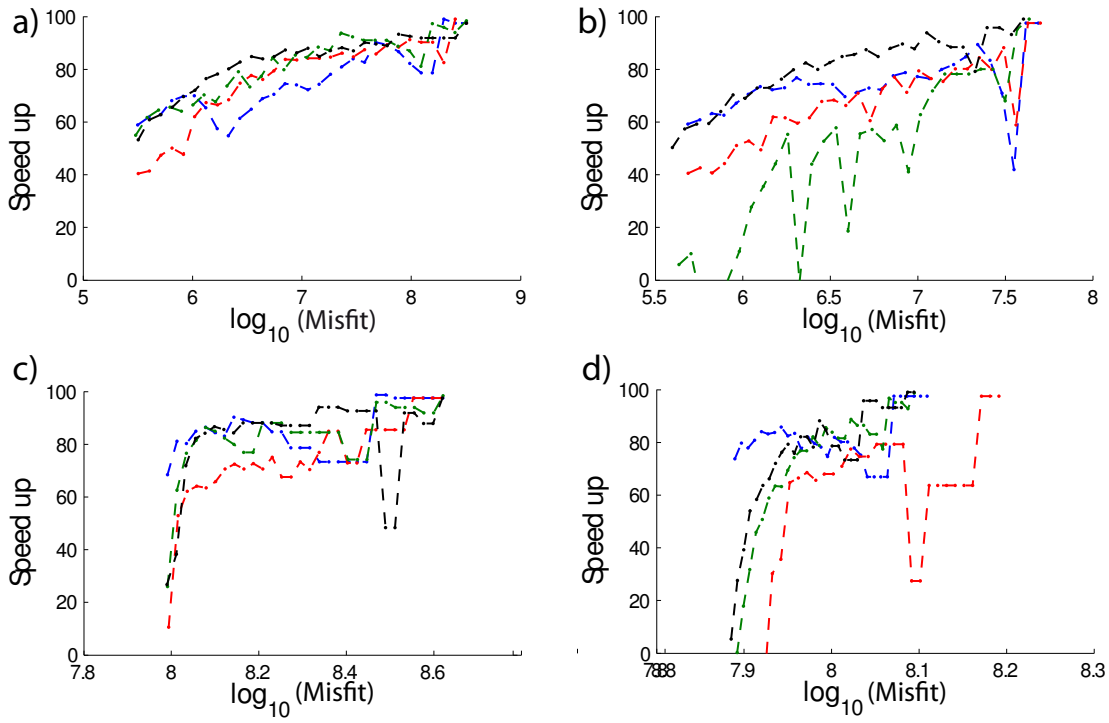


Figure 3.10: BP case study. Speed-up versus data misfit function for the two frequency groups for the case without noise (a-b) and with noise (c-d). Figures a-b summarize the results for the speed-up without noise for each optimization method shown in Figure 3.9. Figures c-d summarize the results for the speed-up with noise for each optimization method shown in Figure 3.13. Note how the speed-up drops abruptly in the case of noisy data, as the convergence rate slows down near the minimum of the misfit function. $nl-$CG/SD: blue lines. l-BFGS: red lines. GN: green lines. FN: black lines.

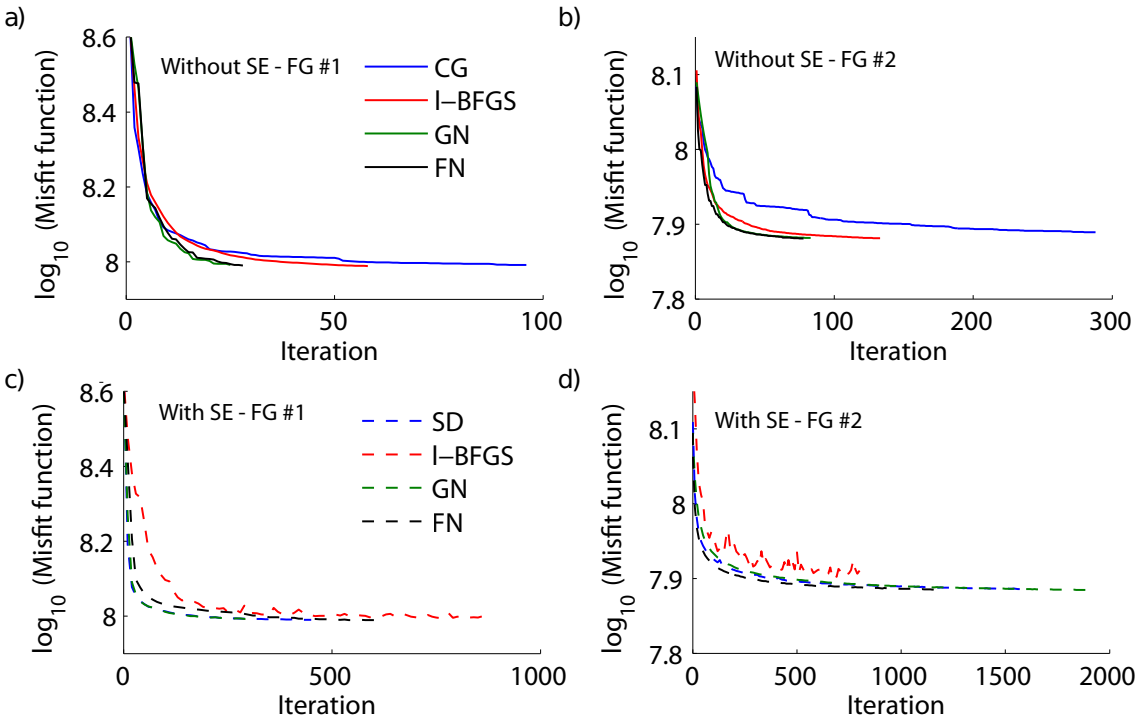


Figure 3.11: BP case study with noise - Convergence rate. Reduction of the misfit function as a function of the iteration number without (a-b) and with (c-d) source encoding. (a, c) First frequency group (FG #1). (b, d) Second frequency group (FG #2). The curves are shown for the four optimization methods. $nl-$CG: blue lines. l-BFGS: red lines. GN: green lines. FN: black lines.

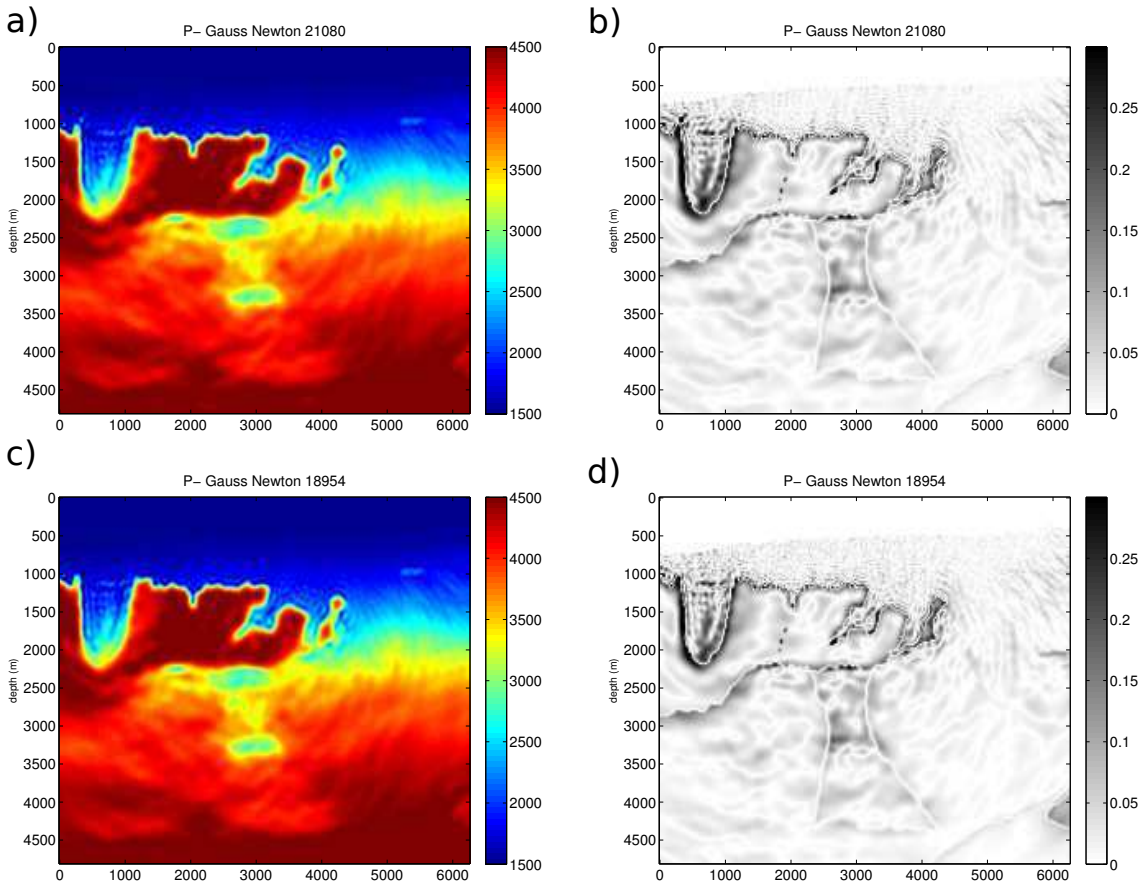


Figure 3.12: BP case study with noise. (a, c) Final FWI model for the GN optimization method used without (a) and with (c) source encoding ($K = 3$). (b, d) Velocity model error (difference between the final FWI model and the true subsurface model).

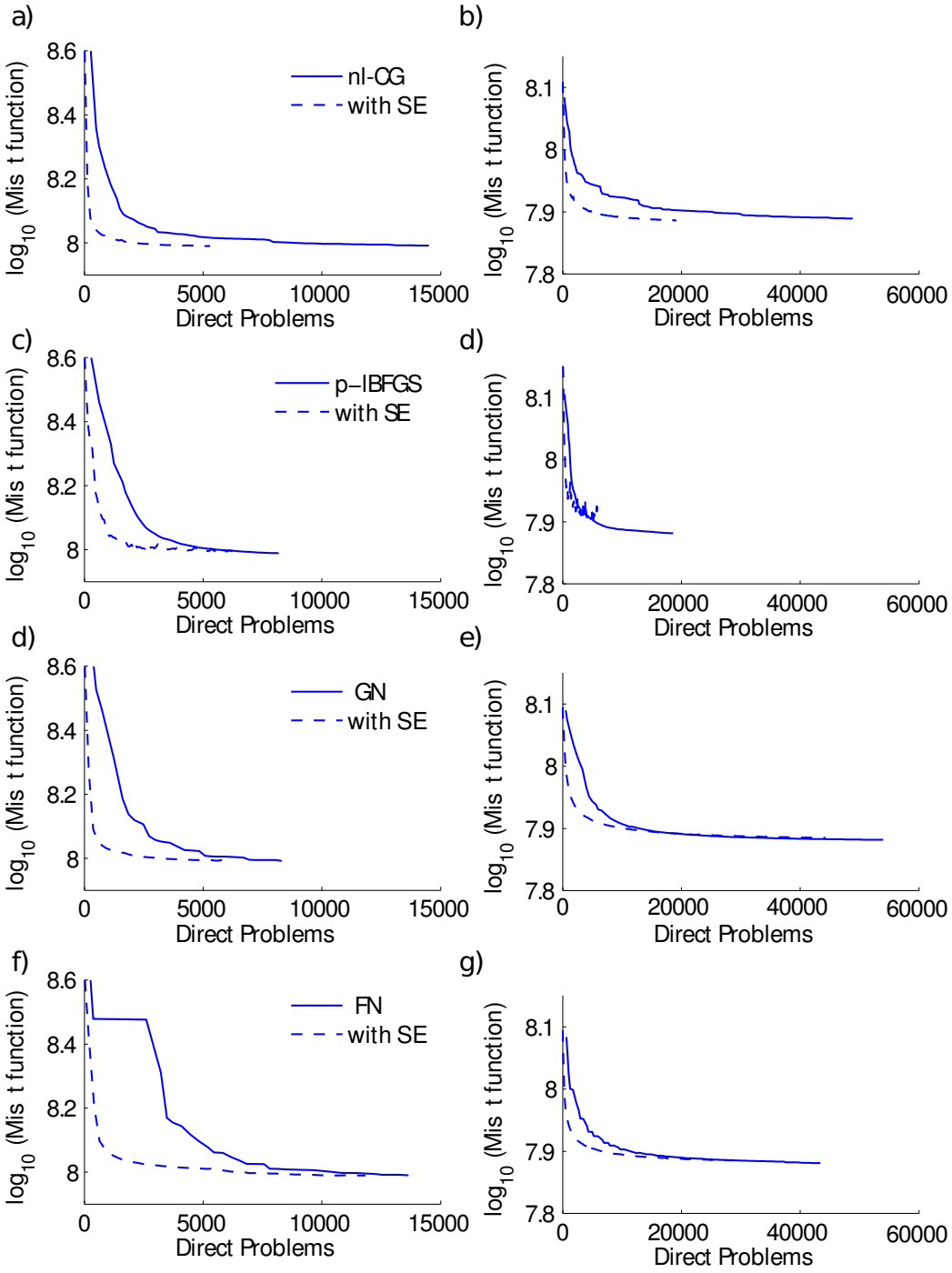


Figure 3.13: BP case study with noise. Assessment of the computational saving provided by source encoding - Reduction of the misfit function as a function of the direct problems, for the first (a,c,e,g) and second (b,d,f,h) frequency groups. (a-b) $nl-$CG optimization method. (c-d) $l-$BFGS optimization method. (e-f) GN optimization method. (g-h) FN optimization method. The computational gain is provided by the difference between the number of direct problems performed with (dash lines) and without (solid lines) source encoding for a given misfit function value.

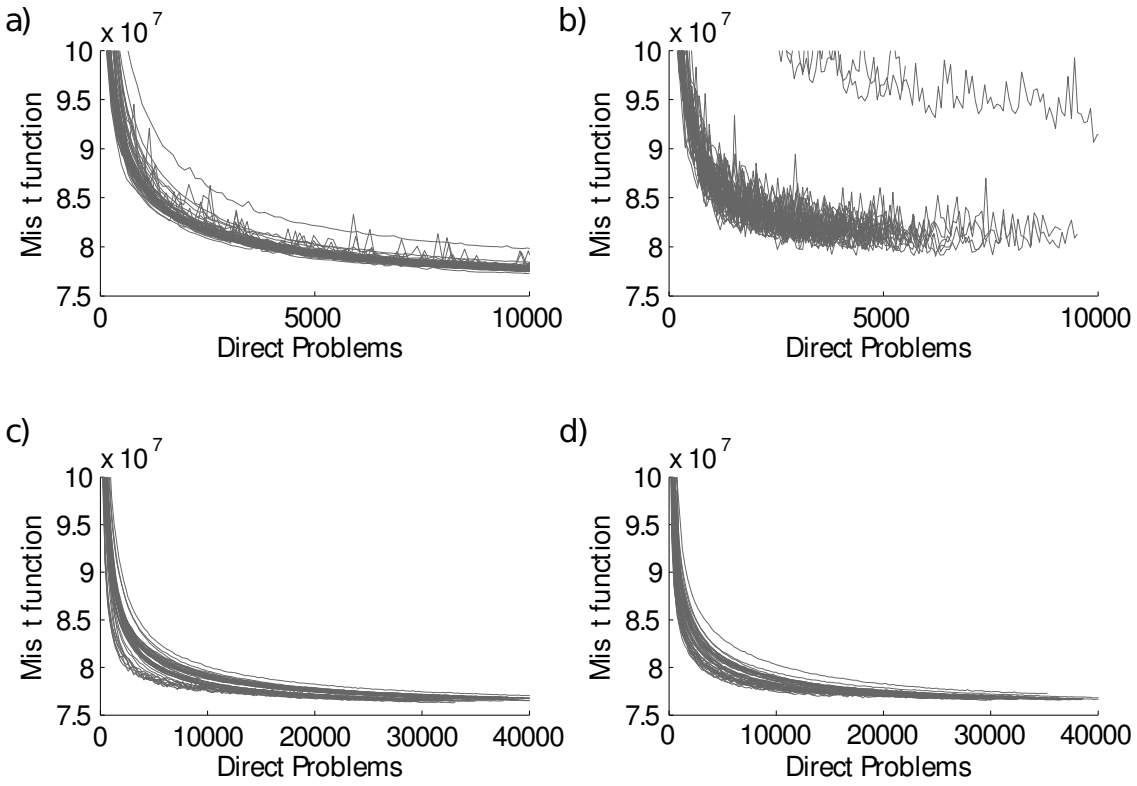


Figure 3.14: BP case study with noise. Misfit function value as a function of the number of direct problems for the second frequency group, for 50 independent realizations, using source encoding. a) SD b) l-BFGS c) GN d) FN

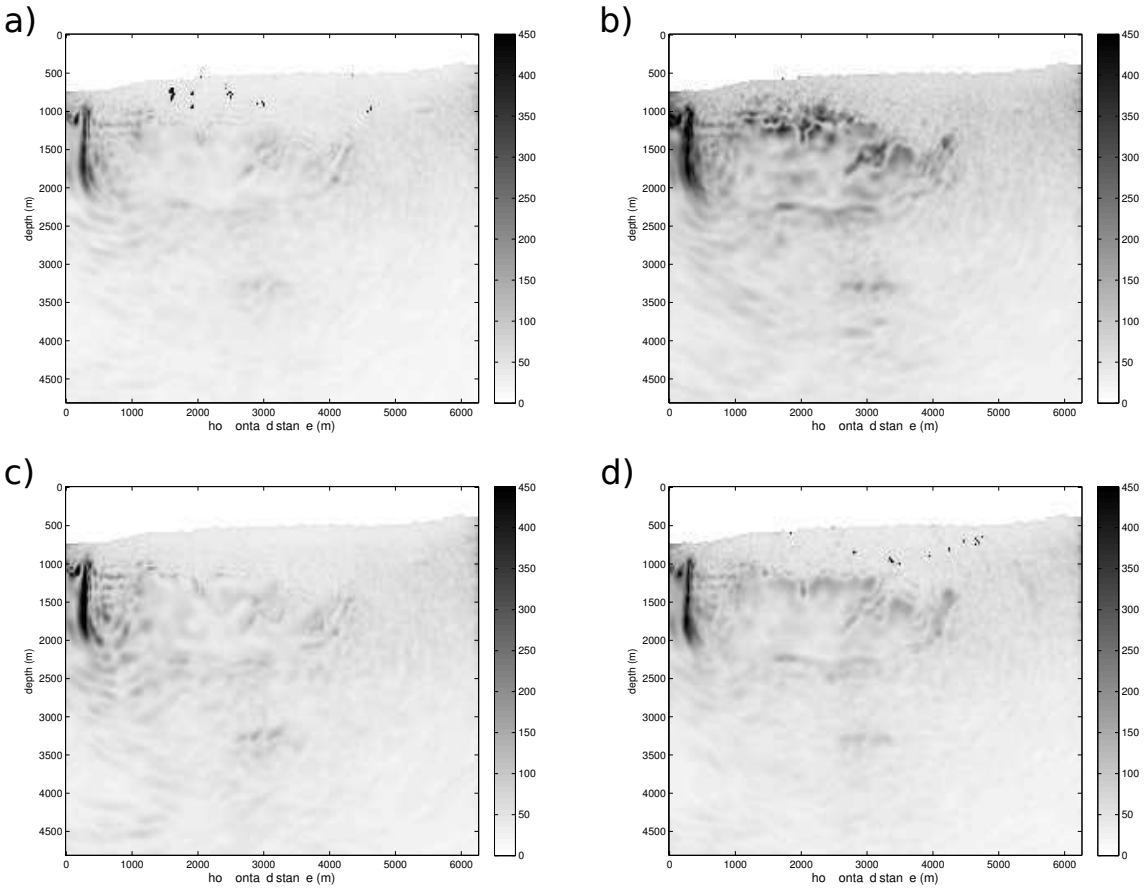


Figure 3.15: BP case study with noise. Standard deviation of the final velocity model for 50 realizations using source encoding. a) SD b) l-BFGS c) GN d) FN

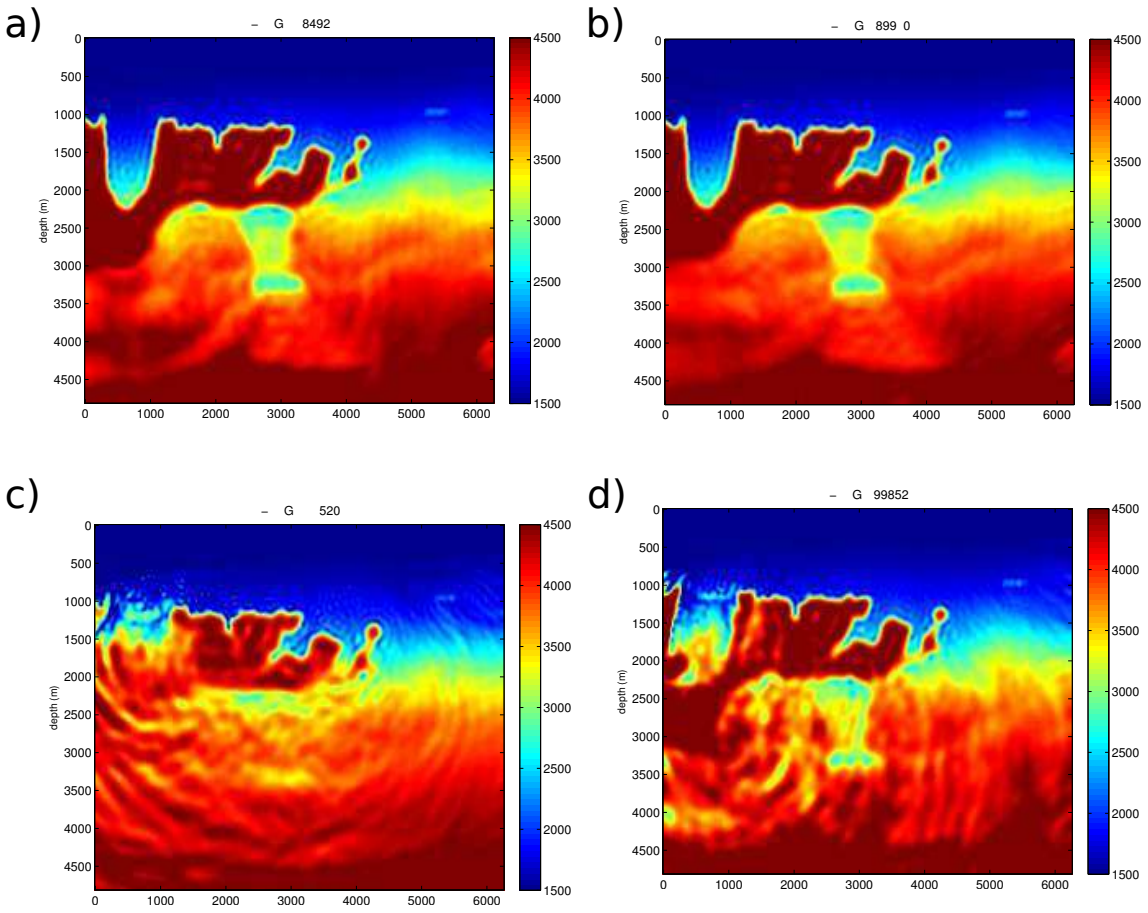


Figure 3.16: BP case study without noise. (a-b) Final velocity models obtained by inverting a single frequency group containing nine frequencies between 1Hz and $9Hz$ without (a) and with source encoding (b) and with the initial velocity model shown in Figure 3.5b. (c-d) Same as (a-b) with the smoother initial velocity model shown in Figure 3.5c. Note how source encoding allows to reach an improved local minimum.

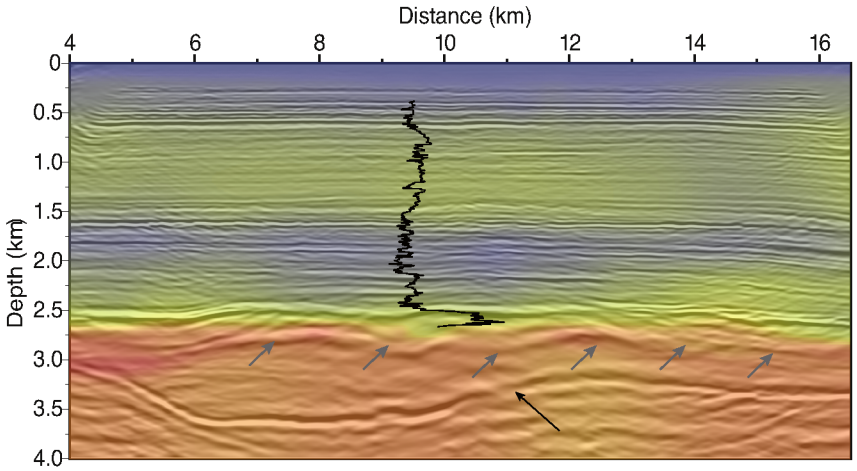


Figure 3.17: Valhall case study. Reverse time migrated image computed in the initial vertical velocity model, which is superimposed with a transparency. A sonic log located at 9.5 km in distance is also superimposed on the migrated image. The gray arrows delineate the base of the reservoir. The black arrow points a location in depth where the image of a deep reflector lacks continuity.

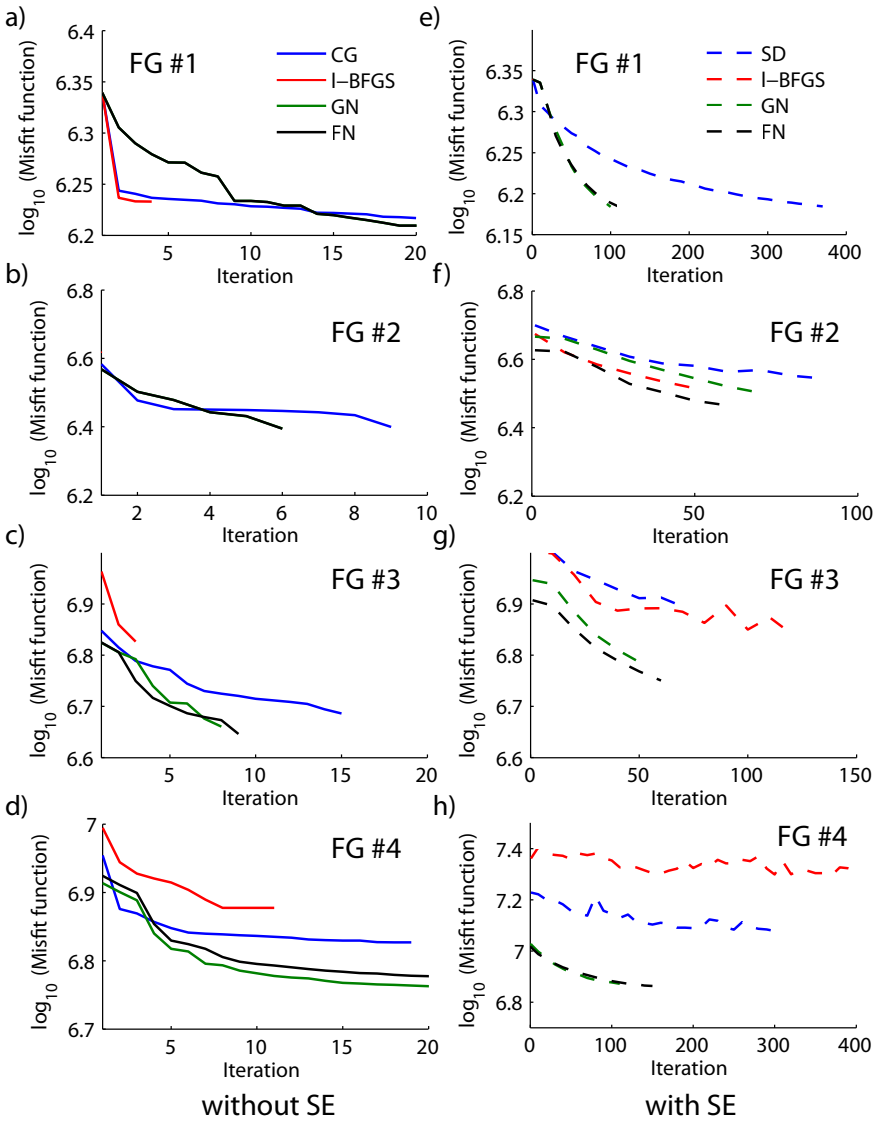


Figure 3.18: Valhall case study. Convergence of FWI. (a-d) Misfit function versus iteration number for each frequency group without SE. (e-f) Same as (a-d) when SE is used. Blue lines: $nl-$CG optimization method. Red lines: $l-$BFGS method. Green: GN optimization method. Black lines: FN optimization method. (a, e) First frequency group. (b, f) Second frequency group. (c, g) Third frequency group. (d, h) Fourth frequency group.

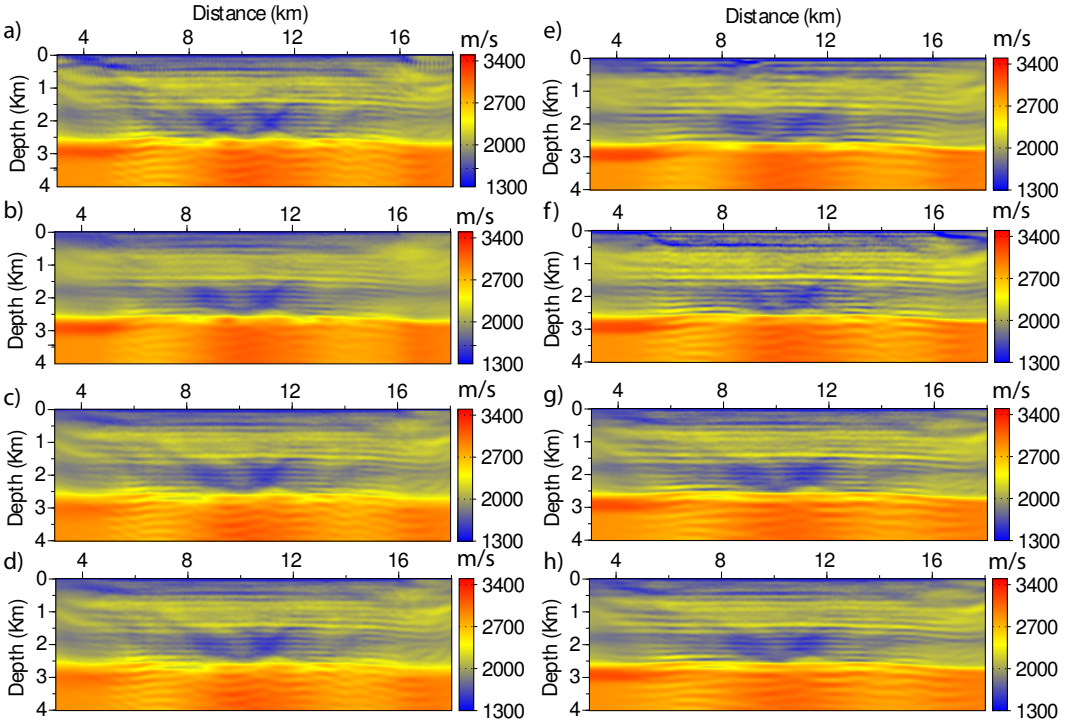


Figure 3.19: Valhall case study. Final FWI velocity models obtained without (a-d) and with (e-h) source encoding. (a,e) $nl-$CG/SD optimization method. (b, f) $l-$BFGS/BFGS$_r$ optimization method. (c, g) GN optimization method. (d, h) FN optimization method.

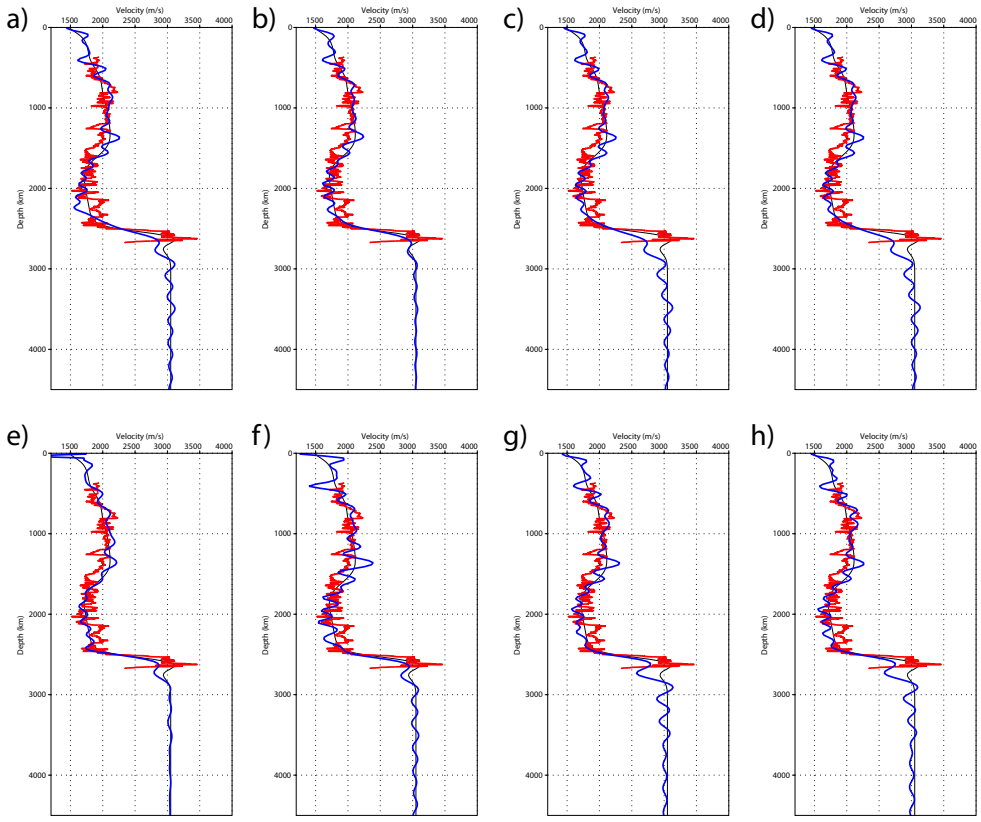


Figure 3.20: Valhall case study. (a-b) Logs of the final velocity models (blue lines) obtained without source encoding at a horizontal distance x=9.5km for $nl-$CG/SD (a), $l-$BFGS/BFGS$_r$ (b), GN (c) and FN (d). (e-g) Same as (a-b) when source encoding is used during FWI. The sonic log is plotted with a red lines and the log of the initial model is plotted with a black line.

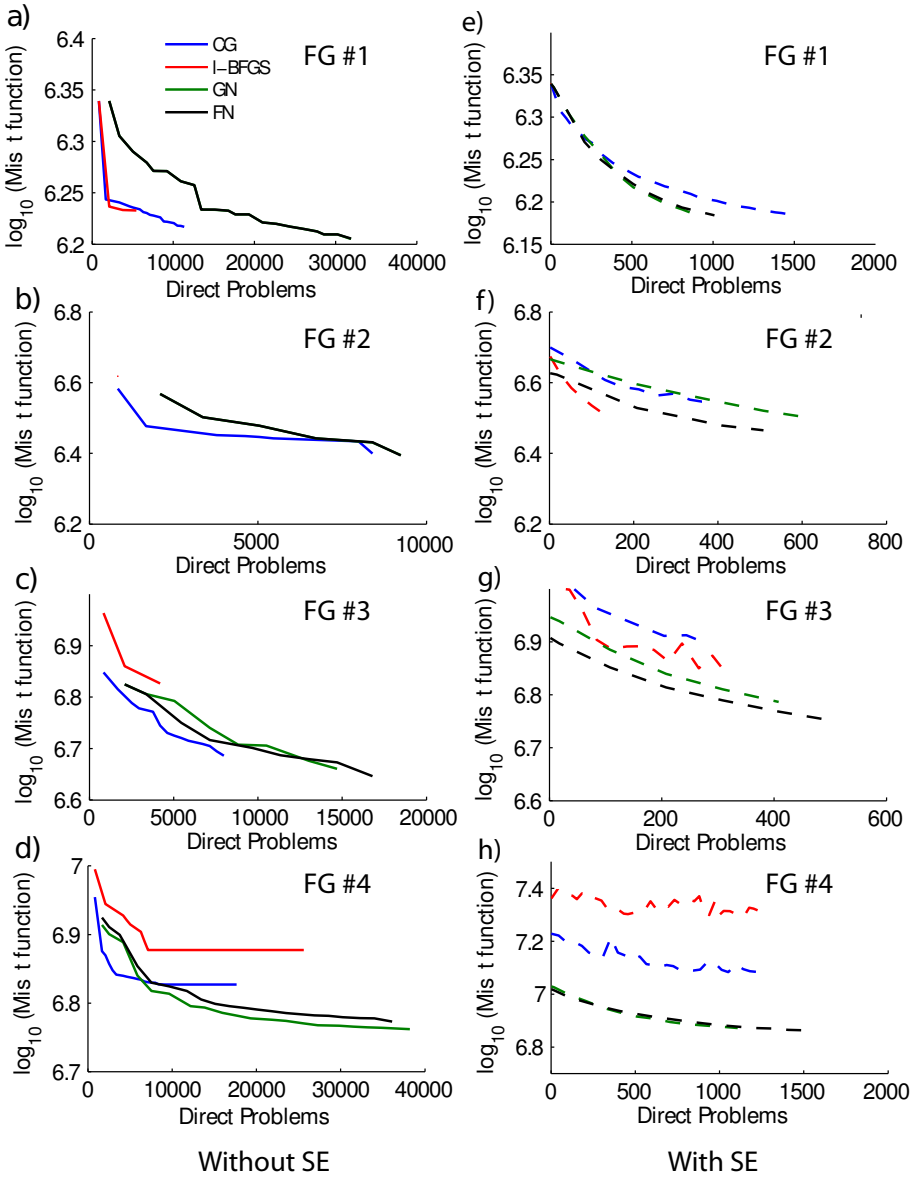


Figure 3.21: Valhall case study. Computational cost. Misfit function versus number of direct problems for each optimization method and for each frequency group when all the sources are processed independently (a-d) and when source encoding is used (e-h). Blue lines: SD. Red lines: $l-BFGS$. Green lines : GN. Black lines: FN. (a, e) First frequency group. (b, f) Second frequency group. (c, g) Third frequency group. (d, h) Fourth frequency group.

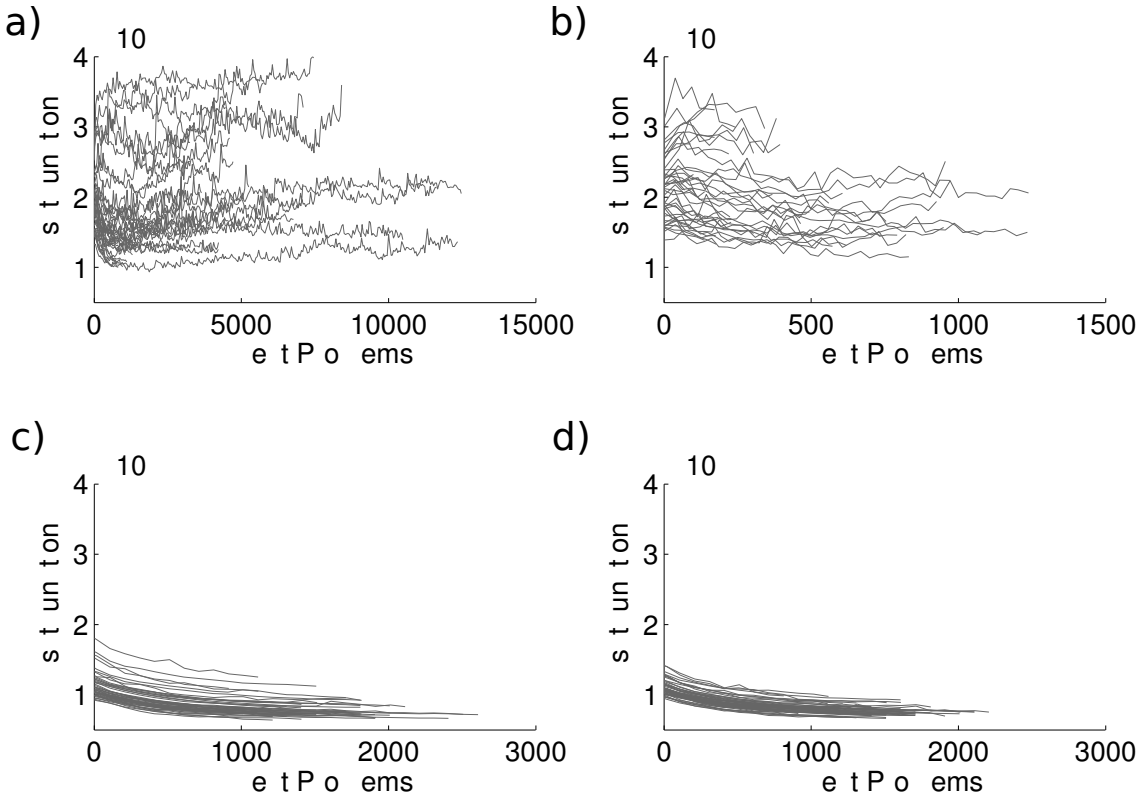


Figure 3.22: Valhall case study. Reduction of the misfit function of the last frequency group versus the number of direct problems for 50 realizations of FWI with source encoding. a) SD. b) $l-\text{BFGS}_r$, c) GN, d) FN. Random variable follow a normal distribution.

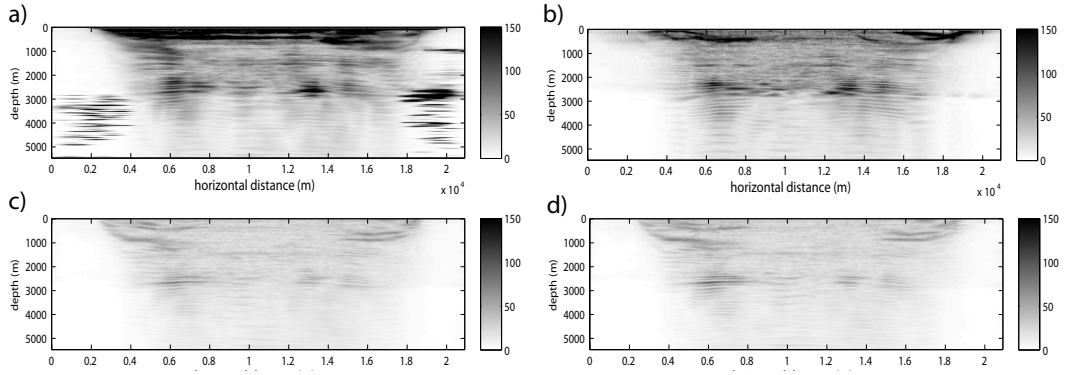


Figure 3.23: Valhall case study. Standard deviation of the final velocity model for 50 realizations of FWI with source encoding. a) SD. b) $l-\text{BFGS}_r$, c) GN, d) FN. Random variable follow a normal distribution.

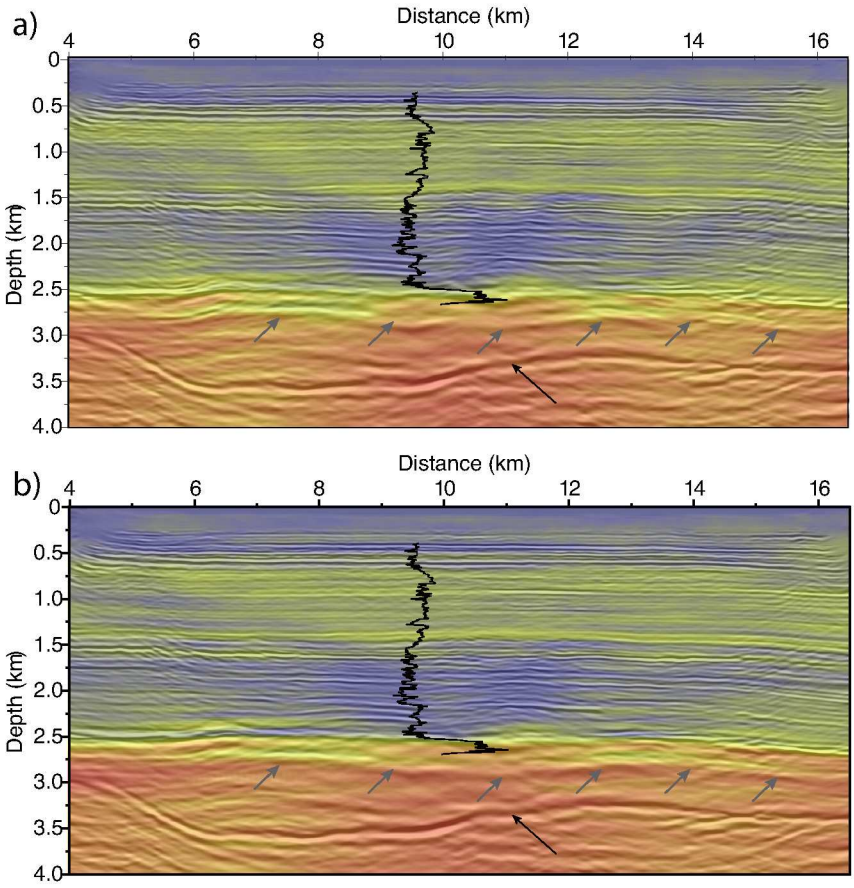


Figure 3.24: Valhall case study. Reverse time migrated image computed in the FWI model inferred from the FN optimization method without (a) and with (b) source encoding. The vertical velocity model within which the reverse time migrated image is computed is displayed with a transparency allowing one to check the consistency between the background velocities built by FWI and the reflectors mapped by the migration. The sonic log as well as the black and gray arrows shown in Figure 3.17 are superimposed.

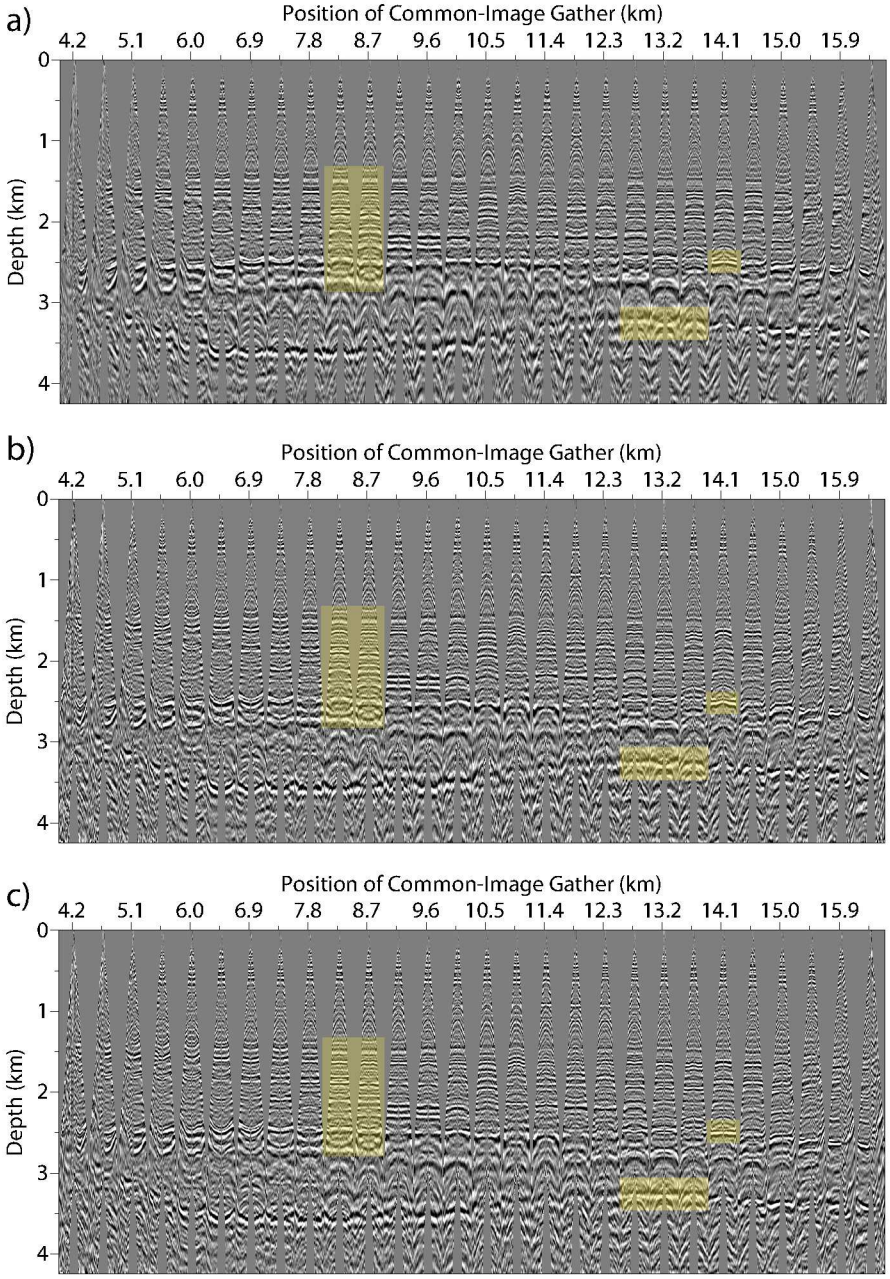


Figure 3.25: Valhall case study. Common image gathers in the offset-depth domain computed by reverse time migration performed in the initial model (a) and in the final FWI models inferred from the FN optimization method without (b) and with (c) source encoding. Reflectors which are flatter in the CIGs computed in the FWI model obtained with source encoding are highlighted.

Chapter 4

Regularization techniques for FWI

This last chapter is devoted to a few aspects concerning the inverse problem. In Section 1.1, we compare the Total Variation (TV) regularization and l_2 norm regularization on synthetic and real data. We find that the regularization term with the TV norm provides adequate earth models, including with the real data. We attempt to improve the quality of the final model in Section 1.3 by applying a TV denoising technique. We perform a slight adaptation to carry out a local TV denoising based on incorporation of the information provided by a migration analysis. The modified TV denoising technique shows the capacity to denoise the models and maintain the critical reflectors unchanged. Finally, in Section 2 we provide some observations regarding the spectral content of the measured data for the BP-2004 salt model using a surface acquisition. We identify a difference in the spectrum of the short offsets, receivers close to the reflecting boundaries, and receivers far from the reflecting boundaries. We see the reflected and transmitted waves have a different spectral content, and we observe a gap in the reflected wave spectrum. The impact this has on the inversion is yet to be studied with more detail, but we believe that this gap enlarges the model null space that may be otherwise reduced by including a wider range of frequencies in the inversion.

1 Regularization

If the inverse problem is not well posed, as discussed in Chapter 1, we must include additional information to find a feasible solution through a process called regularization. Regularization reduces the null space and stabilizes the solutions by including a priori information of the expected solution. For example, the regularization term must reflect if the desired solution should be smooth, highly contrasted, or with few non zero elements. The known additional information is transferred to the optimization problem by imposing additional constraints. For example, we could impose the parameter solution to have minimal l_2 or total variation (TV) norms. Alternatively, we could ask for maximum sparsity in some basis representation or to minimize the difference with respect to a previous model. Each choice privileges some characteristics of the desired solution. The theory of regularization for linear inverse problems is well developed

(Tikhonov and Arsenin, 1977; Engl et al., 2000; Kaltenbacher et al., 2008; Whitney, 2009). We extract a few key notions from Engl et al. (2000).

Full waveform inversion is an example of an ill posed inverse problem, where regularization is needed. Some of the major causes of the ill-posedness are due to:

- Limited aperture sampling that creates regions that are not sampled by the waves. Regions not sampled by waves belong to the null space.
- Regions beneath interfaces with a high reflection coefficient (such as salt bodies), are poorly sampled by the waves. Model parameters in weakly illuminated regions have a very low influence on the data. These model parameters can thus be considered as part of the null space.
- Noise in the data. The none uniqueness of the solution increases when the data is noisy. That is, more model parameters provide the same misfit function in the data space.
- For a complete reconstruction of the subsurface image, the complete wavenumber spectrum must be reconstructed. In order to do so, a wide range of frequencies in the data and offsets are necessary. The finite bandwidth of the source spectrum implies that certain frequencies are not present in the measured data. In practice, the aperture of the acquisition is limited. Therefore, the measured data is insufficient and does not constrain all the model parameters, of all scales.

1.1 Tikhonov regularization

In 1963, Andrey Tikhonov introduced a regularization algorithm in an attempt to stabilize the solution. The regularization algorithm, which now bears his name, is widely used. A general linear least squares problem may have an infinite number of solutions. For example, when we consider that the data are noisy, there may not be a solution that fits exactly the noise. The simplest form of Tikhonov regularization adds information to the optimization problem by restricting the solutions to those that have the minimum l_2 norm. Intuitively, it is possible to see that this restriction might be closer to the solution we are seeking, because although we want the minimum misfit, we do not want to fit exactly all the perturbations produced by noise in the data.

The same idea can be applied in the framework of FWI. Recall that the misfit function is defined by

$$\phi_0(m) = \frac{1}{2}\sum_{\omega_i}^{N_f}\sum_{s=1}^{N_s}\sum_{r=1}^{N_r}\|Pu_{s,r}(x,\omega_i) - d_{s,r}(x,\omega_i)\|_2^2.$$

To simplify, let us consider a level of noise uniform among frequencies, sources and receivers, denoted by σ. Then, the optimization problem with Tikhonov regularization consists in solving

$$\mathbf{P_1} \qquad \min ||Rm||_2^2 \tag{4.1}$$

$$\text{such that} \quad 2\phi_0(m) \leq N_f N_s N_r \sigma^2, \tag{4.2}$$

where R is an operator controlling the kind of solution favored by the regularization. For example, if $R = \mathbb{I}$, the regularization term favors minimum norm solutions, but requires no spatial regularization. If $R = \nabla$ or $R = \nabla^2$ smooth solutions will be privileged.

Using Lagrange multipliers, we can show (see for example Engl et al. (2000)) that, for an appropriate choice of $\lambda > 0$, this constrained optimization problem can also be expressed as

$$\mathbf{P_2} \quad \min\left\{\phi_0(m) + \lambda||Rm||_2^2\right\}. \tag{4.3}$$

Optimization problem (4.3), is known as the unconstrained optimization problem with Tikhonov regularization, and λ is known as the regularization parameter.We will denote by *regularized mis-*

fit function (ϕ) the objective function of (4.3), i.e.

$$\phi(m,\lambda) = \phi_0(m) + \lambda||Rm||_2^2.$$

We will therefore try to minimize, simultaneously two functionals and the value of λ determines the relative weight between the two. Thus the actual outcome depends on the choice of λ. Good criteria to choose the λ depend on whether the value of the level noise σ (or at least an educated guess for it) is available or not.

1.1.a *Choice of λ for the known noise level case*

One of the most widely used criteria when the noise level σ is know is the *Morozov discrepancy principle (MDP)*. It accounts for choosing λ such that, at the optimal, $||Pu(m) - d|| \approx \sigma$. The logic behind this principle is that it is unreasonable to ask for a solution with a discrepancy $||Pu(m)-d||$below σ, due to the effect of the noise. Since little regularization implies less stability, the best possible regularization will be obtained with the largest regularization parameter that gives a discrepancy of the same order of σ. Note that an overestimation of the noise level might lead to a loss of accuracy, and an underestimation might render the solution unstable. In practice, the MDP would imply to solve the problem with an arbitrary value of λ and then progressively modify it until the discrepancy is sufficiently close to σ.

This principle can be refined under further assumptions. For example, the popular χ^2 criterion (Nolet, 2008) is based on a maximum likelihood estimation under the assumption that the errors are uncorrelated amongst samples, and follow a Gaussian distribution.

Let χ^2 be a function of the model, defined as the data misfit normalized by the variance of the noise σ^2, i.e.

$$\chi^2(m) = \frac{2\phi_0}{\sigma^2}. \tag{4.4}$$

If there are no sources of bias, the expected value $\mathbb{E}(d)$ is equal to the error-free value. Now, we have that the density function of each data is

$$P(d_{s,r}(.,\omega_i)) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{|d_{s,r}(.,\omega_i) - \mathbb{E}(d_{s,r}(.,\omega_i))|^2}{2\sigma^2}\right), \tag{4.5}$$

and by independence, the joint density is

$$P(d) = \prod_{i=1}^{N_f}\prod_{s=1}^{N_s}\prod_{r=1}^{N_r} \frac{1}{(\sigma\sqrt{2\pi})^{N_f+N_s+N_r}} \exp\left(-\frac{|d_{s,r}(.,\omega_i) - \mathbb{E}(d_{s,r}(.,\omega_i))|^2}{2\sigma^2}\right). \tag{4.6}$$

Replacing the expected values, with those predicted by the model

$$P(d|m) = \prod_{i=1}^{N_f}\prod_{s=1}^{N_s}\prod_{r=1}^{N_r} \frac{1}{(\sigma\sqrt{2\pi})^{N_f+N_s+N_r}} \exp\left(-\frac{|d_{s,r}(.,\omega_i) - Pu_{s,r}(m,\omega_i)|^2}{2\sigma^2}\right). \tag{4.7}$$

In the last step, the (perhaps unrealistic) assumption that is made is that no errors are introduced in the modelling due to linearisations, unaccounted for physics of anisotropy, elasticity, source function errors, etc. The goal is to find the model that is associated with high probability with its predicted data vector. Therefore, a likelihood function is defined,

$$\mathcal{L}(m|d) = P(d|m) \propto \exp\left(-\frac{1}{2}\chi^2\right). \tag{4.8}$$

Maximizing the likelihood $\mathcal{L}$ involves minimizing χ^2. That is, minimizing the sum of squares of data misfit (method of least squares). In practice data may suffer from outliers that deviate

from the Gaussian. A possible solution is to pre-treat the data to remove these outliers. Note that the χ^2 function is essentially a statistical measure of the goodness of fit.

☛ In the case that every datum is satisfied with a misfit of one standard deviation, the misfit criteria gives $\chi^2 = N := N_f + N_s + N_r$. The confidence interval, given that the errors follow a Gaussian distribution, is $\chi^2 \approx N \pm \sqrt{2N}$.

1.1.b *Choice of λ for the unknown noise level case*

Some heuristic parameter rules may be employed when we lack information on the noise level, or when it is not reliable. These error free parameter choice rules cannot be proved to converge, but in many cases they provide acceptable results. We focus on one of these methods, the *L-curve* method by Hansen (Hansen, 1998). Even though this method is widely used, it does not have a solid mathematical foundation, as is explained in Engl et al. (2000). It consists in plotting ϕ_0 vs. $||Rm||$, for different values of λ. See for example Figure 4.3 e).

The logic of the *L-curve* is that if λ is below the optimal value, then the norm of the discrepancy, which is around σ changes very little, with respect to changes in λ. Since this corresponds to a situation without regularization, the norm of the solution will be very big. In the graph, this can be seen as a very steep slope. The other situation is when λ is considerably greater than the optimal value. In this case, the dominant term in the misfit function will be the norm of the approximation, and the residual norm may be arbitrarily large. The L-curve is frequently used with Tikhonov regularization, but the same logic for the L-curve is preserved for other regularization methods. However, it is not always guaranteed that the curve will always have an L shape.

☛ Typically, the value near the corner of the L-curve is preferred, which represents a compromise between minimizing the residual norm and the penalty term. Attempts have been made to find the corner of the L-curve with numerical optimization routines (Engl et al., 2000).

1.1.c *Example 1: No regularization and overfitting the data*

Consider a velocity model consisting of an homogeneous background velocity of $4000m/s$ and a rectangular heterogeneity of $4300m/s$ as shown in Figure 4.1. There are 38 sources and 72 receivers per source, around the target area. Absorbing boundary conditions are imposed on all the edges of the domain. Starting from an homogeneous background velocity model of $4000m/s$, we perform an acoustic inversion in the frequency domain. Using only one frequency group, we invert frequencies from $2\ Hz - 6\ Hz$, with $1Hz$ interval. We add 5% of uncorrelated Gaussian noise to the data. *In this example, since we have coverage of sources and receivers around the target area and the structure of the model is simple with low velocity contrasts, the major source of ill-posedeness comes from the fact that the data is noisy.* The inversion algorithm will thus try to fit the noise.

We add noise to the data by generating independent random Gaussian variables η with zero mean and variance σ^2,

$$\eta_i \backsim \mathcal{N}(0, \sigma^2) = \sigma\mathcal{N}(0, 1).$$

Let $\bar{d}$ be the true observed data without noise. The noisy data is simply, $d_{s,r}(.,\omega_i) = \bar{d}_{s,r}(.,\omega_i) + \eta_{s,r}(.,\omega_i)$.

$$P_\eta = \mathbb{E}[\eta^*\eta] = \mathbb{V}[\eta] = \sigma^2, \tag{4.9}$$
$$P_d = \mathbb{E}[d^*d]. \tag{4.10}$$

For this example, we choose σ^2 such that

$$P_{noise} = \gamma \cdot P_{data}, \tag{4.11}$$

where $\gamma = 0.05$. Since we are working in the frequency domain, a different σ value is used for each frequency to satisfy (4.11). We perform the *inversion without regularization*, and use the values of σ we chose, and replace them in the χ^2 criteria in equation (4.4). Additionally, we normalize by the number of data samples N, such that the minimum desired value is 1. The reduction of the χ^2 function as a function of iterations is plotted in Figure 4.2a. The red dotted line indicates the value of $\chi^2 = 1$. As we can see, the χ^2 function decreases below the value of 1, approximately after iteration 5. This means that past the fifth iteration, the inversion is most likely to be fitting the noise. Figure 4.2b plots the model error $m_{err} = \dfrac{\|m_{true} - m_i\|_2^2}{\|m_{true}\|_2^2}$ as a function of the iterations. Shortly after iteration 5, the model error starts to increase, confirming that the inversion is fitting the noise in the data. Therefore, in knowledge of the noise level, the inversion should be stopped when $\chi^2 = 1$. However, our only stopping criteria for this example is a maximum number of iterations which is equal to 30. The velocity model at iteration 5 is shown in Figure 4.2c, and the velocity model at the end of the inversion is shown in Figure 4.2d. Clearly, the velocity model at $\chi^2 = 1$ is better than the final velocity model at iteration 30, despite having a smaller data misfit.

1.1.d *Example 2: Regularization and use of the L-curve*

We now add a regularization term $\|\nabla m\|_2^2$ to the misfit function. Assuming we have no a priori knowledge of the noise level in the data, we will perform the inversion with different values of λ and from an L-curve determine the best regularization parameter. The y-axis of the L-curve in Figure 4.3e corresponds to the final data misfit, and the x-axis of the L-curve is the value of $\|\nabla m\|_2^2$ for the final model. In Figure 4.3a - 4.3d we show the final velocity models using decreasing values of λ. Figure 4.3a corresponds to the highest value of λ, giving more relative weight to the regularization term, thus providing the smoothest model. Figure 4.3d corresponds to the smallest regularization weight λ and the final velocity model is not smooth and resembles the final velocity model obtained without regularization. The misfits for the final models with different values of λ indeed form an L-curve. Strong regularization weights have high data misfits and correspond to very smooth models, while low regularization weights have low data misfits and models that are not necessarily smooth. In this example, the model in Figure 4.3b is in the corner of the L-curve and provides the best compromise between data fit and smoothness.

1.1.e *Norm for the regularization term*

Choosing the norm for the regularization term depends on the previous information that is available on the solution m. Let us first define the norms.

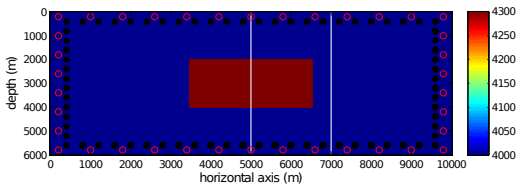


Figure 4.1: True velocity model. 38 sources (circles), 72 receivers (cross marks) per source, around target area. Absorbing boundary conditions around all the area.

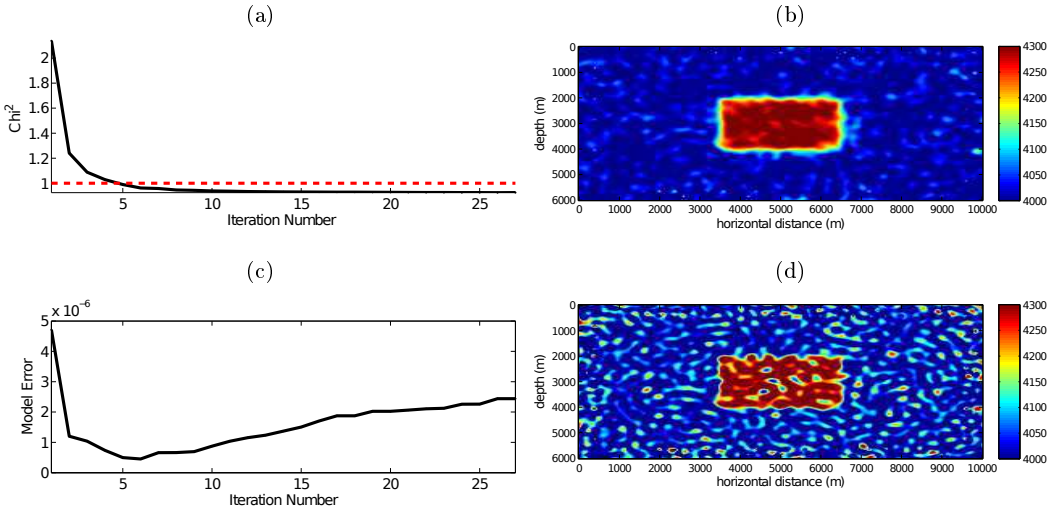


Figure 4.2: Inversion without regularization using one frequency group with frequencies in 2 $Hz-$ 6 Hz, with $1Hz$ interval. a) χ^2 fit versus iteration number. The red dotted line indicates the optimal value of $\chi^2 = 1$. b) Model error $\dfrac{\|m_{true} - m_i\|^2}{\|m_{true}\|^2}$ as a function of the iterations. c) Intermediate model when $\chi^2 = 1$. d) Final velocity model.

In continuous formulation, m is a function of $\vec{x} \in \mathbb{R}^d$, where d is the dimension and for this part we have made explicit that x is a vector. The gradient of the model is also a vector,

$$\nabla m = \begin{pmatrix} \nabla_{x_1} m \\ \nabla_{x_2} m \\ \vdots \\ \nabla_{x_d} m \end{pmatrix}. \tag{4.12}$$

Using as a regularization term the l_2 norm of the gradient of the model in a 2D gives,

$$||Rm||_2^2 \quad = \quad \frac{1}{2}\int_{\mathcal{D}} \|\nabla m\|_2^2 d\vec{x} = \int_{\mathcal{D}} \left((\nabla_{x_1} m)^2 + (\nabla_{x_2} m)^2\right) d\vec{x}. \tag{4.13}$$

☛ When imposing this l_2 norm in the regularization term, we require that the function $\nabla m \in L^2(\mathcal{D})$. Functions of at least two variables where $\nabla m \in L^2(\mathcal{D})$ do not admit for m to have discontinuities along curves, only at points. Thus when imposing regularization term (4.13), we are using the hypothesis that the model or function m we are seeking to reconstruct is smooth, and has no discontinuities except for some points. However, it may not always be true that we wish to reconstruct a continuous model, but on the contrary a model with discontinuities. The question arises as to which is the norm that best describes the model.

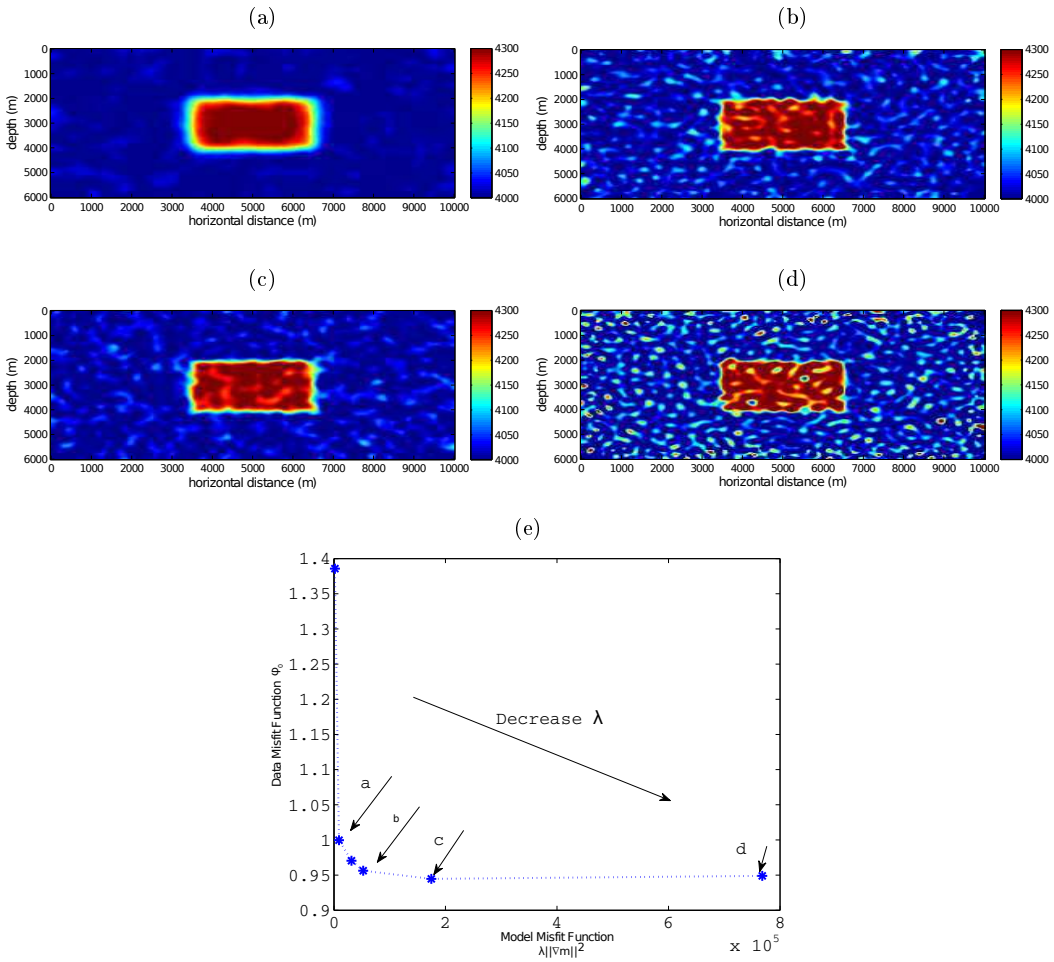


Figure 4.3: Final models with regularization $\lambda\|\nabla m\|^2$. Decreasing values of λ from Figure a) - d). e) L-curve, showing the optimal value close to Figure b).

To get a graphical intuition of the effect of the norm on the regularization term, suppose we are using a regularization term $\int_\Omega |\nabla m|^p dx$. Since the inverse problem is ill posed, it is possible that the two very distinct models in Figure 4.4a and Figure 4.4b fit the data equally well. It will be the model with minimum regularization energy that will be chosen. For the model in Figure 4.4a, $\int_\Omega |\nabla m|^2 dx \approx 5\,(50)^p$. The regularization energy for the model in Figure 4.4b will be, $\int_\Omega |\nabla m|^2 dx \approx 250^p$. Therefore, if $p > 1$, the chosen model will be the smooth model in Figure 4.4a. On the other hand, if $p < 1$, the model with minimum energy will be Figure 4.4b. Note that even though both models have the same total change (both change form 0 to 250), they do not change in the same way. The smooth model makes small changes of 50 for each position and model 4.4b is *piece-wise constant* and has just an abrupt change in one position. In other words, 4.4b is a *sparse model*. For this particular example, both models provide equal energy for $p = 1$.

1.1.f *Norm for the data*

Conventional variational methods to solve *linear* inverse problems involve a least squares l_2 misfit function because this leads to solving linear equations, the previously mentioned normal

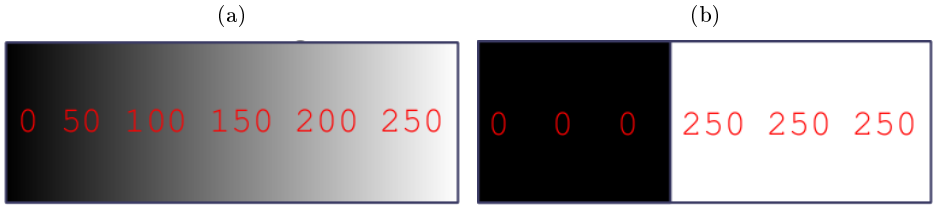


Figure 4.4: Figure from Wittman (2012). Schematic comparison of regularizations of the form $\int_\Omega |\nabla m|^p dx$. a) $|\nabla m| = 50, 50, 50, 50, 50$. $\int |\nabla m|^p \approx 5\,(50)^p$. b) $|\nabla m| = 0, 0, 250, 0, 0$. $\int |\nabla m| \approx 250^p$. For this example, if $p > 1$ smooth models will have the smallest regularization energy. On the other hand, if $p \leq 1$, models with a few changes (sparse), will have the minimum energy.

equations. In addition, the l_2 norm is also frequently used to minimize the misfit in the data because of the statistical argument that the least squares estimator is the best over the ensemble of possible answers. That is, the least squares corresponds to the maximum likelihood estimator, if the experimental errors have a normal distribution. For *non-linear* inverse problems, we still assume the errors in the data are normally distributed, so the least squares is still the maximum likelihood estimator. The difference lies in the fact that for non-linear inverse problems, the least square problems do not have a closed form solution and are solved by general iterative optimization methods, in which the solution is successively improved by small increments in each iteration. In other words, in each iteration the model increment is assumed to be small, $m_{k+1} = m_k + \Delta m$, so that the objective function for m_{k+1} can be correctly approximated by a second order Taylor expansion. This procedure amounts to linearizing the inverse problem around m_k. In either case (linear or non linear inverse problem), if the distribution of the errors in the observed data is not Gaussian, but for example has outliers, the least squares estimations is not the best estimator, since the l_2 norm of the outliers is large. In this case, some other norms may be used (Djikpéssé and Tarantola, 1999; Guitton and Symes, 2003; Ha et al., 2009; Pyun et al., 2009; Brossier et al., 2010), to attempt to find a better statistical estimator for the errors in the data. However, note that if the norm of the data is changed, the imaging condition will change. That is, if the data norm is not the l_2 norm, the gradient will not be the expression that was explained in Chapter 2: the source term of the adjoint wavefield will be modified since the residuals are not the same. Therefore, changing the norm of the data has deeper consequences than merely seeking a statistical estimator for the noise.

1.2 Total Variation Regularization

Let us start by defining the Total Variation norm. For continuous functions, the total variation is related to the l_1 norm of the gradient of the function. As we are interested in admitting discontinuous functions, its rigorous definition in the continuos framework requires a weak formulation [1]. As in practice we deal with discrete approximations, we use here the simpler discrete version. For a function $m(x)$, we define

$$||m||_{TV(\Omega)} = \sum_{i=1}^{N_z} \sum_{j=1}^{N_x} |\nabla m_{i,j}| \Delta x \Delta z. \tag{4.14}$$

Here, the gradient is understood to be a discrete approximation of the true gradient operator. TV is particularly appropriate for recovering blocky or possibly discontinuous functions

1. Here we define the TV norm in a discrete setting, useful when solving a discrete approximation of the definition of TV. In a continuous setting for a function $m \in L^1(\Omega)$,$\Omega \subset \mathbb{R}^2$, the total variation of m is defined by $TV(m) = \sup \left\{ \int_\Omega m(x) \nabla \zeta(x) dx \ : \ \zeta \in C_c^1(\Omega; \mathbb{R}^2),\ |\zeta(x)| \leq 1 \ \forall x \in \Omega \right\}$. See Chambolle (2004)

by preserving sharp boundaries, from noisy data. The total variation (TV) quasi norm allows discontinuities in the model (image), and therefore allows to recover the edges of the original model. Smooth regions are approximated by piecewise constant models (Osher et al., 2005; Caselles et al., 2011). The downside is that total variation has a tendency to discard small variations, known as the texture, in the image. The resulting images using this regularization term have a cartoon like appearance, meaning the edges are clear, and the smooth regions are very smooth. This has led to numerous studies proposing modifications of the total variation regularization term (Bertalmio et al., 2003; Vese and Osher, 2003). We use the traditional total variation with the knowledge that although we may reconstruct sharp edges, fine earth model details may be lost.

1.3 TV denoising algorithms for seismic imaging : A modified ROF method.

Denoising methods exploit the fact that there is an inherent regularity in natural images (for example, earth models). Therefore, denoising algorithms must differentiate the noise from the true image information. Rudin et al. (1992) introduced an influential image de-noising method which consists in minimizing the total variation of the image subject to certain constraints related to the noise statistics, which were imposed through Lagrange multipliers. The edge preserving noise removal algorithm is commonly referred to as ROF, which stands for the initials of the authors Rudin, Osher and Fatemi. Using TV as a regularization term in the imaging problem was basically motivated by its ability to recover image discontinuities. The remarkable results obtained in image de-noising with total variation since its introduction (Rudin et al., 1992), have made of this an active research area (Chambolle, 2004; Caselles et al., 2011). Proofs of the existence and uniqueness of the solution for the ROF model have been done (Chambolle and Lions, 1997), and there is a vast amount of literature on optimization schemes and numerical schemes to solve this problem (Vogel and Oman, 1996; Chambolle and Lions, 1997; Vogel, 2002; Chambolle, 2004; Osher et al., 2005).

The original ROF variational problem was originally stated as an image denoising problem. Let $\mathcal{D}$ be the image domain, $m(x)$ be the ideal undistorted image represented as $m(x) : \mathcal{D} \to \mathbb{R}$, let $\hat{m}(x)$ denote the noisy observed image where $\hat{m}(x) : \mathcal{D} \to \mathbb{R}$, and let $f(x)$ denote a blurring kernel such that $f(x) : \mathcal{D} \to \mathbb{R}$, and ξ some additive white noise with zero mean and variance σ'^2. *We remark that here, we are considering the noise acting directly on the image and not on the data as before.* In practice, if we are considering a 2D image, $\mathcal{D} = \mathbb{R}^2$. The observed measured image is,

$$\hat{m} = f * m + \xi. \tag{4.15}$$

The denoising problem consists in recovering u from f, which amounts to a linear ill posed inverse problem. In particular, for the denoising case, we take f to be the Dirac delta.

One possibility to solve (4.15), consists in adding an L_2 norm regularization, keeping in mind that this restricts the solutions to those that do not have discontinuities along measurable curves. This was the motivation to introduce the TV norm in the regularization term, where the hypothesis we are making when introducing a TV regularization term is that the functions with bounded variation are a reasonable functions for the images (in our case of earth models) we are seeking to reconstruct. Functions with bounded variation may have discontinuities along rectifiable curves. The discontinuities along curves may be identified with edges. The capability of total variation regularization to recover edges that we have highlighted in the previous subsection is one of the main characteristics to assume that this is the appropriate regularization. The original ROF method (Rudin et al., 1992) was stated as a constrained optimization problem,

$$\min_m ||m||_{TV(\Omega)} \quad \text{such that} \quad ||\hat{m} - m||_2^2 = \sigma'^2 \tag{4.16}$$

Of course, this assumes that we have knowledge on the noise level σ'^2. For smooth images, the total variation of the function is equivalent to the L^1 norm of the derivative, and so can be related to the amount of oscillations in m. As before, the constrained optimization problem (4.16) can be reformulated as an unconstrained problem

$$\min_m \quad ||m||_{TV(\Omega)} + \lambda||\hat{m} - m||_2^2, \tag{4.17}$$

Minimizing (4.17), it is possible to find that the optimality condition is satisfied when [2]

$$\begin{cases} 2\lambda(m - \hat{m}) - \nabla \cdot \left(\dfrac{\nabla m}{|\nabla m|} \right) = 0 & \text{in } \mathcal{D} \\ \dfrac{\partial}{\partial \hat{n}} m = 0 & \text{on } \partial\mathcal{D}. \end{cases}$$

The second condition, a Neumann boundary condition, is satisfied when dealing with isolated systems. The first equation is the gradient of the objective functional.

Note that the first term in (4.17) is convex, the second is strictly convex, and therefore the sum is convex. Thus, we are dealing with a case of *convex optimization*. However, the term $\nabla \cdot \left(\dfrac{\nabla m}{|\nabla m|} \right)$ has a singularity at $\nabla m = 0$. To deal with the singularity a small factor may be included in the denominator,

$$2\lambda(m - \hat{m}) - \nabla \cdot \left(\frac{\nabla m}{|\nabla m| + \delta} \right) = 0 \tag{4.18}$$

However, if δ is taken too small the gradient may be almost singular and the convergence rate will be slowed down. On the other hand, if $\delta > |\nabla m|$, the edge detection will not be made and the gradient will resemble a standard smoothing constraint, $\nabla^2 m$. Therefore, with an inappropriate choice of δ, the optimization will be either inefficient or unstable. In addition, $\nabla \cdot \left(\dfrac{\nabla m}{|\nabla m|} \right)$ is highly non linear, because of the dependence on $|\nabla m|$ in the denominator. These two difficulties have given birth to new optimization methods specific to problems the TV norm (Osher et al., 2005), like the iterative lagged scheme (Vogel and Oman, 1996; Vogel, 2002), primal-dual methods (Chambolle and Lions, 1997; Chambolle, 2004).

As pointed out before, one of the disadvantages of TV denoising is that it removes texture and fine details of the image in order to attain a piecewise constant and sparse image, with sharp edges. Alternative solutions have been proposed to avoid loosing the texture and fine details of the image, by performing sequential denoising problems where the denoising weight is lowered (Vese and Osher, 2003).

When denoising algorithms are applied to velocity models on the exploration scale at the end of the inversion, there is the risk that the small details such as thin velocity layers (objects appearing with a small area) will be deleted. In addition, deep structures sometimes have a weak amplitude in the image (compared to the amplitude of shallow structures), and may be also removed in the denoising process. But, ideally, the denoising of velocity models will not remove any important geological characteristic. We therefore modify the denoising process to include the

2. Equation (4.18) can be obtained by applying the Euler-Lagrange equation.

use of the information on the reflectivity provided by other imaging techniques such as migration.

We create a masking matrix that does not apply any denoising in the regions where, according to the migrated image, there are important structures. The modified expression that must be satisfied at the minimum is,

$$2\lambda(m-\hat{m}) - (1-M)\left(\nabla\cdot\left(\frac{\nabla m}{|\nabla m|+\delta}\right)\right) = 0, \tag{4.19}$$

where $M \in \{0,1\}$ is a local thresholding value provided by the migrated image,

$$M(x) = \begin{cases} 1 & \text{if the migrated image identifies a reflector} \\ 0 & \text{otherwise.} \end{cases} \tag{4.20}$$

It could be interesting to reformulate this modified algorithm as minimization problem as in (4.17). However, in general, this is not a straightforward due to the fact that M depends on x.

1.4 Numerical Implementation of regularization and denoising (l_2 and TV norms)

Let model m is discretized on a $N_z \times N_x$ grid, $m \in \mathbb{R}^{N_z} \times \mathbb{R}^{N_x}$ using Neumann boudary conditions $\partial m/\partial\hat{n} = 0$. The model parameter $m_{i,j}$, $1 \le i \le N_z, 1 \le j \le N_z$, will be updated through the component (i,j) of the gradient. The regularization term in the misfit function using the l_2 norm will be

$$\|\nabla(m)\|_2^2 = \sum_{i=1}^{N_z}\sum_{j=1}^{N_x}|\nabla m_{i,j}|^2 \tag{4.21}$$

$$= \sum_{i=1}^{N_z}\sum_{j=1}^{N_x}(D_x m_{i,j})^2 + (D_z m_{i,j})^2\,, \tag{4.22}$$

where D_x, D_z are discrete derivative operators. For the optimization, we will only use gradient or quasi-Newton methods, thus we only need the derivative of the regularization term.

$$g_R = \frac{\partial R(m)}{\partial m} = \nabla^2 m \tag{4.23}$$

$$= D_{xx}m_{i,j} + D_{zz}m_{i,j}. \tag{4.24}$$

The TV norm regularization term in discrete form is,

$$TV(m) = \sum_{i=1}^{N_z}\sum_{j=1}^{N_x}|\nabla m_{i,j}| \tag{4.25}$$

$$= \sum_{i=1}^{N_z}\sum_{j=1}^{N_x}\sqrt{(D_x m_{i,j})^2 + (D_z m_{i,j})^2} \tag{4.26}$$

The gradient of the TV is,

$$g_R = \nabla\cdot\frac{\nabla m}{\|\nabla m\|_1} = \frac{D_{xx}m_{i,j}(D_z m_{i,j})^2 - 2D_x m_{i,j}D_z m_{i,j}D_{zx}m_{i,j} + D_{zz}m_{i,j}(D_x m_{i,j})^2}{((D_z m_{i,j})^2 + (D_x m_{i,j})^2 + \delta)^{3/2}}, \tag{4.27}$$

where a stability parameter δ has been added in the denominator to avoid division by zero.

The discretizations of the derivatives used are (other choices of discretizations are possible):

$$D_z m_{i,j} = \frac{m_{i+1,j} - m_{i-1,j}}{2h} \tag{4.28}$$

$$D_x m_{i,j} = \frac{m_{i,j+1} - m_{1,j-1}}{2h} \tag{4.29}$$

$$D_{zz} m_{i,j} = \frac{m_{i+1,j} - 2m_{i,j} + m_{i-1,j}}{h^2} \tag{4.30}$$

$$D_{xx} m_{i,j} = \frac{m_{i,j+1} - 2m_{i,j} + m_{i,j-1}}{h^2} \tag{4.31}$$

$$D_{zx} m_{i,j} = \frac{m_{i+1,j+1} + m_{i-1,j-1} - m_{i+1,j-1} - m_{i-1,j+1}}{4h^2}, \tag{4.32}$$

where h is a uniform discretization length. For example, with this discretization, the TV norm regularization term is

$$TV(m) = \frac{1}{2h} \sum_{i=2}^{N_z-1} \sum_{j=2}^{N_x-1} \sqrt{(m_{i+1,j} - m_{i-1,j})^2 + (m_{i,j+1} - m_{i,j-1})^2}. \tag{4.33}$$

1.4.a *TV Regularization : numerical examples*

We will now compare the difference between applying the TV norm ($\|\nabla m\|_1$) and the l_2 norm ($\|\nabla m\|_2^2$) as regularization term in the inversion.

Rectangular inclusion numerical test

Consider the true velocity model shown in Figure 4.5a. The data contains five percent additive Gaussian noise. Using the initial velocity model shown in Figure 4.5b, a 2D isotropic acoustic inversion in the frequency domain is performed using l-BFGS. There are 38 sources and 72 receivers per source around the target area. The inversion stops when it reaches a maximum of 30. The final velocity model without regularization is shown in Figure 4.6a. We compare the outcome of using the l_2 and the TV regularizations where the parameter λ is obtain fusing an *L-curve*. Using the l_2 norm, the best model obtained is shown in Figure 4.6b. The best model obtained using the TV norm is in Figure 4.6c. A vertical velocity log at $x = 5000\ m$ is shown in Figure 4.6d. The black line corresponds to the true velocity, the green line is the initial velocity. The red line is the velocity model obtained with the l_2 regularization term, and the blue line is the velocity model obtained with the TV norm. The comparison of the logs brings out the differences between the regularization terms. Both regularization terms provide satisfactory final velocity models. The l_2 norm provides a smooth transition between of velocities. On the other hand, the TV norm provides a smooth model, but does not smooth out the edges and allows the reconstruction of a function with abrupt changes. This velocity model is ideal to use with a TV regularization term, as it is clearly piecewise constant.

The regularization gradients at different iterations using the l_2 are shown in Figure 4.7. The magnitude of the gradient is large whenever the model is not smooth. This is to be expected, because the gradient is the laplacian of the model $\nabla^2 m$, which provides non zero values when there are spatial variations in the model $m(x)$. Therefore, the largest values of the gradients are at the borders, where the value of the model is changing. The regularizing gradients at different iterations using the TV norm are shown in Figure 4.8. The regularizing gradient smooths the regions where there are no abrupt changes, and does nothing to the regions of the model where abrupt changes are present. In Figure 4.8c-d, where the box model has already started to form, the magnitude of the gradient around the borders is smallest (in absolute value). Since no

smoothing is applied to the borders, the final velocity model has sharp velocity contrasts.

Before choosing the best models obtained with each regularization term shown in Figure 4.6, several values of regularization weights λ had to be tried out, to reconstruct an L-curve. In Figure 4.9, the first row shows the final velocity models using the l_2 norm with an increasing value of the regularization weight λ from left to right. The second row shows the final velocity models using the TV norm with an increasing value of the regularization weight λ. As discussed earlier in the description of the L-curve, when the regularization weight is too small the inversion fits the noise. When the regularization weight is too large, the inversion does not fit the data to a full extent and the final model retains the imprint of the initial model with the constraint of the regularization term.

BP-2004 Salt model numerical test

We now move to the BP-2004 salt model shown in Figure 4.10 a), and attempt to reconstruct it from the initial velocity model in Figure 4.10 b). There are 62 sources and 248 receivers per source at 250 m depth. A free surface boundary condition is used on top, and absorbing boundary conditions are implemented with PMLS on the other borders. A 2D acoustic frequency FWI is performed using l-BFGS. Eight frequencies are used sequentially in the range from $2Hz-9Hz$, with $1Hz$ interval. When no noise is added to the observed data, the final velocity model without regularization is shown in Figure 4.11a. The second row of Figure 4.11 shows the final models using a regularization term with the l_2 norm with increasing values of λ from left to right. The third row of Figure 4.11 contains the final velocity models using the TV norm regularization term, for increasing values of λ from left to right. The L-curve is not plotted, but the models in the middle column, Figures 4.11 c and 4.11 f, provide the best trade-off. For these two models, vertical velocity logs at three different positions are shown in Figure 4.12. The true velocity is black, the initial velocity is green, the velocity using an l_2 regularization term is in red, and velocity using TV is in blue. For shallow depths ($z < 1000m$) the TV norm provides a smoother solution, with very few variations. The velocity anomalies in the true model that reach $4500m/s$ are reconstructed almost equally using either regularization term. For the deep parts of the model ($z > 2000\ m$) the velocity reconstruction is similar, although the TV norm has a less oscillating behaviour and shows a much better contrast at the boundary of the salt structure. If we would have liked to achieve very smooth solutions in the shallow part of the model using the l_2 norm regularization term, the amplitude of the deep anomalies would not have been satisfactory.

When adding 25% of Gaussian noise to the data, the final velocity model without any regularization is shown in Figure 4.13 a. The second row of Figure 4.13 shows the final models using a regularization term with the l_2 norm with increasing values of λ from left to right. The third row of Figure 4.13 contains the final velocity models using the TV norm regularization term, for increasing values of λ from left to right. The L-curve is not plotted, but the models in the middle column, Figures 4.13 c and 4.13 f, provide the best trade-off. For these two models, vertical velocity logs at three different positions are shown in Figure 4.14. The true velocity is black, the initial velocity is green, the velocity using an l_2 regularization term is in red, and velocity using TV is in blue. The conclusions are similar to the case without noise, although the TV norm does not provide solutions as flat as for the case without noise. For shallow depths, there are oscillations in the both velocity models. Nonetheless, the logs confirm that TV norm still provides piece wise constant constant models, with a smoother behaviour than those provided by the l_2 norm.

Valhall OBC data

Using one line of the OBC Valhall data set, we wish to compare the differences in the final models provided with the l_2 norm and TV regularization terms. The data consists of 320 sources located $5m$ below the water lever, with a spacing $\Delta_s = 50m$. On the ocean bottom there are 210 hydrophone receivers with a spacing $\Delta_r = 50m$. For a more detailed description and other results see Chapter 3. We perform an anisotropic acoustic 2D inversion in the frequency domain, using l-BFGS as optimization algorithm. The initial models of the vertical wavespeed, Thomsen parameters (δ and ϵ) provided by BP are shown in Gholami et al. (2013a). The background density model is inferred from the initial vertical wave-speed model by Gardner's law and the quality factor is fixed at a constant value of 200. We perform a mono-parameter FWI for the vertical velocity v_{P0} keeping the Thomsen's parameters δ and ϵ, the density and the quality factor fixed. We use three overlapping frequency groups, ranging from 3.5Hz to 5.25Hz: [3.5, 3.78, 4], [4, 4.3, 4.76], [4.76, 5, 5.25] Hz. *The only stopping criteria used is when the inversion reaches 30 iterations per frequency group, or if the line search fails to find a direction to update the model.*

The final velocity models using l_2 norm are shown in Figure 4.15a - 4.15c. The regularization weight is slowly increased from 4.15a to 4.15c. When the TV norm is used, the final velocity models are shown in Figure 4.15d - 4.15f. The regularization weight is slowly increased from 4.15d to 4.15f. When the regularization factor is small (Figures 4.15a and 4.15d), the final velocity models are relatively similar which is to be expected as the regularization term has a relative less important weight. However note that with a strong regularization weight, the deep structures are less pronounced, thus loosing precision. The models in the middle (Figures 4.15b and 4.15e) provide the best compromise in explaining the data and satisfying the regularity constraint. The differences in the velocity models can be seen more clearly through the vertical velocity logs. Figures 4.16a - 4.16c plot the velocity logs for the models in Figure 4.15a - 4.15c, obtained using the l_2 norm. Below, in Figures 4.16d - 4.16f, the velocity logs for the models found using TV (Figure 4.15d - 4.15f) are plotted. The velocity logs with the strongest TV constraint (Figure 4.16f) has a step-like behaviour. As the TV regularization weight is decreased, this behaviour becomes naturally less pronounced.

The comparison of the logs for preferred velocity model using the l_2 and the TV norm shown in Figures 4.16b - 4.16d, have several differences, particularly in the near surface ($z < 500\ m$) and below 2500 m. However, both have oscillations in the logs. Despite this, in the logs with the TV regularization, the position of the deep reflectors can be inferred more accurately.

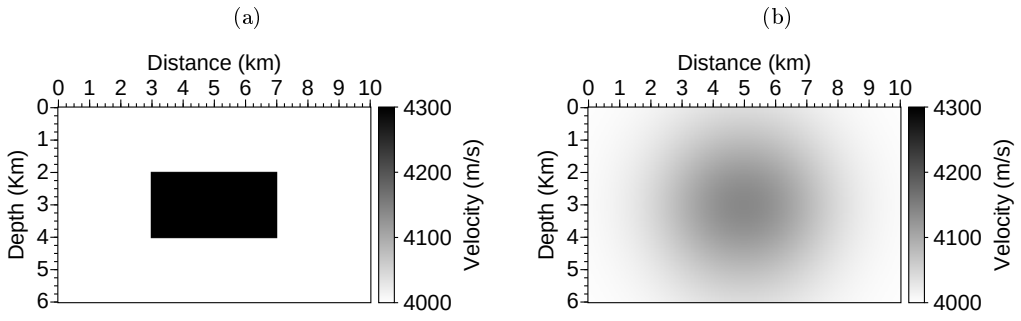


Figure 4.5: a) True velocity model b) Initial velocity model.

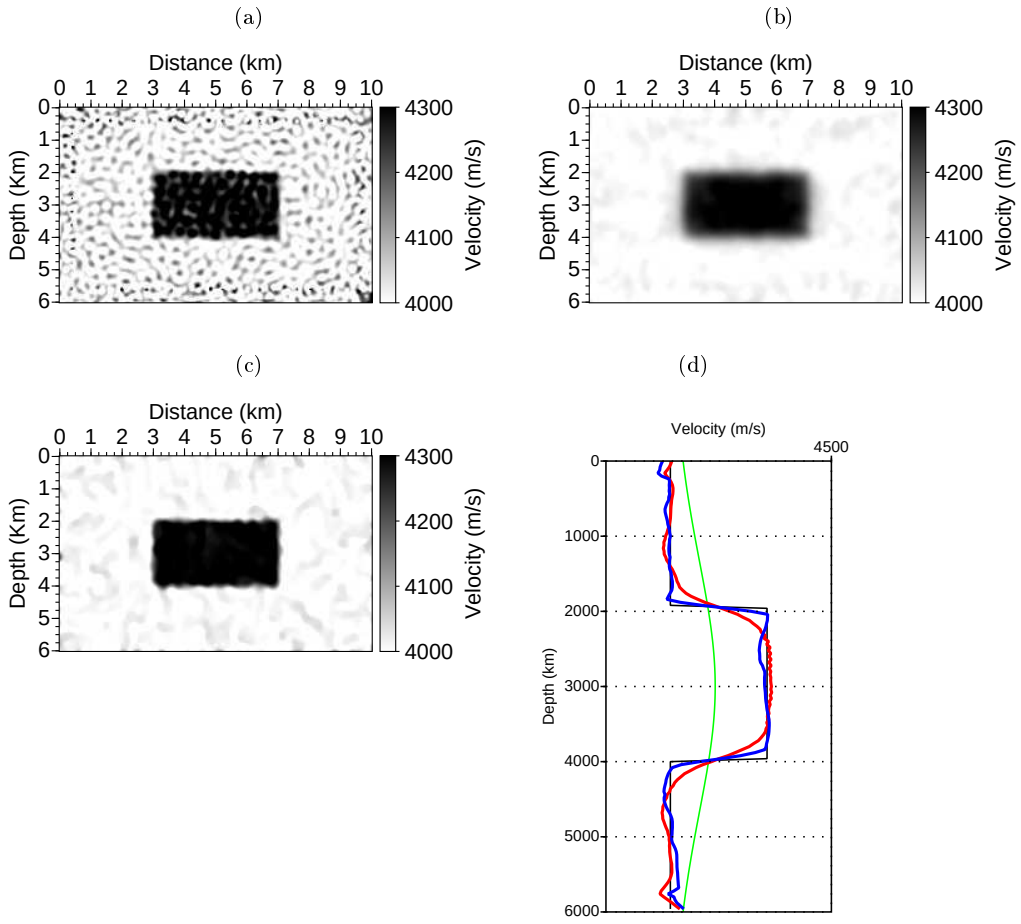


Figure 4.6: a) Final velocity model without regularization b) Final velocity model using $\|\nabla m\|_2^2$. c) Final velocity model using $\|\nabla m\|$. d) Vertical velocity log at $x = 5000\ m$. As expected, the TV norm reconstructs the sharp boundaries.

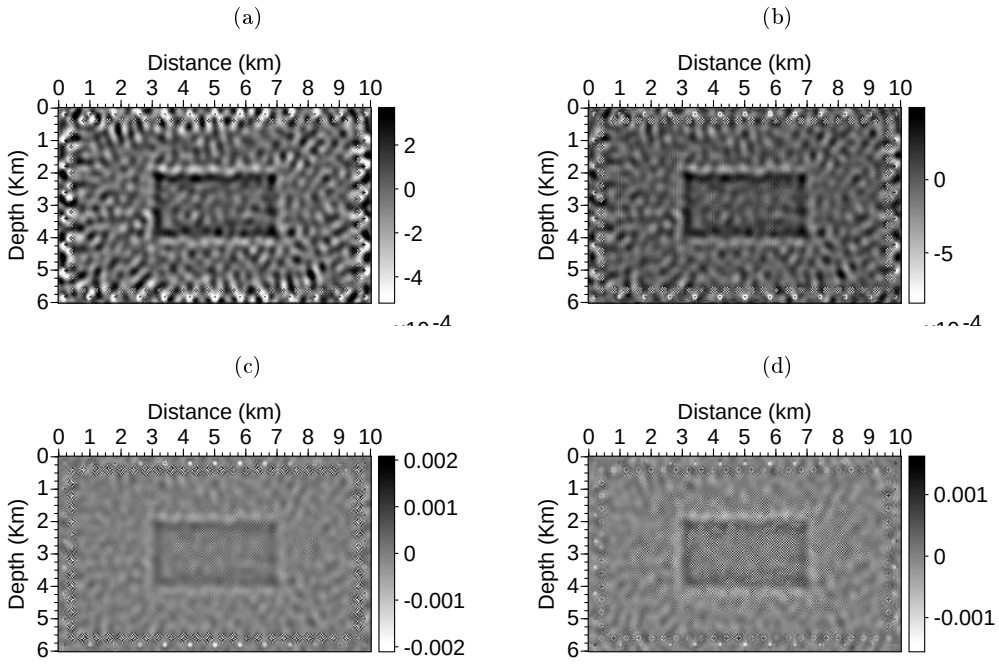


Figure 4.7: Regularization gradients using $\|\nabla m\|_2$ for different iterations. a) iteration 1 b) iteration 2 c) iteration 5 d) iteration 10

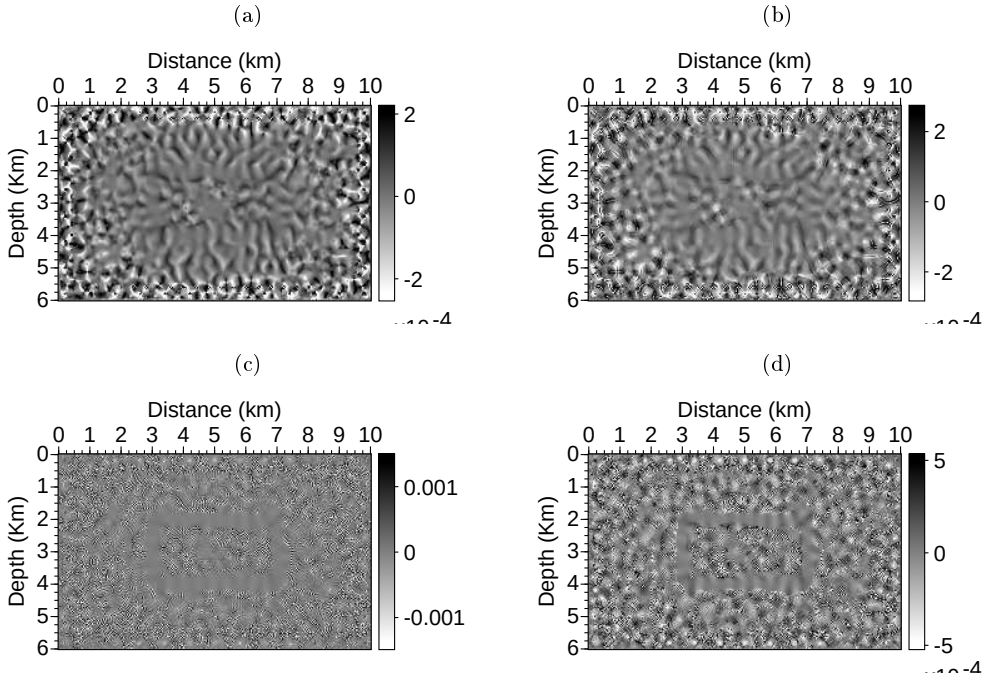


Figure 4.8: Regularization gradient using $\|\nabla m\|_2^2$ for different iterations. a) iteration 1 b) iteration 2 c) iteration 5 d) iteration 10. Notice that in Figure c) and d), the box has already appeared and the gradient is close to zero at the edges.

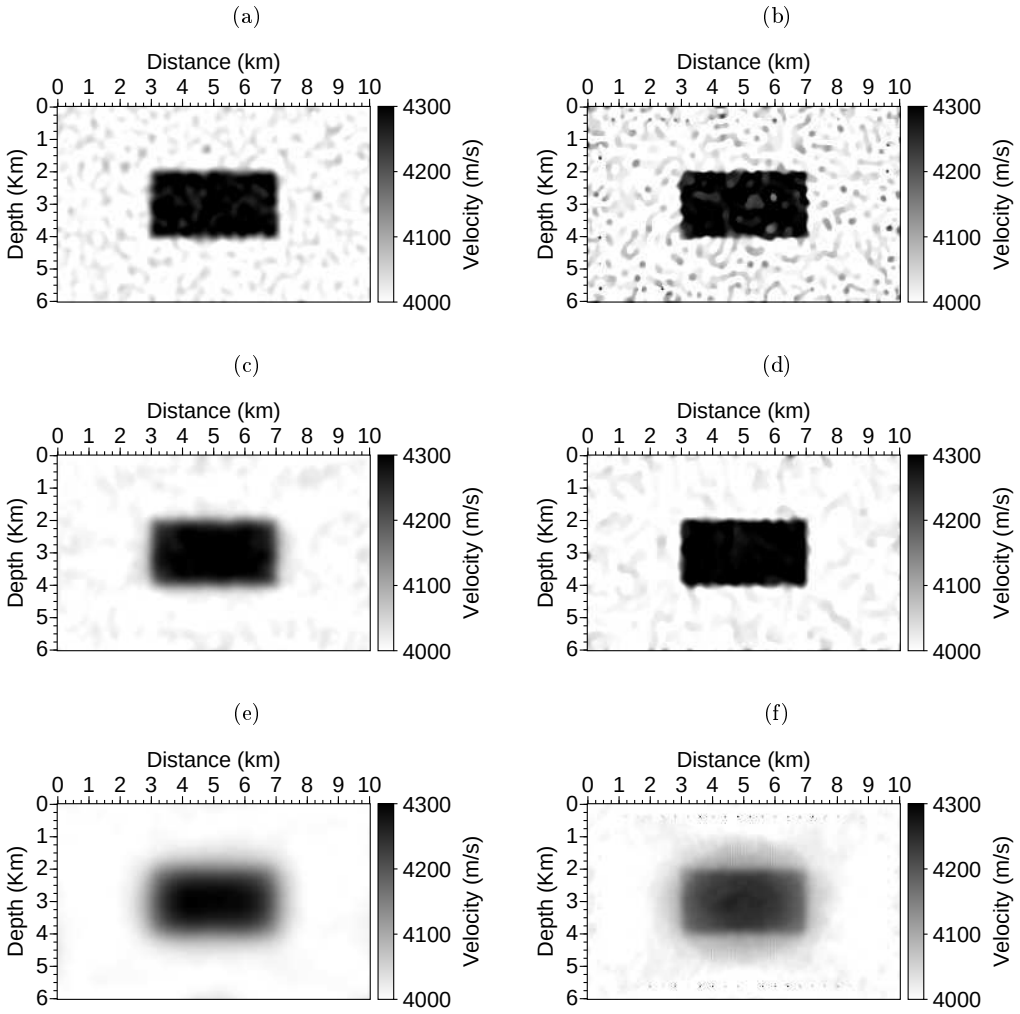


Figure 4.9: Data with 5% of additive Gaussian noise. Comparison of inversion results with $\|\nabla m\|_2^2$ and TV regularization. (a,c,e) : regularization using $\|\nabla m\|_2^2$, with increasing values of λ from left to right. (b,d,f) : regularization using $\|\nabla m\|$, with increasing values of λ from left to right. a) $\lambda = 10^6$ b) $\lambda = 10^7$ c) $\lambda = 10^8$ d) $\lambda = 10^5$ e) $\lambda = 10^6$ f) $\lambda = 5 \cdot 10^6$.

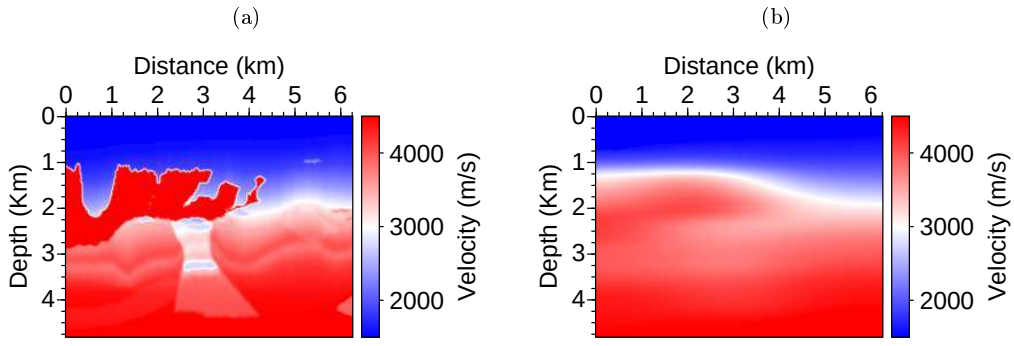


Figure 4.10: a) BP-2004 salt model b) Initial model.

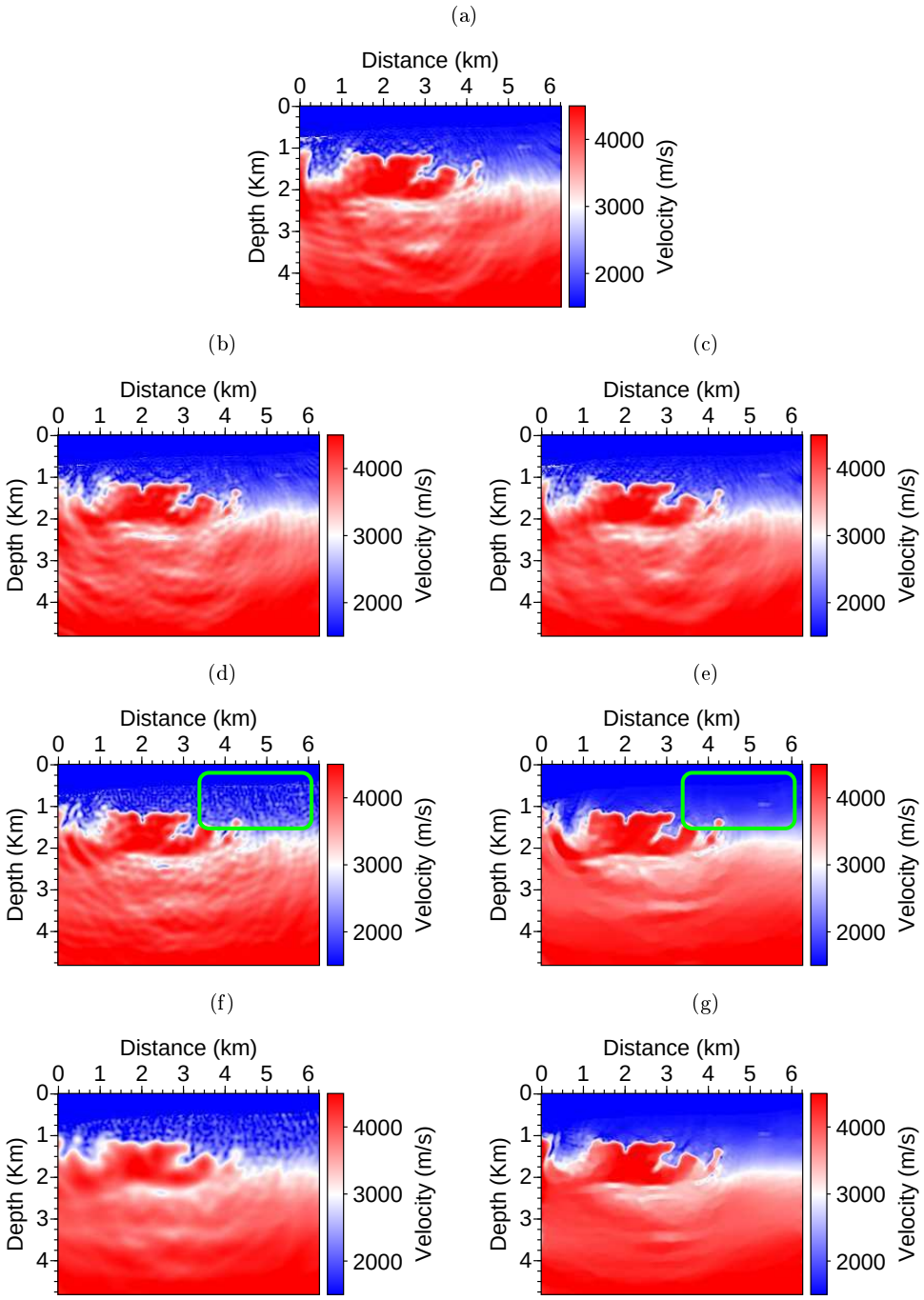


Figure 4.11: Data without noise. Comparison of inversion results with $\|\nabla m\|_2^2$ and TV regularization. a) Final velocity velocity model without regularization. (b,d,f): Inversion with l_2 norm regularization. Increasing regularization factor from up (Figure b) to down (Figure f)). (c,e,g) : Inversion results with TV regularization. Increasing regularization factor from up (Figure c) to down (Figure g)).

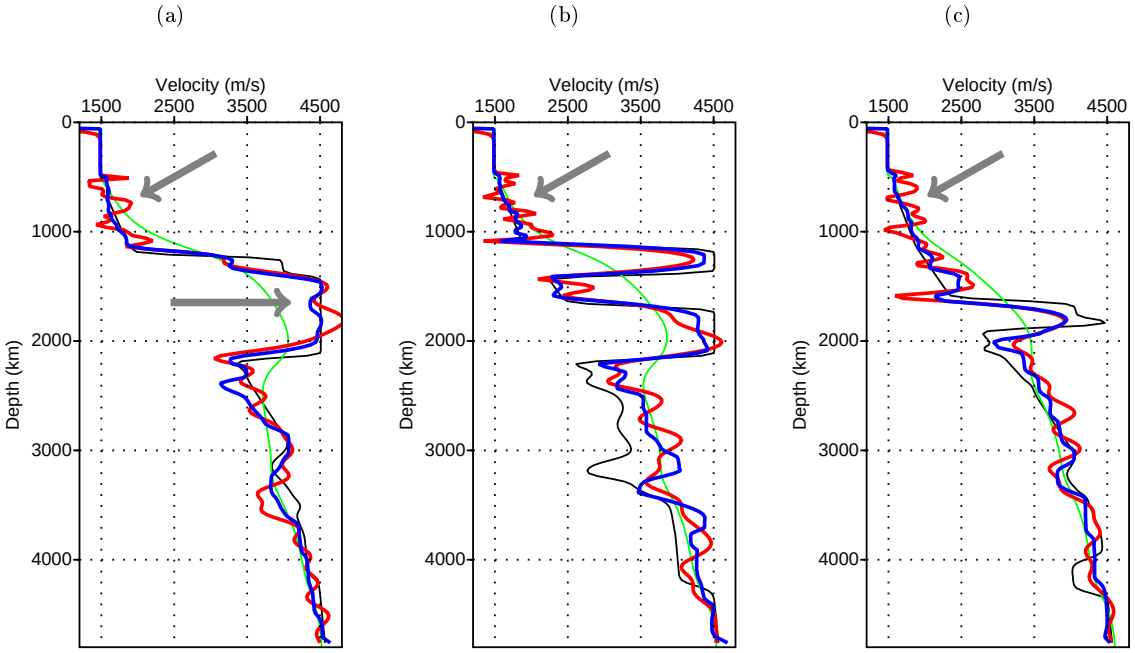


Figure 4.12: Data without noise. Comparison of final model results with $\|\nabla m\|_2$ in Figure 4.11c and with TV regularization in Figure 4.11f. Vertical velocity logs for a) $x = 3200\ m$ b) $x = 4800\ m$ c) $x = 6400\ m$. The true velocity is black, the initial velocity is green, velocity using a regularization term $\|\nabla m\|_2^2$ is in red, and velocity using TV is in blue.

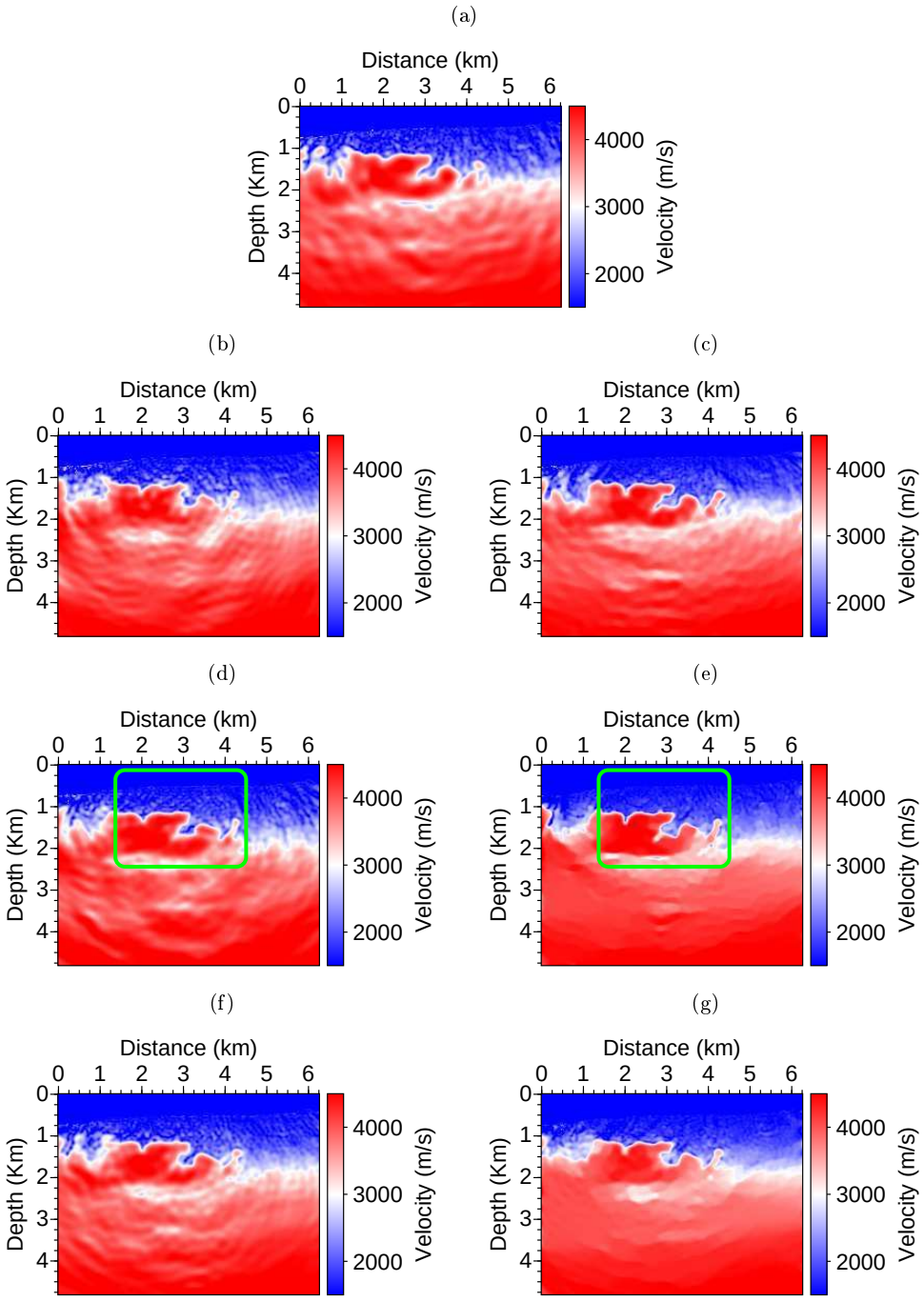


Figure 4.13: Data with 25% of additive Gaussian noise. Comparison of inversion results with $\|\nabla m\|_2$ and TV regularization. a) Final velocity model of the inversion without regularization. (b,d,f): Inversion with l_2 norm regularization. Regularization weight is increasing from up (Figure b) to down (Figure f). (c,e,g): Inversion results with TV regularization. Increasing regularization factor from up (Figure c)) to down (Figure g)).

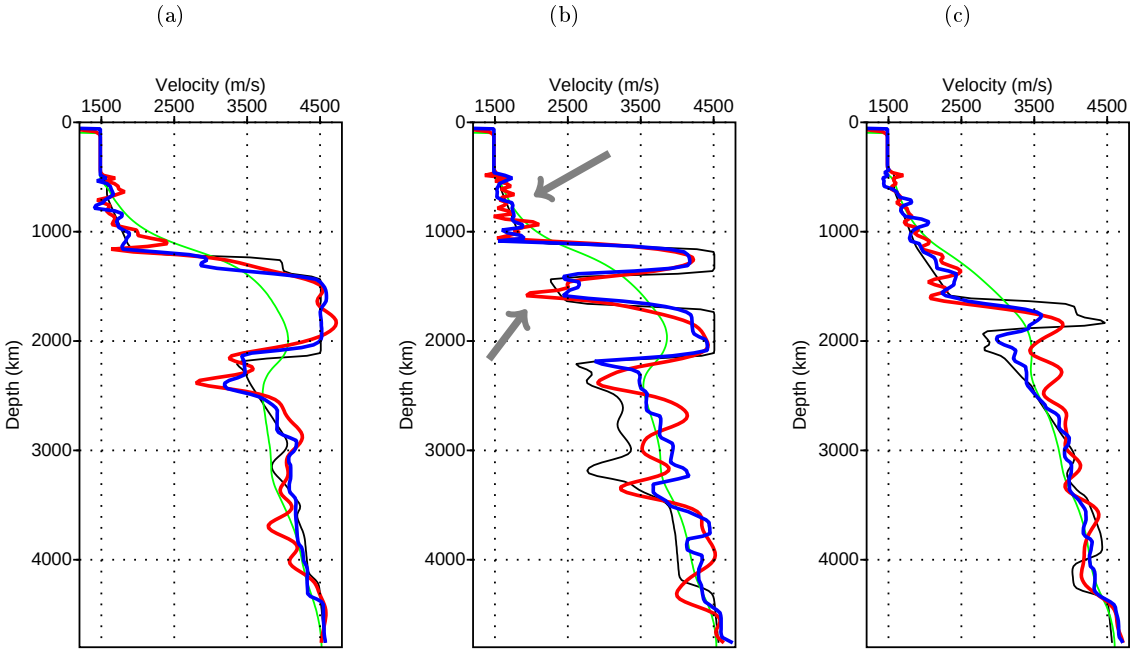


Figure 4.14: Data with 25% noise. Comparison of final modes using l_2 regularization in Figure 4.13c and TV regularization in Figure 4.13f. Vertical velocity logs for a) $x = 3200\ m$ b) $x = 4800\ m$ c) $x = 6400\ m$. The true velocity is black, the initial velocity is green, velocity using a regularization term $\|\nabla m\|_2$ is in red, and velocity using TV is in blue.

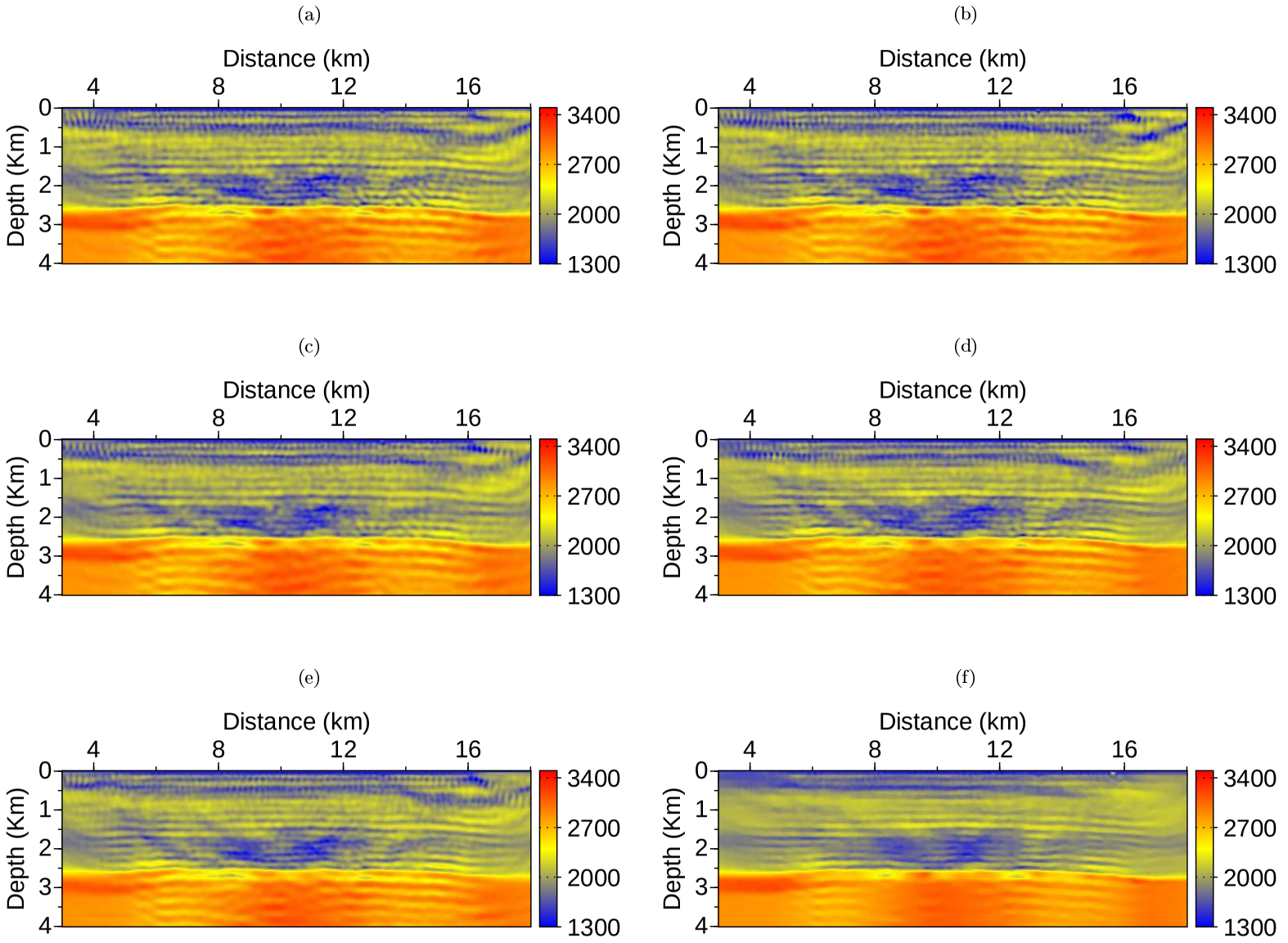


Figure 4.15: Comparison of inversion results with $\|\nabla m\|_2$ and TV regularization. (a,c,e) First row: Inversion with l_2 norm regularization, increasing regularization weight from left to right. (b,d,f) Second row : Inversion results with TV regularization. Increasing regularization factor from left to right.

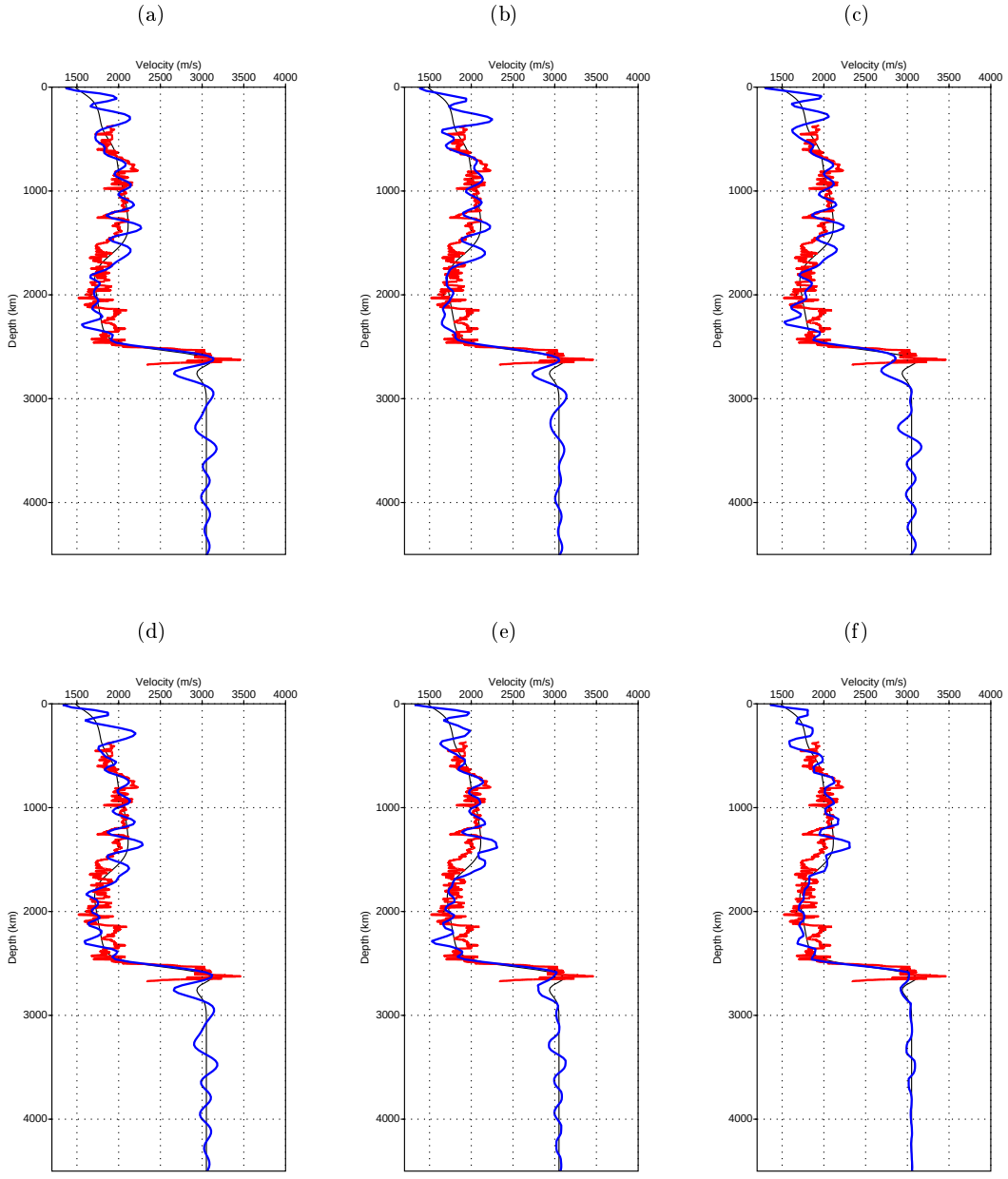


Figure 4.16: Vertical velocity logs at $x = 9.5\ km$. The red line corresponds to a sonic log, and the black line corresponds to the initial model. First row: Figures a) - c) show the velocity logs for the final velocity models in Figure 4.15a - c using l_2 norm regularization, increasing regularization weight from left to right. Second row: Figures d) - f) show the velocity logs for the final velocity models in Figure 4.15d - f using TV norm regularization, increasing regularization weight from left to right.

1.4.b *Denoising : numerical examples*

The ROF denoising algorithm is applied to the final model of FWI, with the purpose of denoising the model while preserving the edges. We use the algorithm in equation (4.18). Let $E(m)$ be the functional (an energy) to be minimized (Chan and Shen, 2005),

$$\min_m E(m) = \min_m \left\{ 2\lambda(m - \hat{m}) - \nabla \cdot \left(\frac{\nabla m}{|\nabla m| + \delta} \right) \right\}. \tag{4.34}$$

To minimize m we can move in the direction of steepest descent,

$$\begin{aligned} m_0 &= \hat{m} \\ m_{t+1} &= m_t + dt \left(2\beta(m_t - \hat{m}) - \nabla \cdot \left(\frac{\nabla m_t}{|\nabla m_t| + \delta} \right) \right), \end{aligned} \tag{4.35}$$

where t is simply an artificial time parameter to iterate. The term $\nabla \cdot \left(\frac{\nabla m_t}{|\nabla m_t| + \delta} \right)$ is discretized in each time step using (4.27). This time marching algorithm (4.35) stops when it reaches a steady state, or a final time T (Chan and Shen, 2005). The value β controls the weight of the fidelity term $m_t - \hat{m}$. For higher the value of lambda, less denoising will be performed because the model m_t is restricted to stay close to $\hat{m}$. When $\beta \to 0$, no fidelity restriction is imposed, and only the total variation of the image is minimized.

Instead of taking the steepest descent direction, there are faster optimization algorithms to minimize the TV energy using for example Newton methods or graph cuts (Chan et al., 2001; Chambolle, 2004; Darbon and Sigelle, 2005; Ng et al., 2007; Chambolle et al., 2011). Amongst the fastest and most popular is perhaps the split Bregman method (Goldstein and Osher, 2009).

However, in our application the running time of the denoising algorithm is meaningless compared to the running time of FWI. For example, for the real data model of Valhall, $N_z = 219, N_x = 837$. On a single processor, the denoising algorithm is of the order of 30 seconds. Therefore, there is no interest in our case to seek to improve the convergence time. Moreover, we can solve many denoising problems, with different values of β without any computational or time limitations.

For the numerical test, we use the model previously found in Figure 4.15a, and consider this as the initial noisy velocity model. In Figure 4.17b we plot the initial velocity perturbation $\hat{m} - m_0$. We fix $\delta = 10^{-2}$, $T = 300$, and apply the denoising algorithm (4.35). The final results are shown in Figure 4.18, for β values decreasing from left to right. Recall that as $\beta \to 0$ more denoising in performed. The velocity model 4.18a is denoised with respect to the original model, but a lot less than 4.18b,c. The denoised models in 4.18b and 4.18c are similar, suggesting that the denoising has reached a steady state. However, notice in the velocity perturbations in 4.18e and 4.18f, that the deep reflector has been partly removed. This is somewhat expected because TV denoising removes texture and small details of the images, to attain the sparsity and sharpness that it provides.

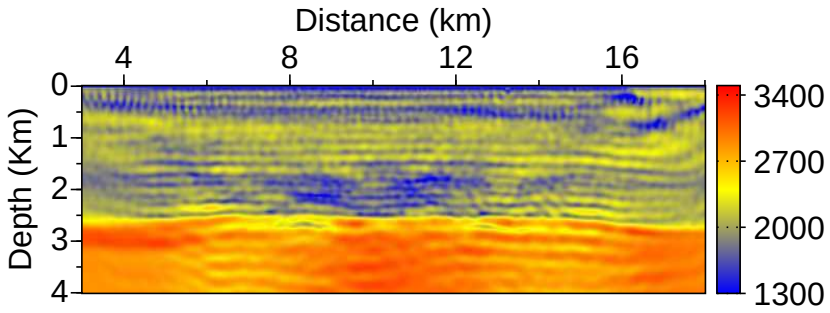
Recall that for the simple case of the inversion of a box model using TV regularization with full acquisition, the gradients (Figure 4.8) revealed that with complete illumination of sources and receivers, the boundaries were perfectly detected and the gradient was zero at the edges. To apply the TV denoising of the FWI velocity models, we can use the migration information in the similar fashion through a thresholding matrix $M(x)$, with the modified algorithm proposed in (4.19).

A migrated image m_{migr} is shown in Figure 4.19a. It is a high frequency image containing positive an negative variations to describe where the reflectors are placed (no the magnitude of the velocity perturbations). Taking the absolute value of the migrated image, we obtain the image in Figure 4.19b, where the regions of the main reflectors can be guessed, but with a little difficulty. To construct the thresholding matrix $M(x)$, we apply a traditional TV ROF denoising algorithm to Figure 4.19b. Finally, we threshold the denoised model 4.19b, and obtain the thresholding operator $M(x)$ in Figure 4.19c. To verify the validity of $M(x)$ and make sure we have not left out any important reflectors in the denoising process, we show $M(x) * m_{migr}$ in Figure 4.19d. Some small reflectors have disappeared but the most important ones remain.

We use 4.19c as $M(x)$ in (4.19) and perform the TV denoising with the same values of β. The final models in Figures 4.20b,c are smooth models but have all the important reflectivity information, as can be seen in 4.20e,f.

4.19 4.20

(a)



(b)

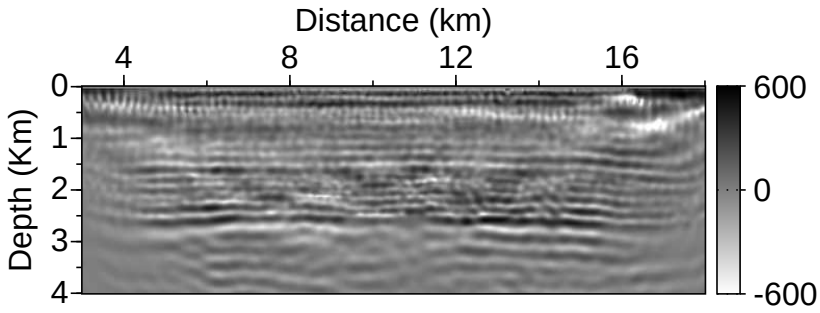


Figure 4.17: a) Original noisy model, $\hat{m}$. b) Velocity perturbation of the noisy model $\hat{m} - m_0$.

1.4.c *Conclusions*

The regularization term in the optimization process plays an important role when solving ill posed inverse problems when the null space is large and when there is noise in the data. As the null space increases and more parameters are unconstrained by the data, the regularization term becomes more important because, for some model parameters, it might be the only restriction that is imposed to them. Similarly, as the noise level in the data increases, so does the level of

indetermination and more sets of model fit the data equally well. The regularization term is used to drive the inversion in the direction of some privileged models, and discard others directions.

We compared the two regularization terms using the l_2 and the TV norm on the synthetic BP-2004 salt model and the Valhall real OBC data set. For the synthetic BP-2004 model we used an initial model that was good enough to avoid cycle skipping, but as far away as possible from the true model so that the regularization had more importance. Due to the reflecting boundary close to the surface, there are some artefacts that appear in that region. For both the case with and without noise, the near surface oscillations are reduced with TV regularization. Without noise, the final velocity model with the TV norm is considerably better. For the real data set application, the final velocity models obtained with the l_2 and TV norm are comparable, with similar overall characteristics. This is mainly due to the fact that in this case, the regularization term is not crucial in the inversion. That is, without any regularization at all, the final velocity is similar. Nonetheless, the step-like behaviour obtained using the TV norm is clear in the velocity logs. In conclusion, our results show that using the TV norm in the regularization term is appropriate for earth models, given that it consists of layered or contrasting surfaces. Although there is not a lot of literature in the comparison of TV as a regularization term, the results here are consistent with the existing literature (Anagaw and Sacchi, 2012; Guitton, 2012).

At the end of the FWI, a TV denoising algorithm was applied. The ROF method removes noise by minimizing the total variation of the model. However, we showed with the Vallhall final velocity model, that the process of denoising may remove some small or deep structures. Therefore, we performed a local TV denoising by incorporating information of the reflectivity provided by a migration image. The modified denoising algorithm was successful in denoising the model but preserving the important reflectors.

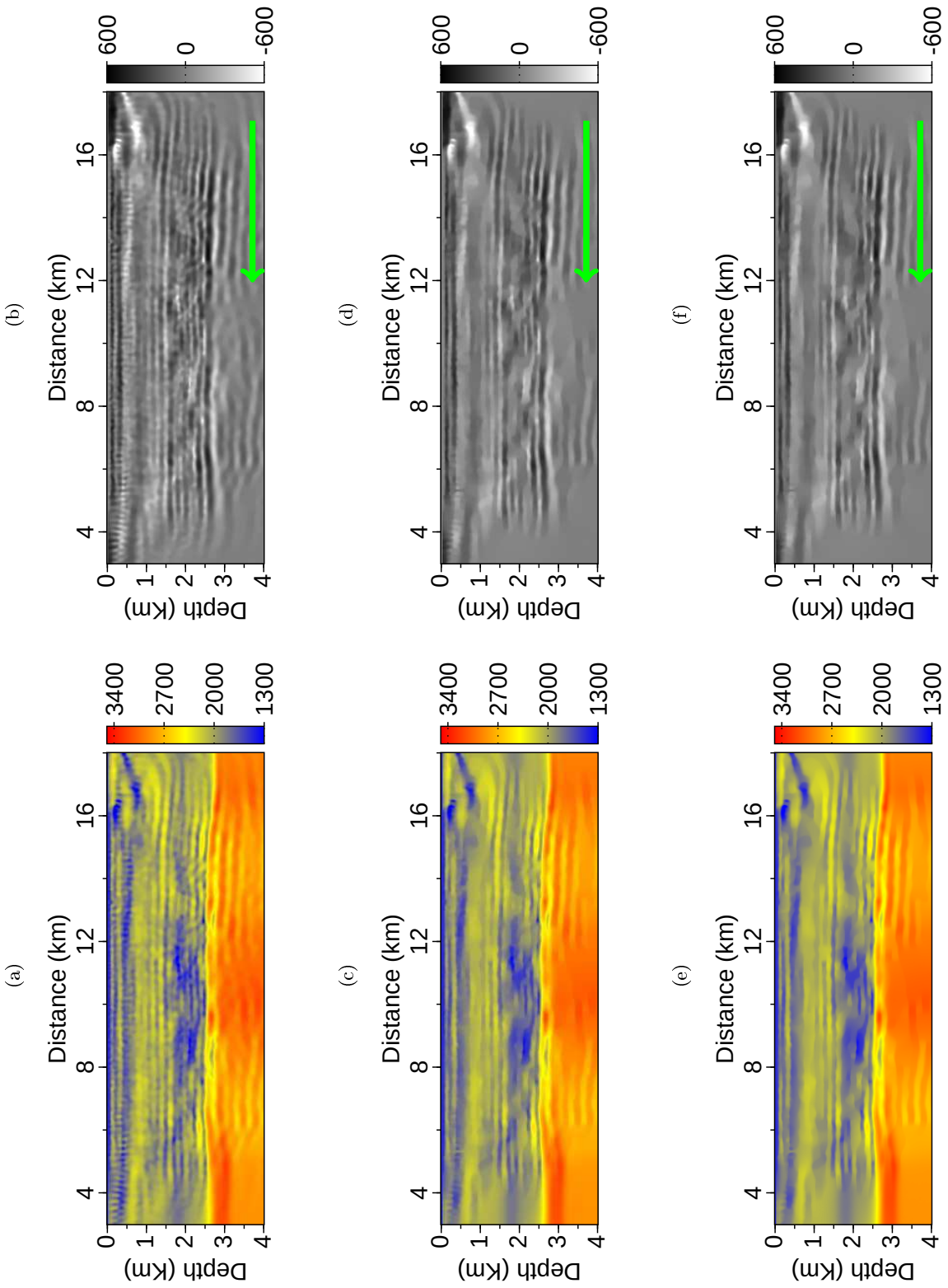


Figure 4.18: Noisy model shown in Figure 4.17. (a,c,e) : final denoised models for decreasing values of λ. (b,d,f) : the denoised velocity perturbations $m - m_0$, for the velocity models in the first row. (a,b) $\lambda = 10^{-2}$ (c,d) $\lambda = 10^{-3}$ (e,f) $\lambda = 10^{-4}$. Note that a big part of the deep reflector has been removed in the denoising procedure in Figures (d) and (f).

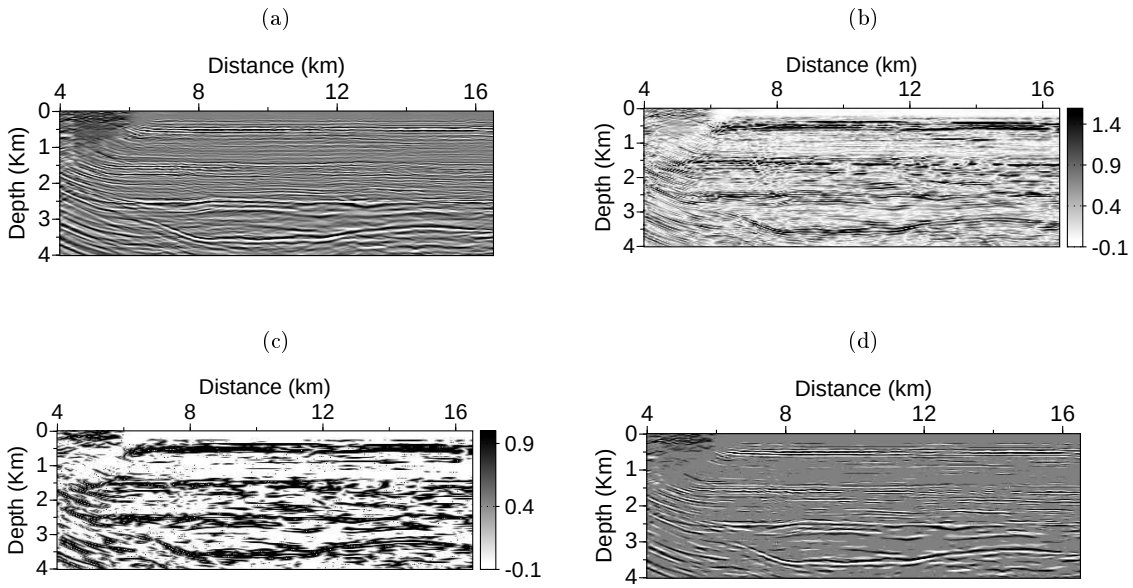


Figure 4.19: a) Migrated image m_{migr} b) Absolute value of the migrated image. c) This matrix corresponds to $M(x)$. The denoised migrated image b) is binarized. d) $m_{migr} * M(x)$ to validation that the matrix $M(x)$ contains all the important reflectors.

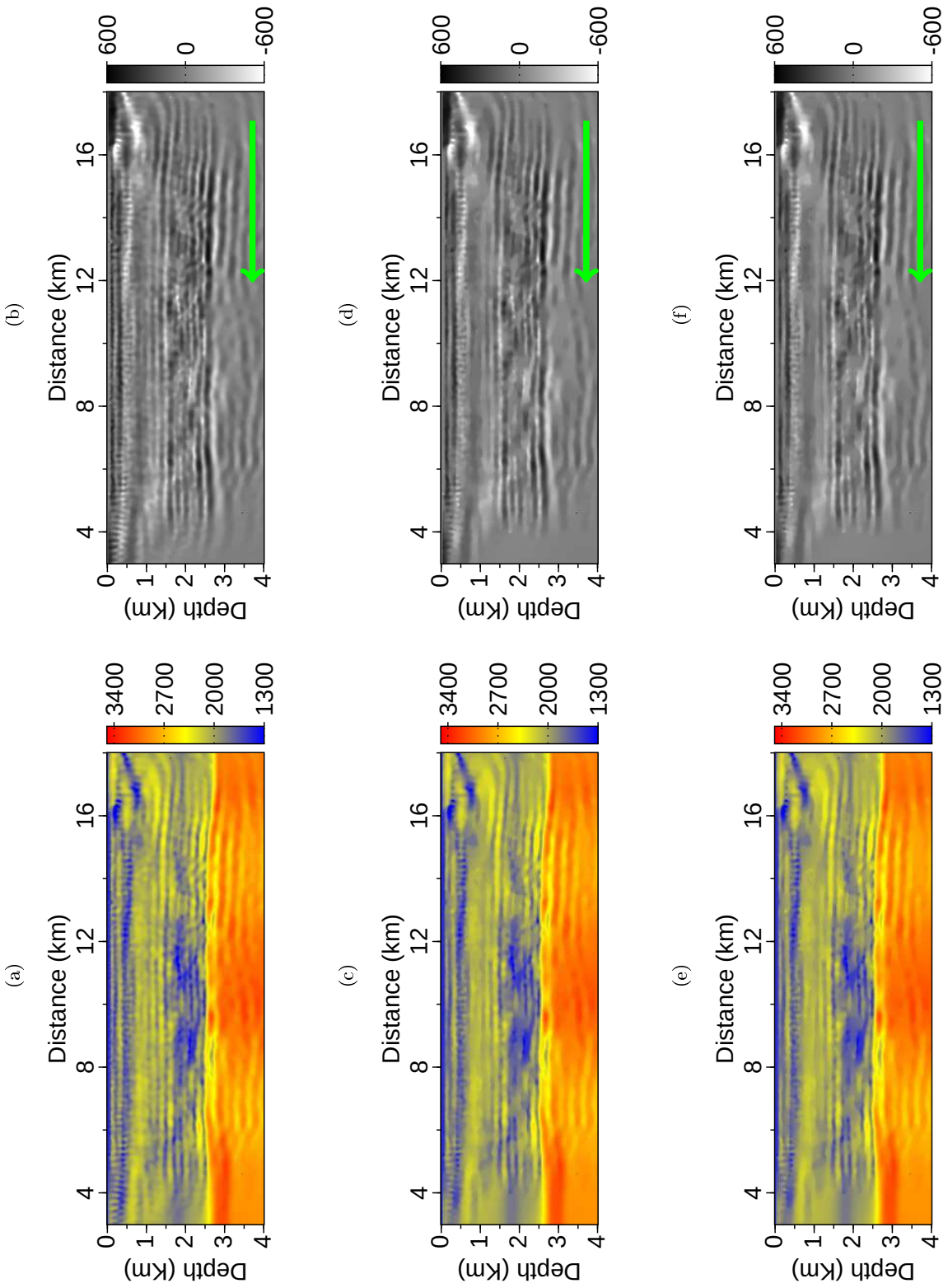


Figure 4.20: Noisy model shown in Figure 4.17. Modified TV denoising. (a,c,e) : final denoised models for decreasing values of β. (b,d,f) : the denoised velocity perturbations $m - m_0$, for the corresponding velocity models. a,b) $\beta = 10^{-2}$ c,d) $\beta = 10^{-3}$ e,f) $\beta = 10^{-4}$. Note that the deep reflector has remained after the denoising procedure in Figures d) and f).

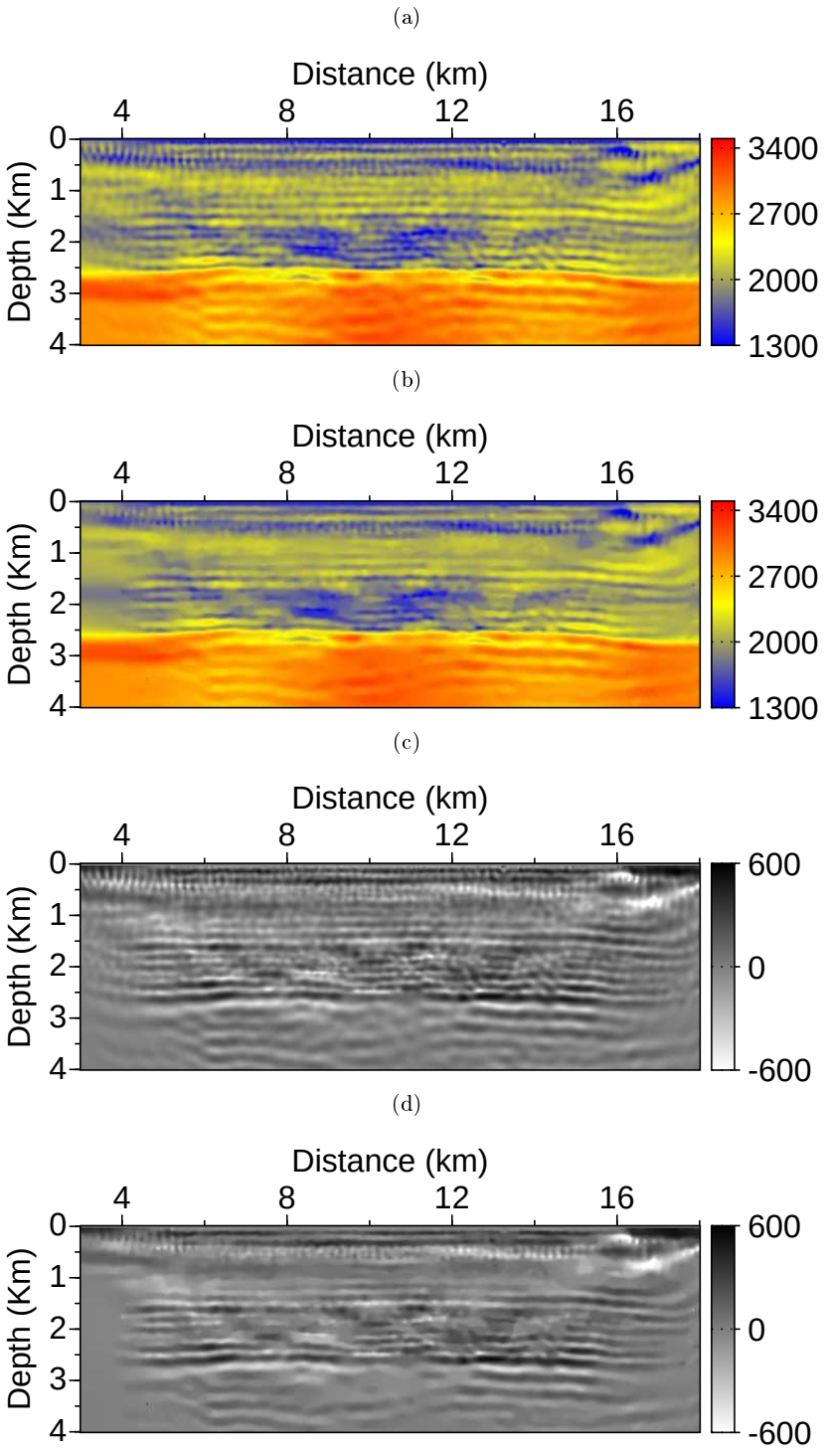


Figure 4.21: Summary of local denoising results : Denoising with thresholding matrix in Figure 4.19. a) Original image b) Denoised image c) Original velocity perturbation. d) Denoised velocity perturbation. d) Vertical velocity log at $x = 9500\ m$. The red line is a sonic log, the blue line is the noisy model and the gray line is the denoised model. The TV denoised model has a better S/N ratio and preserves the information of all the important reflectors. The TV denoised model has a better S/N ratio and preserves the information of all the important reflectors.

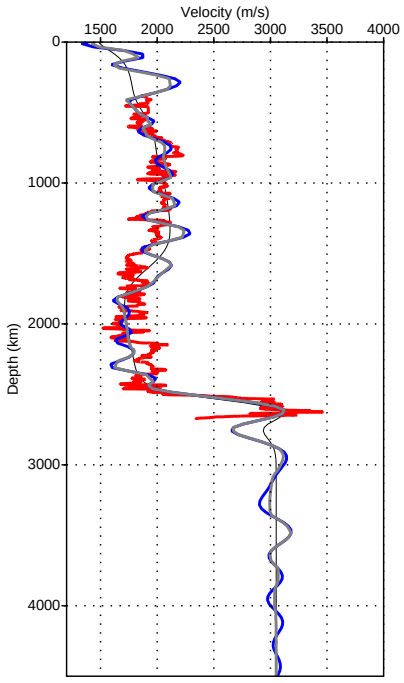


Figure 4.22: Summary of local denoising results : The red line is a sonic log, the blue line is the noisy model (Figure 4.21a) and the gray line is the denoised model (Figure 4.21b).

2 The model null space and the spectral content of the data

If the optimization process is faced with reconstructing model parameters that are in the null space, it is likely the inversion will fail. Regularization techniques introduce a priori model information (unrelated to the data) and may help, but still may be insufficient in some cases. The model null space (Chapter 1) is a tough challenge in optimization because there is no knowledge which are the parameters that cannot be reconstructed. Modifying the data introduced in the inversion process can help to overcome this difficulty. For example, reconstructing only the low wavenumbers first, reduces the null space by reducing the number of model parameters [3].

Our initial motivation for this study was to understand why, under certain circumstances that will be briefly presented, the inversion of simultaneous frequencies provided better final models, than those found by doing hierarchical frequency groups. This assertion appears to be counter-intuitive because including high frequencies from the beginning implies a highly non-convex misfit function to minimize. However, if one of the sequential frequency groups mainly contains frequencies of the data that is not sensitive to model perturbations, the inversion will fail. On the other hand, if the simultaneous frequency group contains some data that is insensitive to model perturbations but also more data that constrains the model parameters, the model null space may be reduced and the inversion may succeed.

Jannane et al. (1989) used a smooth elastic velocity model to analyze which part of the model wavenumber spectrum could be reconstructed from a data set. The procedure consists in perturbing a velocity model with perturbations of different characteristic wavelengths λ. If the data after the perturbation remains unchanged, it means the data is insensitive to model perturbations which implies that the model parameters of a characteristic wavelength λ belong to the null space. Jannane et al. (1989) find that the short an long wavelengths of the model can be reconstructed, and there is a gap in the region of intermediate wavelengths. The data misfit related to long wavelengths are sensitive to travel times misfit of main reflections. The data misfit of low wavelengths is sensitive to changes in amplitudes in reflection data. A schematic representation is shown in Figure 4.23.

We are not going to study the spectrum of the model, but instead we are going to study the spectrum of the data. First we illustrate that due to non-linear effects introduced by the model, the spectrum of the data differs from the spectrum of the source. Moreover, the spectrum of the data may contain gaps. This depends if the transmitted or the reflected energy is dominant in the seismogram. The gaps in the data spectrum correspond to gaps in the model spectrum and, as indicated by Jannane et al. (1989), meaning there are model parameters in the null space.

2.1 Frequency response

The response of the medium in time and frequency in terms of the Green's functions can be expressed as,

$$\begin{aligned} u(x,t) &= G(x,x',t) \otimes s(x-x',t) \\ \mathcal{F}\left(u(x,t)\right) &= \mathcal{F}\left(G(x,x',t)\right)\mathcal{F}\left(s(x-x',t)\right). \end{aligned} \tag{4.36}$$

3. As long as the discretization grid is made coarser

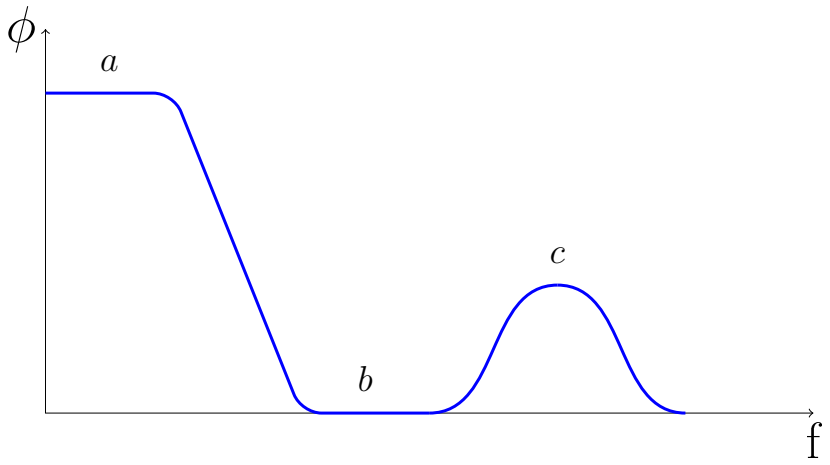


Figure 4.23: Schematic representation of the results in Jannane et al. (1989) illustrating the variation of the misfit in the data as a function of the frequency spectrum in the model. a) The misfit function of the low frequency spectrum of the model is sensitive to misfits in travel times of main reflections. b) The null space of the model c) The misfit function of the high frequency spectrum is sensitive to changes in reflection amplitudes.

Consider a model that consists of one reflector, $G(x, x', t) = \delta(x, x', t - a)$. The Fourier transform is

$$\mathcal{F}\left(\delta(x, x', t - a)\right) = exp(-i\omega a)\delta(x - x').$$

Using (4.36), the magnitude of the spectrum of the wavefield response is

$$|\mathcal{F}\left(u(x, t)\right)| = |exp(-i\omega a)\delta(x - x')||s(x - x', \omega)| = |s(x - x', \omega)|.$$

Therefore, for one reflector, the spectrum of the wavefield equals the spectrum of the source.

Consider now a model that consists of a series of reflectors represented by δ functions separated by intervals of time a,

$$G(x, x', t) = \sum_{n=-\infty}^{\infty} \delta(x, x', t - na).$$

The Fourier transform is also a series of δ functions but with a reciprocal constant,

$$\mathcal{F}\left(\sum_{n=-\infty}^{\infty} \delta(t - na)\right) = \frac{1}{a} \sum_{n=-\infty}^{\infty} \delta\left(x, x', \omega - \frac{2\pi m}{a}\right).$$

Using (4.36), the response in frequency of the wavefield is

$$\mathcal{F}\left(u(x, \omega)\right) = \frac{1}{a} \sum_{n=-\infty}^{\infty} \delta\left(x, x', \omega - \frac{2\pi n}{a}\right) * s(x - x', \omega). \tag{4.37}$$

Therefore, even if the spectrum of the source function is continuous, the spectrum of the wavefield is non-zero only for certain frequencies ($\omega = 2\pi n/a$). This can be understood as the fact that the sum of several arrivals in time $\delta(t - na)$, interfere constructively and destructively and modify the total spectrum.

We will now show an example where we look at the spectrum of a model with one reflector, and the compare it to the spectrum of a model with several reflectors. As shown with (4.37), the sum

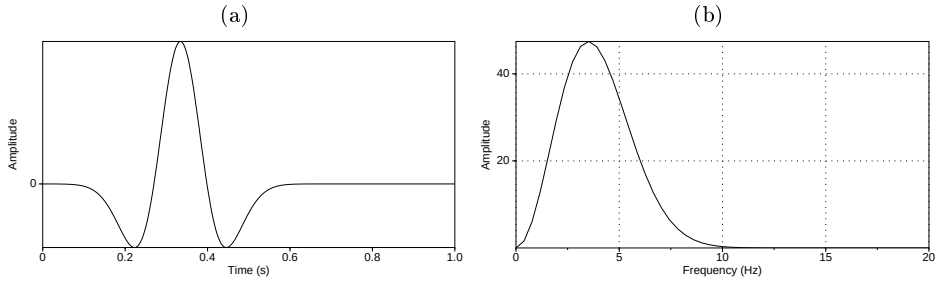


Figure 4.24: a) Ricker source wavelet with central frequency 3.5 Hz. b) Source spectrum.

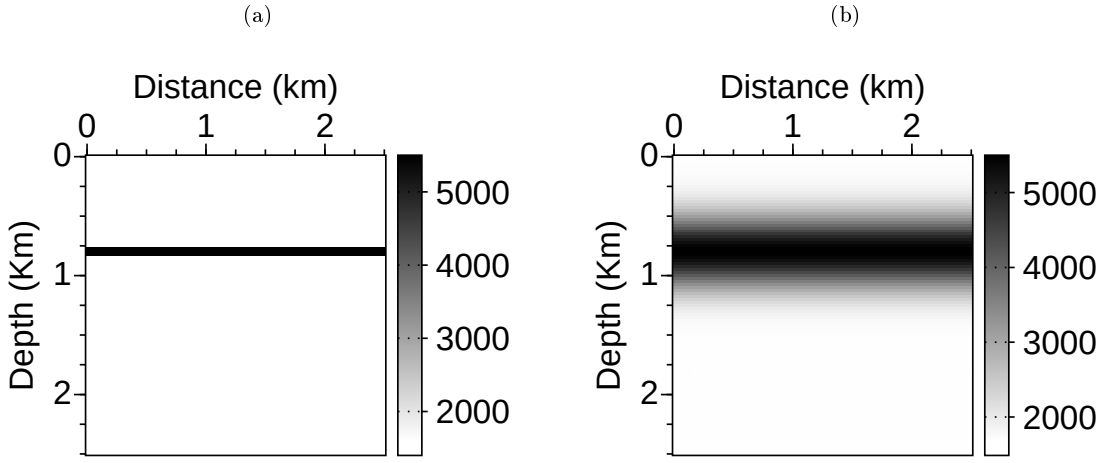


Figure 4.25: a) Velocity model m_1, containing one layer at $z = 800m$, with a homogeneous background of $1500m/s$. b) Smooth initial velocity model.

of arrivals generate interference effects such that the spectrum of the wavefield is not the same as the spectrum of the source, and may contain gaps. Following, for the BP-2004 salt model we analyze the spectral content of the observed data and we illustrate that this impacts the inversion.

Layered model

We consider 2D homogeneous velocity model of $1500m/s$. The model m_1 has one layer of $3500m/s$ at a depth of $z = 800m$ with a width of $50m$ shown in Figure 4.27. The model m_2 has one layer of 5500 m/s at a depth of $z = 950m$ shown in Figure 4.28 with a width of $100m$. We use a Ricker source function at $x_s = 2000m$ with peak frequency $3.5Hz$, shown in Figure 4.24 along with its spectrum. The source and the receivers are placed on the surface. We analyze the spectra of the seismograms generated in each of the models via a frequency time analysis, using wavelet coefficients. All the seismograms and the wavelet coefficients are plotted on the same scale, to allow for a fair relative weight comparison. We use a free surface boundary condition to account for multiples in the data, and absorbing boundary conditions on the other boundaries.

Using the velocity model m_1, the seismograms for different receiver positions and the corresponding wavelet coefficients are shown in Figure 4.27. For example, for a receiver at a position $x_r = 1950m$ (close to the source at $x_s = 2000m$, the observed, initial and residual data are plotted in Figure 4.27a. The wavelet coefficients for the observed, initial and residual data are

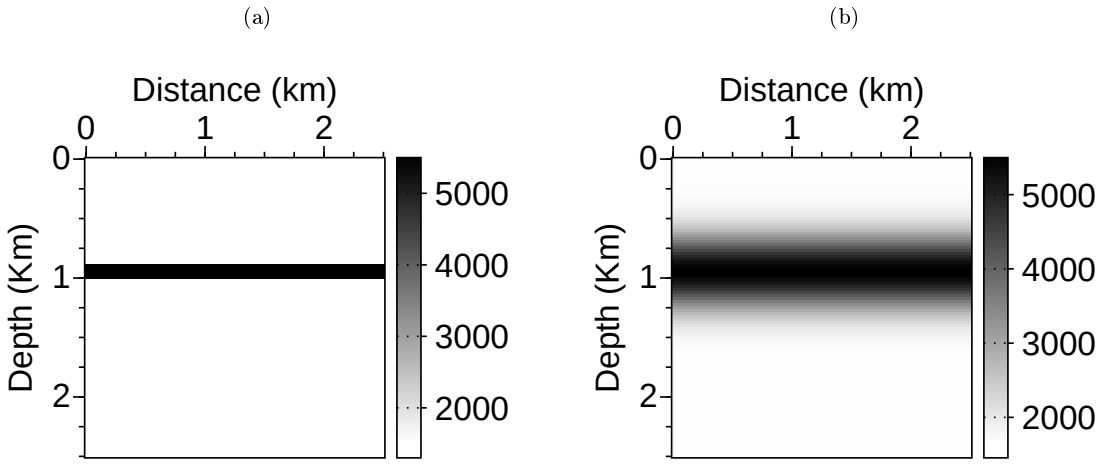


Figure 4.26: a) Velocity model m_2, containing one layer at $z = 950m$, with a homogeneous background of $1500m/s$. b) Smooth initial velocity model.

shown in Figure 4.27b. Because it is at short offset, the seismograms of the observed data register the first refracted arrival, and the succeeding reflection and multiple reflection arrivals separately. The frequency content of the first arrival and the succeeding reflection multiples is similar, with decreasing magnitude (due to the absorbing boundary conditions). As the offset increases (towards Figure 4.27g), the first refracted and reflected waves arrive closer in time, and the frequency spectrum remains the same.

If we analyze the seismograms for the velocity model m_2 consisting of one velocity layer of 5500 m/s at a depth of $z = 950m$ shown in Figure 4.28 with a width of $100m$, the seismograms and frequency time analysis results are similar to those shown in Figure 4.27, except that the reflected waves are now more energetic. Similar results would be obtained any other layer.

When a velocity model consisting of three layers as that shown in Figure 4.29 is used, the seismograms for different offsets and the corresponding wavelet coefficients are shown in Figure 4.30. For the receiver close to the source at position $x_r = 1950m$ in Figure 4.30a, the first arrival and the subsequent reflected wave arrivals are separated in the seismograms of the observed data. Unlike the previous cases for m_1 and m_2, the refracted and reflected waves do not show the same frequency content, and the magnitude is low below $5Hz$. A zoom into this image is shown in Figure 4.31. For other offsets, this same behaviour of the the spectrum of the observed data holds.

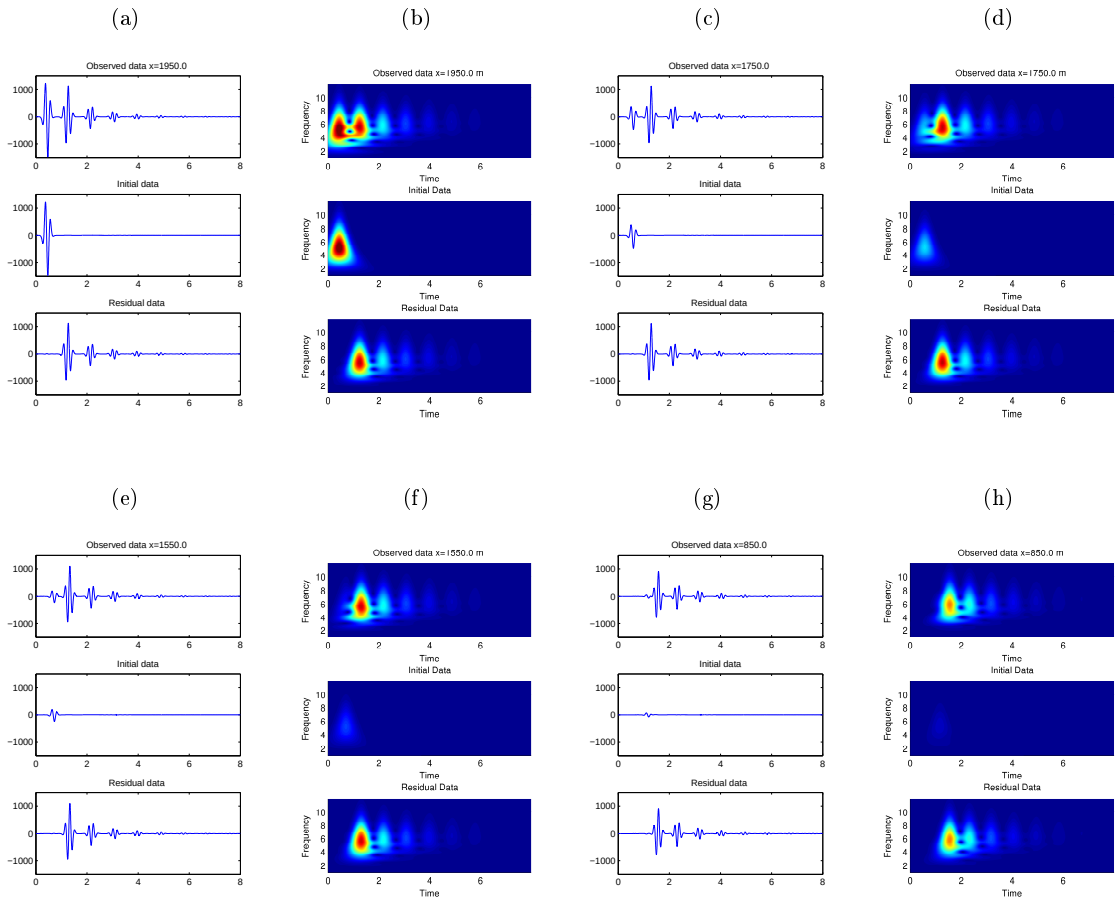


Figure 4.27: Ricker source with central frequency $f_0 = 3.5Hz$ at $x_s = 2000m$. Seismograms and wavelet coefficients for the observed, initial and residual data at different offsets for the velocity model m_1 in Figure 4.27.

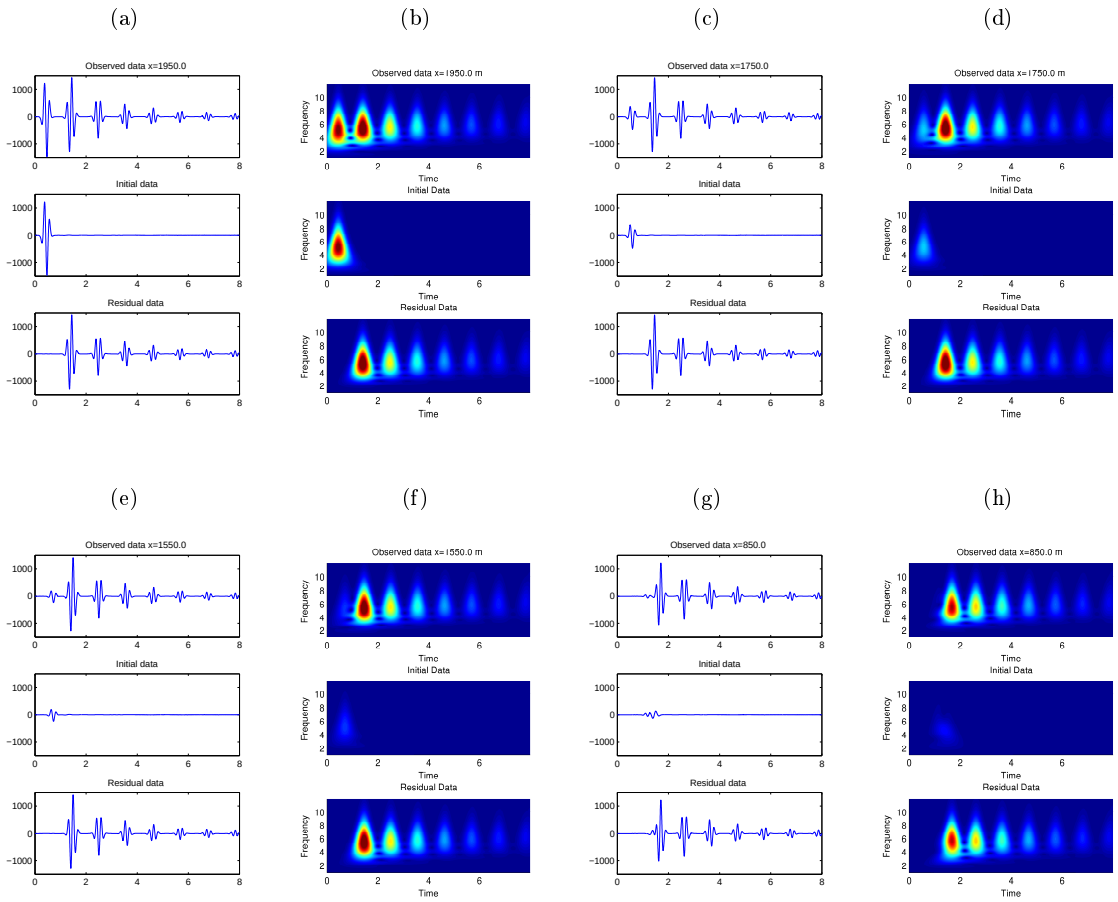


Figure 4.28: Ricker source with central frequency $f_0 = 3.5Hz$ at $x_s = 2000m$. Seismograms and wavelet coefficients for the observed, initial and residual data at different offsets for the velocity model m_2 in Figure 4.28.

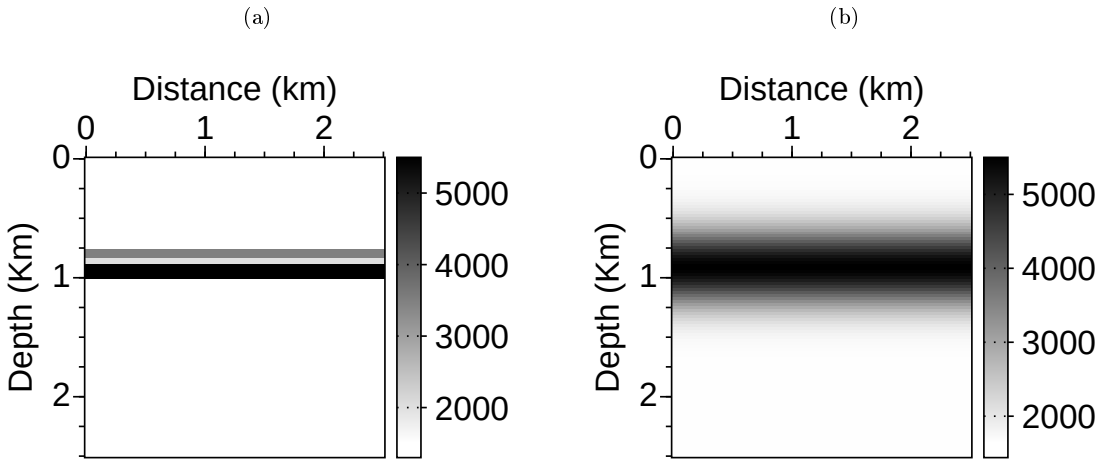


Figure 4.29: a) True velocity model, containing three layers, with a homogeneous background of $1500m/s$. b) Smooth initial velocity model.

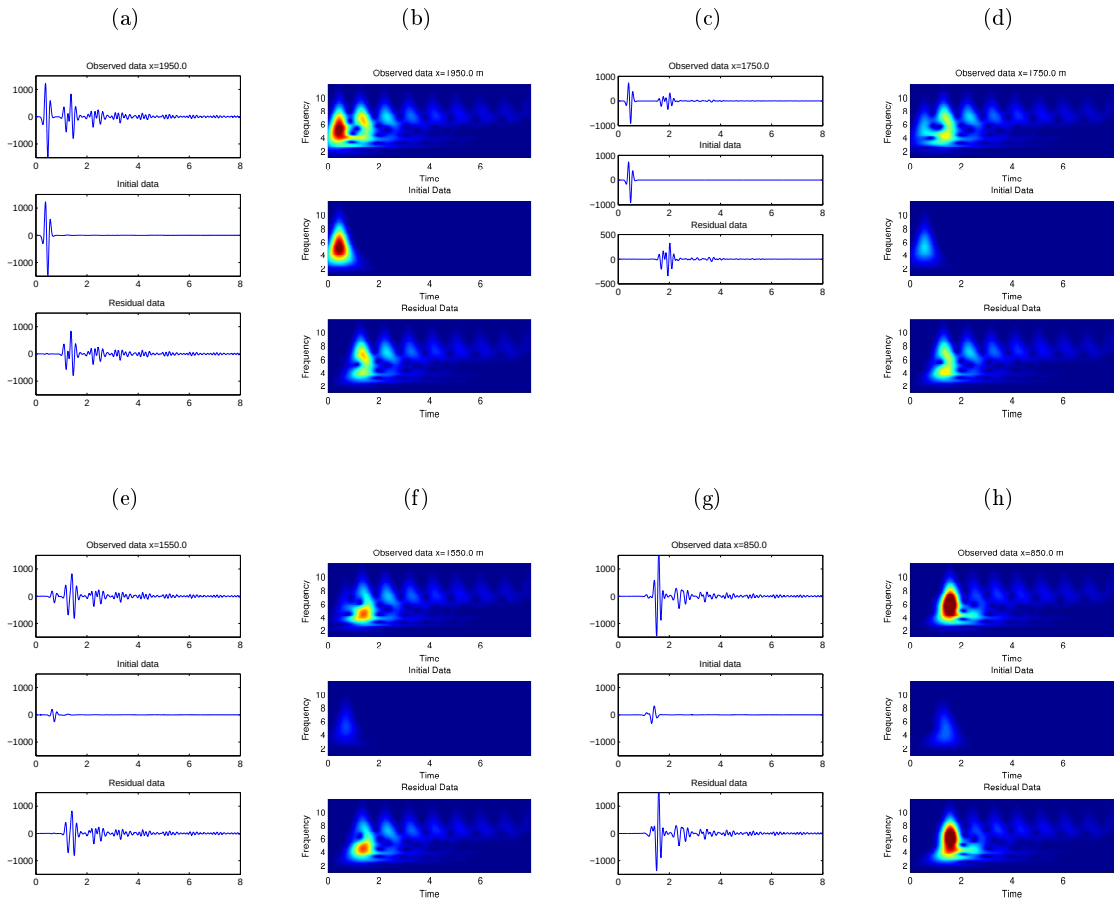


Figure 4.30: Ricker source with central frequency $f_0 = 3.5Hz$ at $x_s = 2000m$. Seismograms and wavelet coefficients for the observed, initial and residual data at different offsets for the velocity model m_3 in Figure 4.29. Notice the difference in the spectral content of the first arrival and reflected waves. A small gap appears in the reflected wave spectrum.

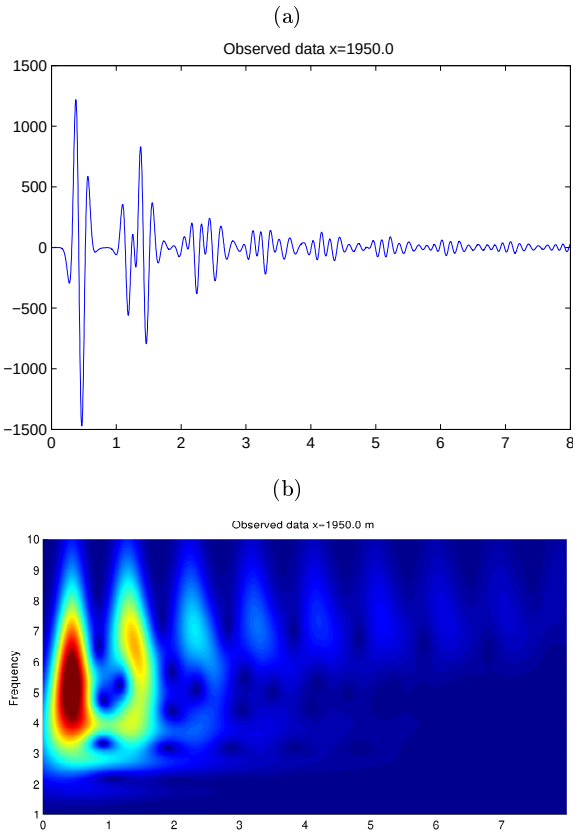


Figure 4.31: Zoom of Figures 4.30a,b. a) The seismogram shows the first arrival and the subsequent reflection arrivals. b) The spectral content is different for the first arrivals and the reflected arrivals.

2.2 Spectral content of the data in the BP-2004 Salt Model

From the previous illustration we established that the frequency content of the measured data may not be the same as the frequency content of the source due to interference in the arrival of waves caused by the model structure. We now perform an analysis of the seismograms using the $2D$ BP-2004 V_p salt velocity model shown in Figure 4.32a, which contains a high velocity contrast at a $1km$ in depth. As before, we only concentrate on the spectrum of the observed data, which is the information we use to reconstruct the model. We use the Ricker source wavelet similar to the one in Figure 4.24, with a central frequency of $4Hz$, located at $x_s = 2000m$. After the analysis of the seismograms, we will proceed to perform an inversion.

The shot gather for one source at $x_s = 2000m$ in Figure 4.33 shows the complexity of the wavefields. The observed, initial and residual data with its corresponding wavelet coefficients are shown in Figure 4.34, for different offsets. All the seismograms and wavelet coefficients are plotted on the same scale. There are some offsets, for which there is a gap in the spectrum of the observed data. Three different behaviours stand out :

- **Transmission and reflection energy : Short offset** (in this case, Figures 4.34a-f)

  For relatively short offset data, the arrival of the first refracted wave and the arrival of the reflected waves are distinguishable, as can be seen in the seismograms of the observed data in Figures 4.34a,c,e. The corresponding wavelet coefficients in Figures 4.34b,d,f show the frequency content of each arrival. The refracted wave contains all the range of frequencies provided by the source wavelet, and the reflected waves contain some low and high frequencies. A zoom of Figures 4.34a,b can be seen in Figure 4.35. This spectrum is shown schematically in Figure 4.36. The red curve represents the spectrum of the refracted wave, and the blue curve represents the spectra of the reflected waves. There are no gaps in the spectrum.

- **Reflection energy** (Figures 4.34a-j)
  The offset is sufficiently large so that the first arrival and the reflected waves are not independently identifiable in the observed seismograms but, on the other hand, interfere with each other. Since at this distance ($3500m < x < 5000m$) the receivers are close to the strongly contrasting salt dome, the reflected waves are more energetic than the refracted waves. The wavelet coefficients show that the spectrum of the observed data is mainly reflected energy, with a spectrum similar to the blue curve in Figure 4.36.

- **Transmission energy** (Figures 4.34k-l)

  The offset is sufficiently large so that the first arrival and the reflected waves are not separable on the observed data seismograms. Since the receiver is farther from the salt dome, the reflection energy is weak. Moreover, as can be seen in Figure 4.34k, the large amplitude of the first arrival shows that barely any energy of this wave has been lost in reflection processes, and there is mainly transmission energy. The spectrum of the observed data in Figure 4.34l confirms that this is mainly transmission data. The spectrum resembles the red curve of Figure 4.36.

In summary, when reflected energy is predominant in the data, we observe a gap in the spectrum of the observed data due to interference effects, as the blue curve in Figure 4.36. When there is transmission energy present in the data the spectrum is continuously covered, as represented in the red curve in Figure 4.36. The question is, how will this impact the inversion?

In the presence of strong reflectors in the model, if only short offset data close to the con-

trasting structures is available, most of the measured data will have gaps in the spectrum. The absence of certain frequencies in the observed data, produces a lack of sensitivity of the misfit function to these frequencies, implying that there are model parameters that cannot be resolved. If long offset data far from the contrasting structures is also available, the transmission energy may be recorded, the spectrum will be continuous and the model null space is reduced. However, if the inversion does not contain enough transmission data and the reflection data is predominant, the large null space problem remains a difficulty in the optimization process.

Despite having this knowledge, it is not clear how this information should be incorporated in the inversion. We now perform an inversion and will attempt to use this information to select the frequencies. The surface acquisition consists of 62 sources with a spacing of $\Delta s = 100m$, and 248 receivers per source with a spacing of $\Delta r = 25m$. The initial velocity model shown in Figure 4.32b is a smooth version of the true velocity model. We perform an inversion using l-BFGS and l_2 norm of the gradient of the model as a regularization. We first use two frequency groups, $(2-5\text{Hz})$, $(6-9\text{Hz})$ with $1Hz$ interval. The final velocity model is shown in Figure 4.37a. The inversion failed to find an adequate final model. We perform a second inversion, using a single frequency group $(2-9\text{Hz})$, with $1Hz$ interval. The final model in Figure 4.37b indicates a superior quality of the inversion, compared to the final result using two frequency groups.

Although the explanations for this unexpected results are not clear, we believe this phenomena is related with the strong contrasting boundary that creates strongly reflected waves that dominate the spectrum of the observed data. The presence of the gap in the measured data, enlarges the model null space and causes the inversion to fail. It seems that as more frequencies are introduced, all the information restraining the model parameters are injected simultaneously, reducing the model null space. However, in general, when introducing high frequencies from the beginning, the risk of cycle skipping and converging to an inadequate local minima is always present.

In this example, we selected the first frequency group $(2-5\text{Hz})$ and the second group $(6-9\text{Hz})$ to be separated by the gap. The first frequency group converges to a final model similar to that shown in Figure 4.37a and the second frequency group only performs a few iterations and stops. This shows that the inversion failed to explore frequencies greater than $4Hz$ when the two frequency groups are processed sequentially, while the full frequency range was successfully exploited by FWI when all the frequencies are processed all at once (Figure 4.37b). This suggests that the optimization takes advantage of the simultaneous access of the low-frequency part and the high-frequency part of the data to properly interpret the small amplitude of the intermediate frequency components as a result of interference effects. Another reason for this disparate behaviour may be related to the non-linear optimization process and the different paths followed by the inversion during the sequential versus simultaneous inversion. The sequential inversion honors the multiscale approach (Sirgue and Pratt, 2004) that is useful to reduce the cycle skipping problem: the long wavelengths are reconstructed before the short wavelengths. In the simultaneous inversion, introducing the high frequencies during the early stages of the inversion might have helped to reduce the above mentioned null-space thanks to the full frequency bandwidth that may have mapped the reflectivity information in the early models, hence providing a reflection regime in the sensitivity kernels which might have helped to avoid cycle skipping.

In conclusion, we analyze the spectral content of the measured data for the BP-2004 salt model using a surface acquisition. We identify a difference in the spectrum of the short offsets, receivers close to the reflecting boundaries, and receivers far from the reflecting boundaries. We see the reflected and transmitted waves have a different spectral content, and we observe a gap in the reflected wave spectrum. The impact this has on the inversion is yet to be studied with more

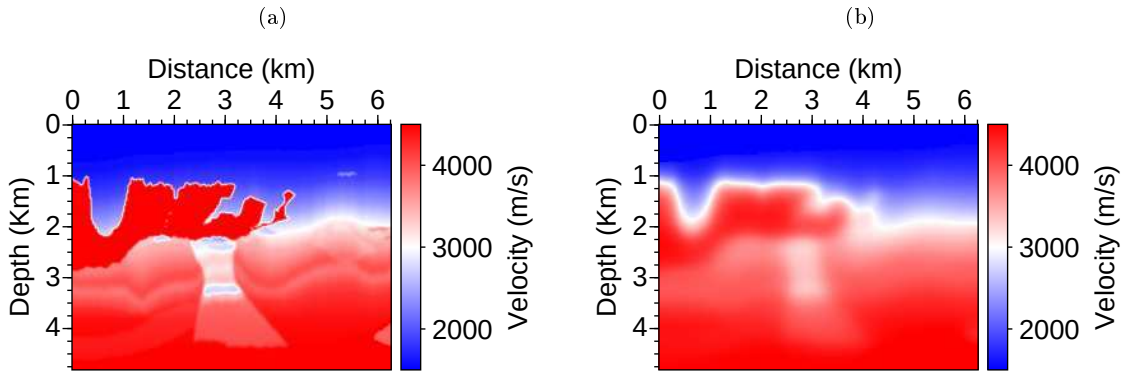


Figure 4.32: Bp-2004 salt model. True and initial velocity models

detail. We believe that this gap enlarges the model null space that may be otherwise reduced by including a wider range of frequencies in the inversion.

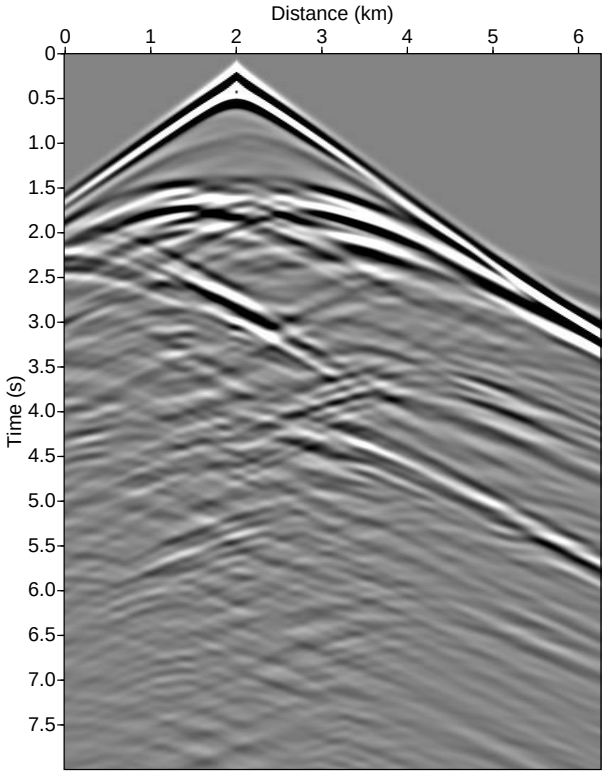


Figure 4.33: Bp-2004 salt model. Seismograms for one source at $x_s = 2000m$.

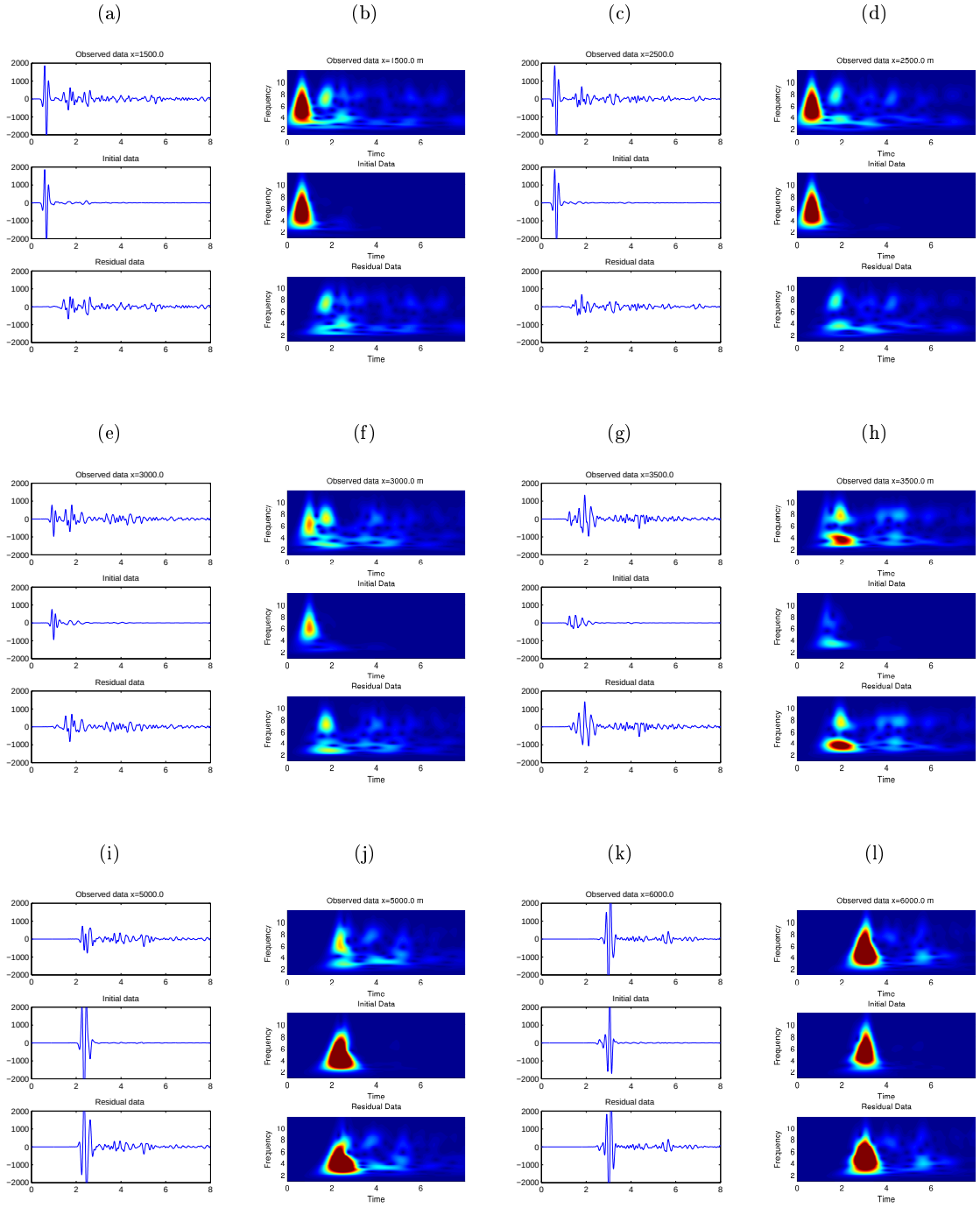


Figure 4.34: Ricker source with central frequency $f_0 = 4Hz$ at $x_s = 2000m$. Seismograms and wavelet coefficients for the observed, initial and residual data at different offsets for the true and initial velocity models in Figure 4.32.

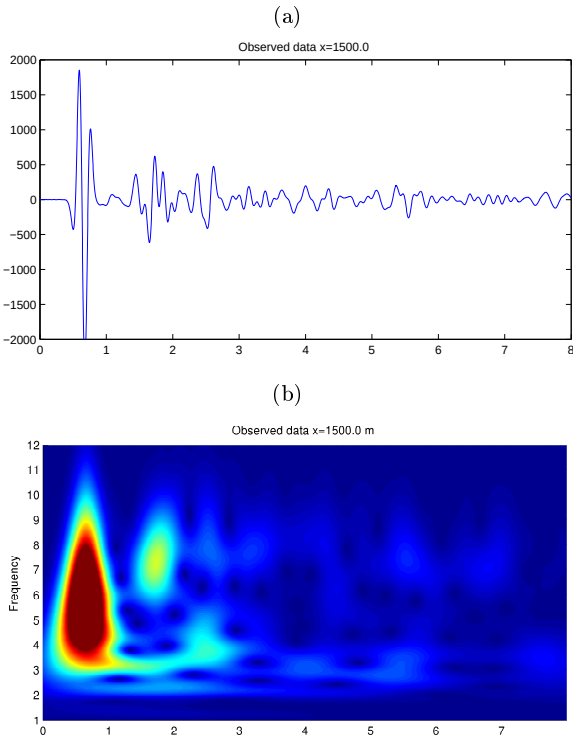


Figure 4.35: Zoom of Figures 4.32a,b. a) The seismogram shows the first arrival and the subsequent reflection arrivals. b) The spectral content is different for the first arrivals and the reflected arrivals.

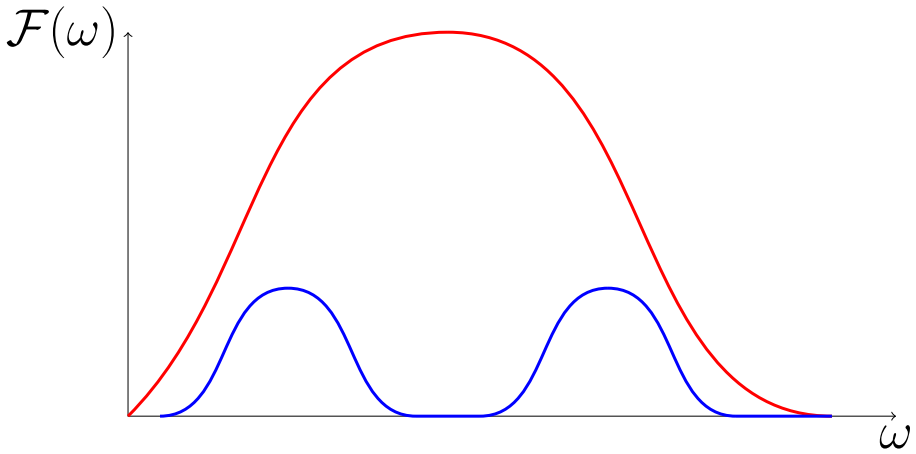


Figure 4.36: Schematic representation of the frequency content of wavefields, with a Ricker wavelet with a spectrum plotted by the red curve. A wave carrying transmitted energy has a spectral content as the line plotted in red. The frequency content of a wave with reflected energy is plotted in blue.

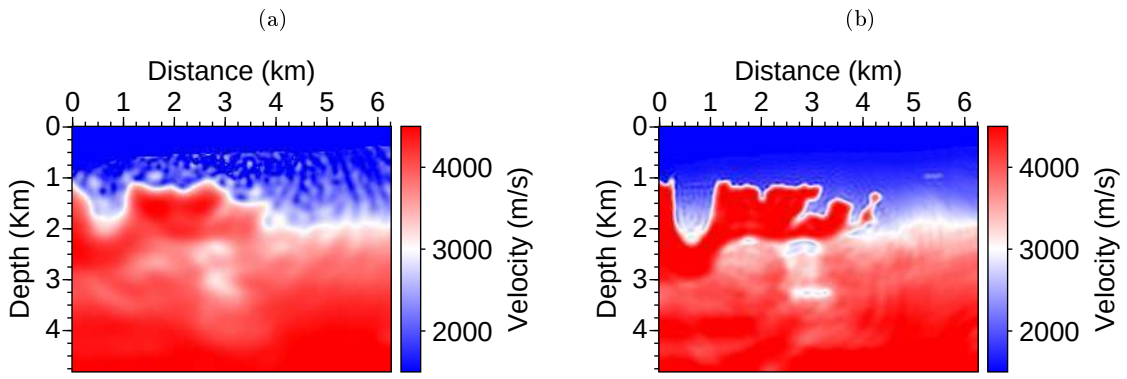


Figure 4.37: a) Final velocity model using a two frequency groups, $2-5$Hz, $6-9$Hz with $1Hz$ interval. b) Final velocity model using a single frequency group from $2-9$Hz, with $1Hz$ interval.

Conclusions and perspectives

1 Conclusions and perspectives

FWI has shown to provide high quality $2D$ and $3D$ images and now constitutes a production imaging tool in reservoir exploration. However, FWI still faces many challenges, specially in the topics of the image resolution, computational cost and non-linear optimization. This thesis concerns some aspects to reduce the computational cost and the use of regularization techniques in the optimization problem.

We presented a self-adjoint formulation of the first order velocity-stress elastic isotropic wave equation that allows to compute the direct and back-propagated wavefield using only one forward modeling operator by only applying a correction term to the source. This has been integrated and is used in our codes. An extension has already been done to the anisotropic velocity-stress equations by Brossier et al. (2013a).

The study combining source encoding with second order optimization methods reveals that the lowest computational cost needed to attain a predefined misfit value is provided by l-BFGS. The random encodings are not regenerated in every iteration, but only periodically. The encoding regeneration and the memory is refreshing of l-BFGS is done periodically in order to have Hessian approximations based on gradients with the same encodings. However, with noisy data, the most accurate and robust direction of descent is provided by Newton methods, with and without source encoding. The computational gain (speed-up) is a function of the misfit value, and so it is not possible to provide absolute values for the computational gain. For our synthetic tests, the highest computational gain was close to 60%, and for our real data application the computational gain was close to 90%.

The computational gains we provided are measured as ratio of the number of forward problems that must be solved with and without source encoding. Therefore, although we worked in 2D frequency domain using direct solvers, our conclusions on the computational gain are expected to be similar in 3D frequency domain applications with iterative solvers. This is because the expected value of the encoded gradient is proportional to the number of sources $\mathbb{E}[\tilde{g}] \backsim N_s$ (See for example Appendix 6), and the variance of the encoded gradient grows as $Var(\tilde{g}) \backsim N_s^2$ (as shown in section 6 of Chapter 3). The expected signal to noise ratio is therefore $\mathbb{E}[\tilde{g}]^2/Var(\tilde{g}) \backsim N_s^2/N_s^2 \backsim c$, proportional to some constant. That is, increasing the number of sources will not change the expected signal to noise ratio of the encoded gradient. However, for 3D acquisitions different strategies for source encoding may be needed. In section 6 of Chapter 3, the variance analysis of the gradient we presented, shows that when different sources generate similar gradients (large cross-talk), simply sub-sampling a subset of sources provides a lower variance than including all sources in the same super source. On the other hand, if there is not a lot of redundancy in the data and the gradients produced by each source are different, assembling all the sources in one super source provides a lower variance. In 2D applications there are not many sources that illuminate the same regions. On the other hand, for a 3D application, given that more sources will be illuminating the same region, more sources will generate similar gradients, there will be more redundancy and there will be more cross talk. Following the variance analysis, in this case we may subsample a few sources. An extension of the estimation of the variance of the encoded gradient with noisy data is yet to be performed. Finally, even though the speed up values in 3D are expected to be similar to our results presented here, for example 50%, the absolute gain in computational time will be greater with iterative solvers than with direct solvers. This occurs because the time spent in the solution of one forward problem is higher with iterative solvers than with direct solvers.

Although our results can be extrapolated to 3D frequency inversion with iterative solvers, they can not be extrapolated to an inversion in the time domain. This is beyond source encoding techniques and is more related to noisy data. When minimizing noisy data, an inversion in the time domain is expected to converge in fewer iterations than an inversion in the frequency domain with only some frequencies. Assuming the noise in each frequency component is independent, integrating over all frequencies in each time step averages the noise and decreases its imprint. On the other hand, performing an inversion of noisy data with a few frequencies will require more iterations to average out the effects of the noise. Therefore, the convergence curves of time are frequency domain inversion without and with source encoding are not the same and our results are not applicable to that setting. Moreover, without considering source encoding, if we are in a setting where it is already a priori known that many frequencies are required for the inversion to converge, it is more useful to perform a time domain inversion a apply source encoding in that that domain. This may be the case for short offset acquisitions or teleseismic FWI.

The quality of the final velocity models after FWI can be improved using digital image processing techniques. Total variation denoising algorithms remove the noise but may also remove small structures in the model. We incorporated the information of the position of the reflectors provided by the migration image, and performed a local total variation denoising that preserves the important information of the image and decreases the imprint of the noise. Other digital image processing techniques may be used to preprocess the data before inversion.

We compared the effect of using the total variation as a regularization constraint with the l_2 norm regularization term. The total variation and l_2 norm regularizations penalize differently the discontinuities of the model, and will shape differently the misfit function and will therefore have a different minima. In the numerical tests we performed, total variation provides a better description of the expected earth structure. However, when the regularization is strong small features may disappear. To continue this work, we would like to include the reflector information provided by the migration image, as was done in the denoising.

2 Conclusions et perspectives (FR)

L'intérêt de la FWI pour la construction d'images de haute qualité du sous-sol en $2D$ et $3D$ a été démontré par plusieurs applications si bien que cette technologie constitue désormais un outil de production pour l'exploration des réservoirs. Cependant, la FWI soulève toujours de nombreux défis, en particulier sur les thèmes de la résolution de l'imagerie, de son coût calcul et de l'optimisation non-linéaire. Cette thèse traite les aspects spécifiques tels que la réduction du coût de calcul de la FWI et l'utilisation de techniques de régularisation dans le problème d'optimisation.

J'ai présenté une formulation auto-adjointe de l'équation d'onde élastique isotrope du premier ordre en vitesse-contrainte qui permet de calculer le champ d'ondes direct et rétro-propagé en utilisant un seul opérateur de modélisation, son implémentation ne nécessitannt que l'application d'un terme de correction à la source de l'équation adjointe. Une extension a depuis été proposée pour les équations anisotropes vitesse-contraintes par Brossier et al. (2013a).

L'étude combinant les encodages des sources avec des méthodes d'optimisation de second ordre révèle que le coût de calcul le plus faible permettant d'atteindre une valeur prédéfinie de la fonction coût est fourni par l-BFGS. Les encodages aléatoires ne sont pas régénérés à chaque itération, mais seulement périodiquement. La régénération de l'encodage et le rafraîchissement de la mémoire est fait périodiquement afin d'avoir des approximations du hessien basées sur des gradients construits avec les mêmes encodages. Cependant, avec des données bruitées, la direction de descente la plus précise et robuste est donnée par les méthodes de Newton, avec et sans l'encodage de sources. Le gain de calcul dépend de la valeur de la fonction côut à laquelle les itérations sont arrêtées, et il n'est donc pas possible de donner une valeur unique du gain de calcul. Pour nos tests synthétiques, le gain de calcul le plus élevé était de près de 60 %, tandis que pour notre application aux données réels, le gain de calcul atteignait près de 90%.

Les gain calcul est défini comme le rapport du nombre de problèmes qui doivent être résolus avec et sans encodage des sources. Par conséquent, même si nous avons travaillé dans le domaine fréquentiel en 2D en utilisant des solveurs directs, nos conclusions sur le gain de calcul doivent aussi s'appliquer à des configurations 3D pour des applications en domaine fréquentiel fondées sur l'usage de solveurs itératifs. En effet, l'espérance du gradient codé est proportionnelle au nombre de sources $\mathbb{E}[\tilde{g}] \backsim N_s$ (Voir par exemple l'appendice 6), et la variance du gradient codé est proportionnelle au carré de nombre de sources $Var(\tilde{g}) \backsim N_s^2$ (comme indiqué dans la section 6 du Chapitre 3). Le rapport signal-bruit attendu est donc $\mathbb{E}[\tilde{g}]^2/Var(\tilde{g}) \backsim N_s^2/N_s^2 \backsim c$, soit proportionnel à une constante. Autrement dit, le rapport signal-bruit est indépendant du nombre de sources. Toutefois, des stratégies différentes pour l'encodage des sources peuvent être nécessaires pour les acquisitions 3D. Dans la section 6 du Chapitre 3 l'analyse de la variance du gradient, montre que lorsque les différentes sources génèrent des gradients similaires (favorisant des phénomènes de cross-talk plus significatifs), le sous-échantillonnage d'un ensemble de sources fournit une variance inférieure au fait d'inclure toutes les sources dans la même super-source. A contrario, s'il n'y a pas beaucoup de redondance dans les données et que les gradients produits par chaque source sont différents, l'assemblage de toutes les sources dans une seule super-source donne une variance inférieure. Dans les applications 2D, il n'y a pas beaucoup de sources qui éclairent les mêmes régions. A contrario, des applications 3D vont générer plus de cross-talk en raison de leur redondance supérieure. L'analyse de la variance montre que dans cette situation in est préférable de considérer un sous-échantillonnage des sources plutôt que leur assemblage. Une extension de l'estimation de la variance du gradient encodé avec des données bruitées reste à faire. Enfin, même si les pourcentages des valeurs des gains du calcul en 3D devraient être

similaires à nos résultats présentés ici, d'environ 50 %, le gain absolu en temps de calcul sera plus grand pour les solveurs itératifs qu'avec des solveurs directs. Cela provient du fait que le temps passé dans la solution d'un problème direct est plus élevé avec un solveur itératif qu'avec un solveur direct.

Bien que nos résultats puissent être extrapolés à l'inversion 3D en fréquence fondée sur des solveurs itératifs, ils ne peuvent pas l'être à une inversion dans le domaine temporel. Cette affirmation dépasse le cadre strict des techniques des encodages de sources. En effet, lorsque l'écart des données bruitées est minimisé, une inversion dans le domaine temporel converge en moins d'itérations qu'une inversion dans le domaine fréquentiel avec un sous ensemble des fréquences. En supposant que le bruit dans chaque composante de fréquence est indépendant, l'intégration sur toutes les fréquences à chaque pas de temps crée une interférence destructive du bruit qui diminue son empreinte. D'autre part, effectuer une inversion de données bruitées avec quelques fréquences, demande plus d'itérations pour moyenner les effets du bruit. Par conséquent, les courbes de convergence des itérations pour l'inversion dans les domaines temporel et fréquentiel avec et sans encodage de sources ne sont pas les mêmes et les résultats ne sont pas directement comparables. En outre, sans prendre en compte l'encodage des sources, si l'expérience sismique est dans une configuration où il est déjà connu a priori que de nombreuses fréquences sont nécessaires pour la convergence de l'inversion, il est plus utile d'effectuer une inversion dans le domaine temporel et d'appliquer l'encodage des sources dans ce domaine. Cela peut être le cas pour les acquisitions à court offset ou pour la FWI en télé-sismique.

La qualité des modèles de vitesses finales après FWI peut être améliorée en utilisant des techniques de traitement numérique d'image. Des algorithmes de dé-bruitage de variation totale suppriment le bruit, mais peuvent aussi éliminer les petites structures dans le modèle. Nous avons donc intégré des informations sur la position des réflecteurs fournie par l'image migrée, et nous avons effectué un débruitage avec un algorithme de variation totale locale qui conserve les informations importantes de l'image et diminue l'empreinte du bruit. D'autres techniques de traitement numérique de l'image peuvent être utilisés pour faire un debruitage des données avant l'inversion.

Nous avons comparé deux normes pour le terme de régularisation: la norme l_2 et la norme l_1. Les deux pénalisent différemment les discontinuités du modèle, et vont impacter différemment la fonction coût et conduisent donc à des minima différents. Dans les expériences numériques que nous avons effectuées, la variation totale fournit une meilleure description de la structure de la terre. Toutefois, lorsque la régularisation est forte, de petits structures (la texture) peuvent disparaître. Pour poursuivre ce travail, j'ai pour projet d'inclure les informations des réflecteurs fourni par l'image de la migration, comme cela a été fait pour le débruitage.

APPENDICES

1 Imaging conditions for other seismic imaging methods

Imaging inverse problems in electromagnetism, medical applications and geophysics generally consist in reconstructing the medium parameters from partial measurements made at the surface of an objects. Sources create waves that are transmitted through the media, that may be reflected (one or several times) by model discontinuities, and then are recorded at the receiver positions. The frequency content of the emitted waves is designed according to the size of the object we with to image. To image small bodies, ultrasound waves are used. To image the earth on a global scale, waves of periods of the order of $500s - 1s$ may be used. For the global scale, the sources are seismic events, such as earthquakes that are energetic enough to travel through long distances of the earth. For geophysical applications on the exploration scale, the sources are usually controlled explosions that generate waves that sample the region of interest, and the frequencies of the generated waves may normally range from $3Hz$ to $60Hz$. Once the data is measured and stored, the inverse problem consists in reconstructing the medium parameters such as the density, or the velocity at which the acoustic or elastic waves propagate. The key question then consists in determining how to transform the measurements into an image that represents the model parameters that accurately explain the recorded data. Usually, we will have an initial guess of the model parameters, and the purpose is to determine where our initial model approximation must be changed. The answer will be given by the imaging condition.

We will refer to the imaging condition as the operator that goes from the data space to the model space, and determines which parameters in the model space must be modified in order to explain the measured data. The question we wish to answer is what is the imaging condition that will allow us to create an image of the earth's subsurface?

The purpose consists in defining an imaging condition that will allow to determine the location of a source. The source may be, for example, an earthquake or a fracture that generates waves that probe the earth. We may also be interested in finding the position of secondary sources. A known primary source is used to illuminate the model, and the secondary sources create reflected and refracted waves. Secondary sources are thus reflectors, interfaces or any large model heterogeneities [4].

In seismic imaging there are several techniques available. The purpose of this section is to bring out the similarities amongst them, and point out the main limitations or assumptions made to reconstruct a perfect image. It is interesting to see that the underlying physical principles are similar amongst various methodologies and thus so are their assumptions and limitations. We will start with the imaging condition determined by time reversal (Fink, 1993), and work our way through the imaging condition of full waveform inversion, which is the center of this study.

1.1 Time Reversal Mirror Imaging Condition

The imaging condition relies on two physical concepts which are time reversal invariance and spatial reciprocity. *Time reversal invariance* means that the underlying physics remains unchanged if time is reversed. The wave equation in a non-dissipative heterogeneous medium only contains a second-order time derivative operator. For example, the acoustic wave equation is,

$$\left(\nabla \cdot \frac{1}{\rho}\nabla - \frac{1}{\rho c^2}\frac{\partial^2}{\partial t^2}\right) u(x,t) = s(x,t) \qquad (x,t) \in \Omega \times [0,\infty), \tag{38}$$

4. The size of the secondary sources must greater than the smallest wavelength of the waves used to probe the media.

where $c(x) = 1/\sqrt{\rho(r)\kappa(r)}$ where κ is the compressibility, ρ is the density, c is the local wave speed and s is a function with compact support. Therefore, if only even orders of the time derivative appear then we can show that if $u(x,t)$ is a solution to the wave equation, then $u(x,-t)$ is also a solution, meaning it is invariant under time reversal (Fink, 1993). If the medium has a frequency dependent attenuation, an odd order time derivative may appear, and time reversal invariance is lost. Modifications may be done to correct this (Ammari et al., 2011). *The spatial reciprocity* means for two points in space x, x_0, the Green's function satisfies that $G(x,x_0) = G(x_0,x)$. In this case, the Green's function $G(x,x_0)$ is the impulse response of the second order wave equation. That is, it is the solution of the wave equation using $\delta(x-x_0)$ as a source function. Spatial reciprocity for two points in space thus means that the impulse measured at a point x due to a source in point x_0 is the same as the impulse response measured at point x due to a source at x_0.

Fink (1993) employed the time reversal and spatial invariance with ultrasound waves to devise an ideal time reversal experiment that locates a reflecting target in an inhomogeneous medium. An array of transmitter-receiver piezoelectric transducers are placed around the target giving rise to what is called a time-reversal mirror (TRM). The pressure field $p(r,t)$ is registered by the transducers, digitalized and stored during an interval T. The pressure field is then time-reversed and retransmitted into the media by the transducers that now behave as sources. The time-reversal procedure refocuses the waves on the source. This process is used to focus on a reflective target that behaves like a (secondary) source. In time reversal experiments one assumes the medium does not have a large attenuating coefficient, otherwise modifications to the original imaging condition must be done. If we assume a low attenuating medium, theoretically there exists a set of waves that precisely retrace the path back to the source, even if the the propagating medium is heterogeneous. With transducers placed as far as half of the smallest wavelength apart, the waves will refocus to the source point and objects of the size of a wavelength can be detailed. Moreover, further experiments showed (Fink, 1993, 2008) that in the presence of a highly heterogeneous media the resolution of the time-reversed beam is better than that in an homogeneous medium. This somewhat paradoxical finding results from the fact that multiple reflections in the media help the redirection of towards parts that would have been missed otherwise with an homogeneous media. Therefore, after the time-reversal the multiple scattering medium behaves like a focusing lens, making the mirror to apparently have an aperture larger than it really is, and thus improving the resolution of the image (Fink, 1993). We will now detail the imaging condition for time-reversal.

Ideal time reversal imaging condition

Let $u(x,t)$ be the solution to the system,

$$\begin{cases} \left(\nabla\cdot\frac{1}{\rho}\nabla - \frac{1}{\rho c^2}\frac{\partial^2}{\partial t^2}\right)u(x,t) = s(x,t) & (x,t)\in\Omega\times[0,\infty) \\ u(x,t) = 0 & t<0 \\ \frac{\partial u}{\partial t} = 0 & t<0, \end{cases} \tag{39}$$

and let $g(x,t)$ be the solution $u(x,t)$ on the boundary : $g(x,t) = u(x,t)$ for all $x\in\partial\Omega$ and $t\in[0,T]$, where T is sufficiently large such that $u(x,t) = 0$ and $\frac{\partial u}{\partial t} = 0$ for $t\geq T$ and $x\in\Omega$.

☛ We must define an imaging condition that allows to reconstruct $s(x)$ from the knowledge of g on $\partial\Omega\times[0,T]$.

Let $v(x,t)$ be the solution to the wave problem

$$\begin{cases} \left(\nabla\cdot\frac{1}{\rho}\nabla-\frac{1}{\rho c^2}\frac{\partial^2}{\partial t^2}\right)v(x,t)=0 & (x,t)\in\Omega\times[0,\infty) \\ v(x,0)=0 & x\in\Omega \\ \frac{\partial v}{\partial t}(x,0)=0 & x\in\Omega \\ v(x,t)=g(x,T-t) & (x,t)\in\partial\Omega\times[0,T]. \end{cases} \tag{40}$$

The imaging functional for time-reversal is defined as (Ammari et al., 2011)

$$\boxed{\mathcal{I}_1(x)=v(x,T),\quad x\in\Omega}. \tag{41}$$

☛ The imaging condition (41) reflects the physical principal of time-reversal imaging. The first step is to store the wavefield $u(x,t)$ generated by $s(x)$ on the boundary $\partial\Omega$, which is done by solving system (39). The second step consists in time-reversing the wavefield on the boundary, which is achieved by defining $g(x,t)$. Finally, function $g(x,t)$ is imposed as a boundary condition and is resent into the media, which generates a wavefield $v(x,t)$ described by system (40). Finally the wavefield $v(x,t)$ will refocus at the source at the end of the propagation which occurs at time T, giving rise to the condition (41).

This imaging condition can be otherwise expressed in terms of the convolution of the Green's function using the Helmholtz-Kirchhoff theorem and doing a far field approximation (Fink, 2008; Ammari et al., 2011),

$$\mathcal{I}_2(x)\approx\int_{t=0}^{T}\frac{2}{\rho c}\frac{\partial}{\partial t}\int_\Omega G(x',x_s,-t)\otimes G(x_r,x',t)\mathrm{d}x'.\,^{5} \tag{42}$$

We can replace the integral with a discrete sum over N_r surface element positions and use the spatial reciprocity which gives rise to (Fink, 2008; Ammari et al., 2011),

$$\boxed{\mathcal{I}_2(x)\approx\int_{t=0}^{T}\frac{2}{\rho c}\frac{\partial}{\partial t}\sum_{i=1}^{N_r}G(x_s,x_i,-t)\otimes G(x_r,x_i,t)}. \tag{43}$$

The summation over an N_r element array is crucial because it is the sum of all contributions that will constructively interfere at the source location at one time, and interfere destructively at other locations and times. When there is full aperture, the time reversal process is a spatio-temporal matched filter (Fink, 2008), leading to the conclusion that the imaging condition is thus optimal. However, as any other imaging method, TRM has several limitations. It should be pointed out that the formulation above is for point like scatterers. Extended targets are analysed by, for example, Hou et al. (2006).

Limitations and applications of TRM

TRM has several limitations that principally concern the imaging resolution and problems arising from the limited aperture of the sampling acquisition (Fink, 1993):

– There is a resolution limitation. The image of a point is not point but rather a point with dimensions that depend on the minimum wavelength.

5. $f\otimes g=\int_{-\infty}^{\infty}f(\tau)f(t-\tau)\,d\tau$, the convolution product

- Anther resolution limitation results when the aperture of the TRM is reduced, meaning that the transducers do not entirely surround the target. In this case some information is lost and as the aperture of the time-reversal mirror (TRM) gets smaller, the focal spot becomes larger, inducing what is called a point spread function.

- A spatial sampling distance of $\lambda/2$ between transducers, where λ is the central wavelength of the pressure field, is necessary to avoid aliasing effects.

- A temporal sampling must have a maximum rate of $T/8$, where T is the central period of the data, to avoid secondary lobes .

- In the presence of a weakly inhomogeneous medium, the Born approximation is satisfied (only single scattering processes are taken into account), and TRM is able to reconstruct the wavefield distortion. However, in the presence of strong inhomogeneities, the efficiency of TRM is limited by multiple scattering events. A first limitations arises from the fact that the recording time interval T must be very long to take into account all the scattered events. Another reason is that in the presence of limited aperture, if the scattered field radiates in all direction due to the strong inhomogeneity, some information will be lost due to the lack of coverage and the time reversed data will be insufficient to optimally focus on the target.

- In the presence of attenuation, and particularly frequency dependant attenuation, the wave equation is no longer time invariant and, although there is still refocusing at the positions of the sources, it is not the same as the original sources. The reason is that frequency dependant attenuation acts as a frequency filter, which is applied twice. Once during the forward propagation, and another time in the back propagation. Assuming that the primary or secondary source was originally a delta function, some frequency components are lost in each step, and this is an irreversible process.

Despite the limitations, time-reversal has been successfully used in applications of all scales. In global seismology, for example, Montagner et al. (2012) worked with an elastic 3-D earth model to characterize the source of an earthquake. They use a normal mode formalism which allows to distinguish between radial and lateral variations of the physical parameters. They show that successful recovery of the Green's functions is only possible for long periods ($>$ 100s). One of the major setbacks is the non-uniform distribution of stations. One possible way to work around this is to assign different weights to the stations. Accounting correctly for the phase information locates accurately the source in time and space. Accounting for the amplitudes correctly, allows the radiation pattern be retrieved. Therefore, the phase information carries the most important information for the time reversal imaging condition. To emphasize the importance of the phase information, a one bit discretization is applied to the seismograms recorder at the stations (Figure 38 c). The purpose of applying a binarization to the seismograms is to conserve the phase information but completely disregard the amplitude information. When the binarized seismograms are time reversed the focusing is comparable to the focusing obtained when the complete seismograms are time reversed (Figure 4 b), d)). The difference lies in the fact that the radiation pattern is not recoverable. Although it was already intuitive that phase carries the most important information for the source location, this example in Figure 38 demonstrates it. More applications are cited in the references of Montagner et al. (2012) and Larmat et al. (2006).

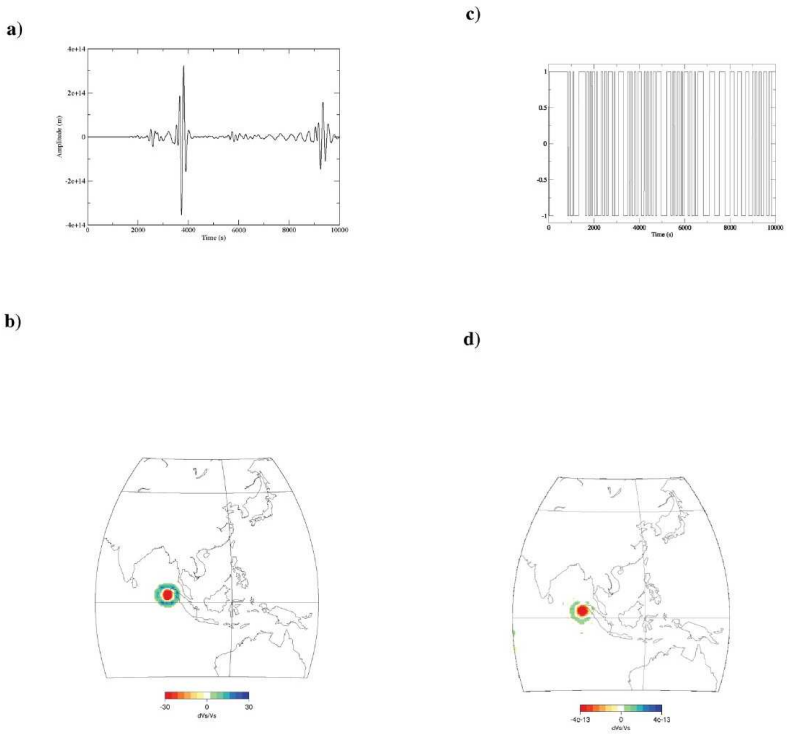


Figure 38: Figure from Montagner et al. (2012) to show that the phase carries the most important information in TRM. Synthetic example of the time reversal method for the source location of an earthquake. a) Seismogram calculated by normal mode summation. b) time-reversal using the complete seismograms. c) One bit seismogram. d) Time-reversal using one bit seismograms. For TRM, the weight distribution of the stations is given according to a Voroni tessellation to compensate for their non-uniform distribution.

1.2 Imaging with noise cross correlations

It has been shown that it is possible to recover the Green's function from the cross correlations of coda waves in a multiple scattering medium (Derode et al., 2003; Campillo and Paul, 2003; Shapiro et al., 2005). To recover the Green's function between two passive receivers one needs to use the reciprocity theorem, time-reversal invariance and the Helmholtz-Kirchoff theorem, thus relying on the same physical principles as time reversal previously described.

Consider a medium with scatterers three transducers A,B,C shown in Figure 39a. Imagine a time-reversal experiment where B sends a pulse that is recorded by C denoted by $h_{CB}(t)$. The point C time reverses it and sends it back. The resulting wavefield at point A is proportional to

$$h_{CB}(-t) \otimes h_{AC}(t), \tag{44}$$

where $\otimes$ represents the convolution operation (we have left out the convolution with the source functions). Because of the commutative property of the convolution, (44) is equivalent to

$$h_{AC}(t) \otimes h_{CB}(-t). \tag{45}$$

The purpose is to reconstruct the direct Green's function h_{AB}, but in general there is no reason for $h_{CB}(-t) \otimes h_{AC}(t)$ to be equal to h_{AB}. However, one can go beyond. Notice that thanks to reciprocity, equation (45) can also be rewritten and reinterpreted as if A and B were two passive receivers and as if point C was a source. In this case this would give

$$h_{AC}(t) \otimes h_{BC}(-t).$$

If many sources C are used and placed in such a way that, with the interpretation of time-reversal experiment in equation (44), they form a perfect time-reversal mirror, then

$$\sum_C h_{AC}(-t) \otimes h_{CB}(t) = h_{BA}(t) + h_{AB}(-t), \tag{46}$$

and the impulse response h_{AB} is recoverable. The Green's function $G_{AB}(t)$ is then available because it is proportional to the derivative of the cross correlation of the wavefields at two points A, B. Of course, if the points C are not placed in a perfect time-reversal mirror, then the approximation of the impulse response will not be good.

A numerical experiment done in Derode et al. (2003) shows that when the sources completely surround the medium, as shown in Figure 39a, the correlation coefficient between $h_{AC}(-t) \otimes h_{CB}(t)$ and $h_{AB}(t)$ is of 97.4%, meaning the approximation is good. If the sources do not completely surround and the medium as in Figure 39b, the correlation coefficient decreases to 81.9% because part of the waves are not recorded by the time-reversal device due to the presence of scatterers outside the the sources.

Limitations and applications

The sources C may be coherent pulses or diffusive noise continuous sources, arising from oceans, wind, etc, which renders the possibility to recover the Green's function between two points by cross correlating the registered noise waveforms. For the correct recovery from noise cross correlation, the noise sources should theoretically satisfy the above stated requirements of being randomly distributed everywhere, and be uncorrelated (Derode et al., 2003; Garnier and Papanicolaou, 2009). Therefore, a pre-processing of the data step should be done to eliminate strong events (such as earthquakes) that violate the conditions of uniformly distributed energy.

However, in geophysical applications the noise sources are not always symmetrically distributed, and may even be completely one sided. Nonetheless, the recovery of the main features of the Green's function is still possible (Stehly et al., 2009; Garnier and Papanicolaou, 2009). Through simulations it can also be seen that the causal and anti-causal correlations are almost never identical in a real experiment, but increase their similarity when there is strong multiple scattering. It is also common to apply a one bit normalization to the seismograms previous to the cross correlation operation to disregard amplitude information, and keep only phase information (Cupillard et al., 2011). Once the empirical Green's function is reconstructed it is then possible to find a background velocity model. This technique is particularly useful in regions that are not seismically active. An example of a global scale application of velocity maps recovered from the reconstruction of the Green's function is shown in Figure 40. Stehly et al. (2009) use one year of seismic noise recordings to perform the cross correlations and recover the empirical Green's function, in the $5s - 10s$ period band. With the Green's function, an inversion is possible and Figure 40 shows the velocity perturbations with respect to a background model in the region of the Alps, where the array density was high, using Rayleigh waves with a period of $5s$. The sensitivity of these waves is in the upper crust. The authors show in Figure 40 that the velocity perturbations found are in correspondance with the known geological properties of the region.

Alternatively, one could also image reflectors by imposing an imaging condition depending on the configuration of the noise sources. Garnier and Papanicolaou (2009) give a theoretical background of different imaging conditions using the cross correlation of noise functions. One possible configuration analized therein is shown in Figure 41a, where the circles represent the noise sources, the triangles represent the receivers and the diamond represents the reflector. The imaging condition specific to each configuration is found by doing a stationary phase analysis and finding the singular component analysis of the cross correlation function. For the configuration in Figure 41a, the cross correlation between points x_1 and x_5 is shown in Figure 41b, where the central part of the cross correlation is set to zero because they do not want to find the background velocity model, but only want to image the position of the reflector z_r. After applying the imaging condition [6] described in Garnier and Papanicolaou (2009) for this configuration, the image of the reflector position is shown in Figure 41c.

1.3 Imaging condition with adjoint methods

Time reversal also has strong connections with adjoint methods introduced in geophysics by Tarantola and Valette (1982), and which are heart of seismic imaging methods such as reverse time migration (Claerbout, 1985), travel time tomography (Tromp et al., 2005) and full waveform inversion (Plessix, 2006). For each of these methods, the imaging condition requires one forward calculation of a direct wave field, and one calculation of and adjoint field.

The imaging condition using adjoint methods is a matching condition everywhere on the physical domain, at each time step, between the source wave field $u_s(x,t)$ and the back-propagated wave field $v_s(x,t)$. The image is given by the zero lag cross correlation of the source wave field u_s and a back-propagated wave field v_s from the receiver positions (Claerbout, 1985; Plessix, 2006),

$$\mathcal{I}(x) = \sum_s \int_0^T v_s(x, T-t) \frac{\partial^2 u_s}{\partial t^2}(x,t) dt \tag{47}$$

6. The equation of the imaging condition is given in Garnier and Papanicolaou (2009).

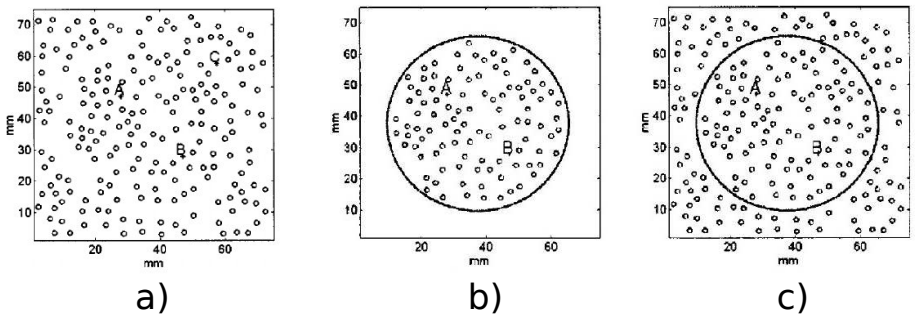


Figure 39: Figure from Derode et al. (2003). Medium with a random collection of 100 scatterers. The boundary conditions on the edges of the square domain are absorbing a) A, B, C are three transducers. b) A and B are passive receivers , 250 sources are placed around the circumference that entirely surround the scatterers. c) A and B are passive receivers , 250 sources are placed around the circumference that only partially surround the scatterers.

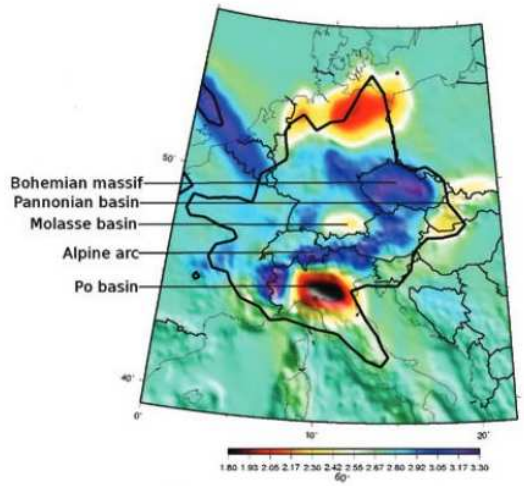


Figure 40: Figure from Stehly et al. (2009). Green's function obtained from the correlation of 1 year of recordings of seismic noise. In this case, the one-bit normalization to the seismograms is not applied but rather the spectrum of the seismograms is whitened. This imposes that the noise records all have the same energy, without changing the phase. The empirical reconstruction of the Green's function is inverted to obtain group velocity measurements. The velocity perturbations for the Alpine region are shown in the figure, using Rayleigh waves with a period of $5s$.

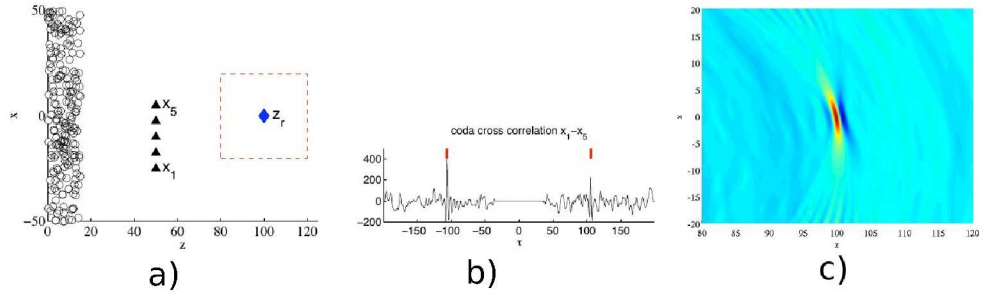


Figure 41: Figure from Garnier and Papanicolaou (2009). a) Configuration of the experiment : the circles represent the noise sources, the triangles represent the receivers and the diamond represents the reflector. Notice that the the noise sources are only on one side of the reflector. b) The cross correlation between points x_1 and x_5. The central part of the cross correlation is set to zero because we do not want to find the background velocity model, but we want to image the position of the reflector z_r. c) An imaging condition is defined specifically for this type of configuration (given in Garnier and Papanicolaou (2009)), which allows to position the reflector in depth.

Let $\lambda_s(x,t) = v_s(x, T-t)$, and we thus obtain :

$$\boxed{\mathcal{I}(x) = \sum_s \int_0^T \lambda_s(x,t) \frac{\partial^2 u_s}{\partial t^2}(x,t) dt.} \tag{48}$$

More general imaging conditions have been proposed (Sava and Fomel, 2006), where not only the term corresponding to the zero lag time correlation is preserved. The idea is to perform time shifts and see if the maximum lies elsewhere than at the zero time shift , as imposed by the standard imaging condition. This modification allows improve depth focusing,

$$\boxed{\mathcal{I}_{RTM}(x,\tau) = \sum_s \int_0^T \lambda_s(x, t+\tau) u_s(x, t-\tau) dt.} \tag{49}$$

Note that applying a time shift of τ to the wave fields is the same as applying a phase shift in the frequency domain, which is imply multiplying by a factor $\exp(2i\omega\tau)$. The imaging condition could also include space shifts,

$$\boxed{\mathcal{I}(x,h) = \sum_s \int_0^T \lambda_s(x-h, t) u_s(x+h, t) dt}, \tag{50}$$

where $h = [h_x, h_y, h_z]$ is a $3D$ vector, meaning there can be two horizontal offsets and one vertical offset. The imaging condition may have a maximum that is not at zero offset ($h = 0$), revealing velocity inaccuracies. In general, we can have an imaging condition including both time and space shifts

$$\boxed{\mathcal{I}(x,h,\tau) = \sum_s \int_0^T \lambda_s(x-h, t+\tau) u_s(x+h, t-\tau) dt.} \tag{51}$$

The imaging condition including time and/or space shifts can be useful because it allows for an angle decomposition. Time shifts are less computationally expensive than space shifts (one variables versus three), but numerical experiments show that the angular resolution is superior

with space shifts (Sava and Fomel, 2005). This is useful because a common method to determine if the data have been correctly imaged is to verify the alignment of the images created with multi-offset data.

Although reverse time migration (RTM) and full waveform inversion (FWI) both use (48) to image, the back propagated wavefields are different because the adjoint source are different and, consequently, so is the image.

1.3.a *Reverse time migration imaging condition*

The imaging condition of RTM is the same as that of FWI. The main difference is in the information of the measured data that is used to image. Since the purpose of RTM is to position the reflector, the measured data is processed and only the (single) reflected data is kept. Then the same explanation of the imaging condition as that explained in the Introduction or in section 1.5 of Chapter 2 applies to this case. The main difference is that the migration technique seeks an image with high wavenumber content, and low wavenumbers represent artefacts in the image. Low fequency artefacts are created in the presence of strong reflectors by the correlation of wavefields with back-scattered energy.

Limitations of RTM

- Assumption in the imaging condition that the seismic data consists of single scattered reflections. Therefore, we need to remove the multiples from the data.
- In the presence of high contrasting interfaces, artefacts appear, as explained in Figure 42. Low frequency artefacts may completely mask the migrated image. A high-pass frequency filter may be applied to the migrated image, or modified imaging conditions may be used (Zhang and Sun, 2009; Liu et al., 2011).

Alternative imaging conditions may be defined to remove the low frequency artefacts (Zhang and Sun, 2009; Liu et al., 2011). It is possible to identify the correlations that will give rise to artefacts in the migrated image. By only keeping the terms of the cross correlation with high frequency content (wavefields travelling in the same direction), the low frequency artefacts are removed (Liu et al., 2011). An example of a migrated image with artefacts is shown in Figure 43a, and without in Figure 43b.

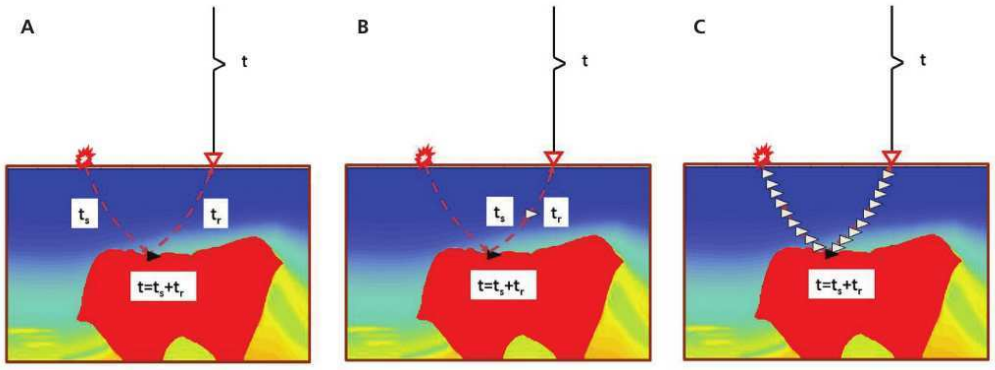


Figure 42: Figure from Zhang and Sun (2009) showing the problems with the imaging condition, in the context of RTM. The imaging condition places a reflector point whenever $t_s + t_r = t$. a) Ideal case : the imaging condition is satisfied only at the reflector point shown with a black arrow. b) Because the reflector creates strong reflected wave fields, there are other points that satisfy the $t_s + t_r = t$, like that shown with the white arrow. c) As a matter of fact, any point along the reflection path satisfies the imaging condition. The migrated image thus shows reflector points all along the ray trajectories, which produce low frequency migration artefacts.

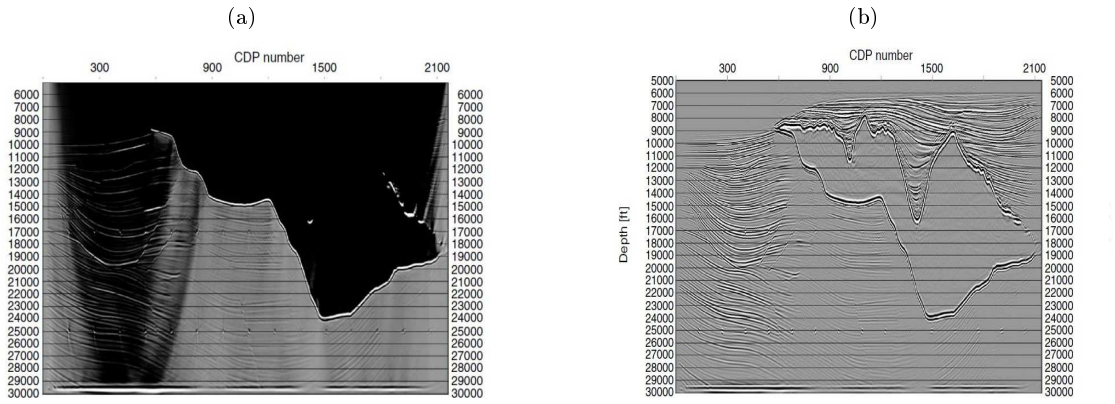


Figure 43: Figure from Liu et al. (2011). a) RTM model with imaging condition (48) showing low frequency artefacts that mask the high frequency image content. b) Migrated model with a modified imaging condition in Liu et al. (2011) that removes low frequency artefacts. The modified imaging condition is attained through a decomposition of the (direct and back-propagated) wavefields in up-going and down-going wavefields via a FFT.

1.4 Time reversal imaging conditions : conclusions

We have seen the imaging conditions mentioned have similarities in the assumptions required to image successfully the primary or secondary sources. When there are sources and receivers completely surrounding the target (complete illumination) no energy is lost, the energy is uniformly distributed and it is all resent back into medium and refocused correctly at the source target positions. However, different imaging techniques treat differently the data and respond to different needs. In geophysical applications, there are conditions for which each of these methods is more suited.

Direct time reversal imaging techniques described in Section 1.1 can mainly image a single source. It will be therefore useful when we want to image the source of an earthquake, as that shown in Figure 38. Passive imaging using the cross correlation of the noise may be useful on a global seismic scale in regions where normally there are no earthquakes and there are normally no waves sampling this region of the earth. On an exploration scale, passive seismic imaging generally produces low resolution images compared to the the images obtained with active sources. For the exploration scale, reverse time migration and full waveform inversion imaging conditions allow to obtain the best resolution of the images.

2 Travel time tomography

Travel time tomography defines the misfit ϕ as a function of the difference between the observed and calculated arrival time delays of a seismic phase ψ,

$$\min_m \phi = \min_m \|\delta T_o^\psi - \delta T_c^\psi(m)\|_2^2, \tag{52}$$

where δT_o represents the observed arrival time delay and δT_c represents the calculated arrival time delay. In ray-based methods, the arrival time of a seismic phase can be considered independent of the frequency content of the data and only the behaviour in the limit infinite frequency case is considered. In *finite frequency travel time tomography*, dispersion effects are taken into

account in the travel time arrival, and consequently a travel time arrival is measured for each frequency.

In travel time tomography, the direct problem is the linearized wave equation The general direct problem can be expressed as

$$\begin{aligned} A(x,m,t)u(x,t) &= s(x,t) && (53)\\ u(x,0) &= 0 \qquad x \in \Omega && (54)\\ \frac{\partial u(x,t)}{\partial t} &= 0 \qquad x \in \Omega, && (55) \end{aligned}$$

where $A(x,m,t)$ is the visco-acoustic wave equation operator, $u(x,t)$ is the wavefield solution, and $s(x,t)$ is the source function. To linearize the direct problem, we first begin by introducing a small perturbation on the model δm, $m = m_0 + \delta m$. We use the Born approximation to assume that wavefield solution u can be expressed as the unperturbed solution plus a first order perturbation term, $u = u_0 + \delta u$. Plugging this into the original forward problem equation and expanding we have,

$$\begin{aligned} A(x, m + \Delta m_0, t)\left(u_0(x,t) + \delta u\right) &= s(x,t) && (56)\\ \left(A(x,m_0,t) + \frac{\partial A}{\partial m}\delta m\right)\left(u_0(x,t) + \delta u\right) &= s(x,t) && (57)\\ A(x,m_0,t)u_0(x,t) + \frac{\partial A}{\partial m}\delta m\ u_0(x,t) + A(x,m_0,t)\delta u + \frac{\partial A}{\partial m}\delta m\delta u &= s(x,t) && (58) \end{aligned}$$

Using the original formulation of the forward problem (53) and excluding second order terms, we obtain

$$A(x,m_0,t)\delta u = -\frac{\partial A}{\partial m}\delta m\ u_0(x,t). \tag{59}$$

Note that (59) has the same form of the original forward problem (53). For (53), a general solution in terms of the Green's function can be expressed as

$$u(x,t) = \int_\Omega G(x,x',t-t') * s(x',t')dx'dt'. \tag{60}$$

Using the term $-\frac{\partial A}{\partial m}\delta m\ u(x,t)$ in (59) as a source function and replacing in (60), we obtain

$$\boxed{\delta u = -\int_\Omega G(x,x',t-t') * \frac{\partial A}{\partial m}u_0(x',t')\ \delta m dx'dt'.} \tag{61}$$

In a more compact form the linear direct problem can be written as,

$$\boxed{\delta u = -\int_\Omega B(x)\ \delta m(x)\ dx,} \tag{62}$$

where

$$\boxed{B(x,t) = G(x,x',t-t') * \frac{\partial A}{\partial m}u_0(x',t').} \tag{63}$$

Equation (61) is the expression for the scattered wavefield in the Born approximation. For travel time tomography it is necessary to select the travel time delay with respect to the arrival time in a background model (Nolet, 2008). In infinite frequency tomography, the travel time

delay is chosen based on the delay of the onset time. However, the onset time is difficult to pick in a seismogram. For this reason, the delay of other characteristics of the seismogram are selected. For example, one may choose to measure the delay of the maximum of the wavefield. Since the time of arrival of the maximum of a wave depends on dispersion effects, tomography that is not based on onset times is called finite-frequency tomography.

For each source-receiver pair, on seeks the maximum of the arrival time of $u(t)$ perturbed by $\delta u(t)$. Let $t = T_{SR}$ denote the time at which u attains its maximum . The shift δT at which $u + \delta u$ has its maximum value is found by taking the deriving with respect to time

$$\dot{u}(T_{SR} + \delta T) + \delta\dot{u}(T_{SR} + \delta T) = 0, \tag{64}$$

where $\dot{u}$ and $\delta\dot{u}$ represent the first time derivatives. Using a Taylor expansion to first order and using $\dot{u}(T_{SR}) = 0$ because it is the maximum, one obtains,

$$\delta T = -\frac{\delta\dot{u}(T_{SR})}{\ddot{u}(T_{SR})}. \tag{65}$$

The derivative of the expression of δu in (61) and the second derivative of u is replaced in (65). By doing this (see Chapter 7 (Nolet, 2008)), one arrives to

$$\boxed{\delta T = \int_\Omega K(x)\delta m dx}, \tag{66}$$

where $K(x)$ is referred to as the Fréchet kernel. This Fréchet kernel contains thus the information of the propagation of the wavefield u and δu in a specific background media. In practice, to solve for the wavefields u and δu one may use finite differences, finite elements or, it is also common to use ray theory and solve the Eikonal equation to solve for the travel time and geometrical spreading in a background media (Mercerat and Nolet, 2012). Independently of the method used to solve for u and δu that conform the kernel, it should be noted that this kernel is only computed once in the background model.

The most common way to find travel time delays is through cross correlations of the perturbed and unperturbed wavefields (Nolet, 2008). Let $\gamma(t)$ denote the autocorrelation of the unperturbed wavefield,

$$\gamma(t) = \int u(t)u(t - t')dt'. \tag{67}$$

The travel time delay will be the maximum of the correlation between the perturbed and unperturbed wavefield,

$$\gamma(t) + \delta\gamma(t) = \int \left(u(t) + \delta u(t)\right)\left(u(t - t')\right) dt'. \tag{68}$$

The maximum of any autocorrelation is always at zero so $\dot{\gamma}(0) = 0$, and the maximum of the perturbed wavefield, at time δT, is

$$\dot{\gamma}(\delta T) = \dot{\gamma}(\delta T) + \delta\dot{\gamma}(\delta T) = 0. \tag{69}$$

Once again, a first order expansion is done and neglecting terms of order higher than two (Nolet, 2008),

$$\delta T = \frac{\delta\dot{\gamma}}{\ddot{\gamma}(0)} = -\frac{\int \dot{u}(t)\delta u(t)dt}{\int \ddot{u}(t)u(t)dt}. \tag{70}$$

This expression can also be written in the form of equation 66. Also, as mentioned previously, any numerical method may be used to solve for the wavefield and the perturbed wavefield. The

Fréchet kernel will have a different expression depending on how the delays are measured, and on the method used to compute the wavefileds. A Fourier transform can be applied to the kernel, to perform an inversion in the frequency domain.

The inverse problem is defined as the minimization of the l_2 norm of the difference between the observed time arrival delays δT_o and the calculated time arrival delay $\delta T_c = K\delta m$.

$$\min_{\delta m} \phi = \min_{\delta m} \|\delta T_c - \delta T_o\|^2 \tag{71}$$

$$= \min_{\delta m} \|K\delta m - \delta T_o\|^2. \tag{72}$$

Taking the derivative of the misfit function with respect to the parameter perturbation we obtain,

$$\frac{\partial \phi}{\partial \delta m} = K^\dagger \left(K\delta m - \delta T\right). \tag{73}$$

In the simplest case, the model update will be given by

$$\boxed{\delta m_{n+1} = \delta m_n - \alpha K^\dagger \left(K\delta m_n - \delta T\right),} \tag{74}$$

where α is a step length. Conjugate gradient algorithms are also commonly employed. The term $K\delta m_n - \delta T$ represents the residuals in the data space. The operator $K^\dagger$ maps the residuals projects back to the model space. For travel-time tomography of the Earth's mantle, the matrix K is stored in disk and is not recomputed. This approximation is efficient whenever the non-linearities are weak. Each row of K corresponds to one specific source-receiver pair, and maps the residual for one source-receiver couple into the model space. It indicates which parameters of the model are sensible to the data for a source-receiver couple (Nolet, 2008).

Travel time tomography or cross-correlation travel time tomography, have the advantage that they are linear inverse problems. If the heterogeneities in the earth are smooth, ray theory is valid and the response of the earth will be linear. That is, if a heterogeneity is doubled in size, the delay time δT will also be doubled. When the heterogeneities have high contrasts and generate highly energetic reflected waves, the linear Born approximation starts to fail. In global tomography the heterogeneities normally do not exceed 10% and the linear assumption thus remains valid (Mercerat and Nolet, 2013). However, this may not be the case for near-surface study cases.

3 The gradient of the misfit function : computation with the adjoint state method

As was remarked in the introduction, one of the tasks that rendered FWI so computationally expensive was the calculation of the gradient of the misfit function. The adjoint state method is a mathematical tool that allowed a huge complexity reduction in this task, rendering FWI implementable.

The method combines an appropriate use of the Lagrangian function and integration by parts in order to re-express the derivative of the functional ϕ. It involves the solution of a backward (or adjoint) related to the forward problem defining the wavefield.

Although in this section we consider only the acoustic wave equation in the frequency domain, a similar development can be applied either to the same equation on the time domain or to different versions of the wave equation, as we show in section .

3.1 The Adjoint State Method

We are aiming to calculate the gradient of the misfit function when working with the acoustic wave equation in the frequency domain, for a fixed frequency ω. It will be convenient to introduce an appropriate Hilbert space in which we will derive the variational formulation. Let $\Omega \subset \mathbb{R}^3$ and $\tau \in \mathbb{R}^+$. Let f, g be two scalar functions from Ω to $\mathbb{C}$. We define the scalar product between two elements f and g of the Hilbert space as

$$\langle f, g \rangle = \int_\Omega f^*(x) g(x) dx, \tag{75}$$

where the operation denoted as * is the conjugate of f, and $\Omega \subset \mathbb{R}^3$ is a bounded set. To render the defined bilinear operator a true inner product, we identify the functions such that if $\langle f, f \rangle = 0$ then f is the null element of the space. Moreover, for completeness, we limit ourselves to the functions for which $\langle f, f \rangle$ is finite $L_2(\Omega)$.

3.1.a *Adjoint operators*

Let us consider a linear operator M applied to the vector g. The adjoint operator $M^\dagger$ of M is defined as,

$$\langle f, Mg \rangle = \langle M^\dagger f, g \rangle. \tag{76}$$

Auto-adjoint operators are defined by $M^\dagger = M$.

For example, suppose we are interested in finding the adjoint of the partial derivative in the $i-$th direction operator, i.e. for $H = \partial_{x_i}$ we search for an operator $H^\dagger$ such that relation (76) is satisfied. Using integration by parts, the scalar product between the function f and Hg can be written as

$$\begin{aligned} \langle f, Hg \rangle &= \int_\Omega f(x)^* Hg(x) dx \\ &= \int_\Omega f(x)^* \partial_{x_i} g(x) dx. \\ &= f(x)^* g(x)|_{\partial\Omega} - \int_\Omega [\partial_{x_i} f(x)]^* g(x) dx. \end{aligned}$$

Assuming zero boundary conditions, we end up with the following definition of the adjoint operator of the derivation operator

$$\langle f, \partial_{x_i} g \rangle = -\langle \partial_{x_i} f, g \rangle.$$

By extension, this means that if $H = \partial_x$, then $H^\dagger = -\partial_x$.

3.1.b *Directional derivatives*

We recall here the definition of the Gateaux directional derivatives applied to our setup as it is useful in the following. Let $\mathcal{W}$ be a Banach space (for example $L_2(\Omega)$). Let $m, \zeta \in \mathcal{W}$. Let $\mathcal{L} : \mathcal{W} \to \mathbb{R}$). If

$$\lim_{\epsilon \to 0} \frac{\mathcal{L}(m + \epsilon\zeta) - \mathcal{L}(m)}{\epsilon}$$

exists, then we say that $\mathcal{L}$ is Gateaux derivable in the direction of ζ and we denote the limit as $D\mathcal{L}(m; \zeta)$.

If the function is Gateaux derivable for all $\zeta \in \mathcal{W}$, by Riesz representation theorem, there exists an element in $\mathcal{W}$, that we denote $\nabla_m \mathcal{L}$ such that

$$D\mathcal{L}(m; \zeta) = \langle \nabla_m \mathcal{L}, \zeta \rangle . \tag{77}$$

3.1.c *The Lagrangian and the adjoint state equations*

We use the same notation and definitions as in Chapter 2, Section 1.2. In particular, we recall that ϕ is the misfit function, u is the wavefield, s are the sources, A is the acoustic wave operator on the frequency domain, P is the projection operator and d are the observed data.

We begin by considering the Lagrangian functional $\mathcal{L}$ associated to the constrained optimization problem. To do so, we will need to take the derivative of the Lagrangian with respect to functions with complex values (e.g. u). However, differentiation with respect to complex functions may be a very strong condition. To see this, let us present an analogue with the standard complex variable theory. For a complex variable z, the differentiation of z^* by z is not defined. We follow the approach developed by Brandwood (1983) that allows us to have a consistent framework: in their work, the authors they propose to treat independently z and z^*. In particular this means that

$$\frac{\partial z}{\partial z^*} = 0 \qquad \frac{\partial z^*}{\partial z} = 0$$

Then, the authors propose to rewrite each function $h(z)$ as a function $w(z, z^*)$ and they prove that either of the conditions $\nabla_z w = 0$ or $\nabla_{z^*} w = 0$ are necessary and sufficient to determine the stationary point of h. Moreover, the gradient $\nabla_{z^*} w$ defines the direction of maximum rate of change of h with respect to z.

Hence, we follow a similar approach, and redefine our Lagrangian each time as a function of the variables and their adjoints, assuming they are independent of one another. The stationary points will be found by deriving with respect to the conjugate, or adjoint, variable. We will therefore perform derivatives such as

$$\lim_{\epsilon \to 0} \frac{\mathcal{L}(u^\dagger + \epsilon\zeta) - \mathcal{L}(u^\dagger)}{\epsilon} = \langle \zeta, \nabla_{u^\dagger} \mathcal{L} \rangle \,, \tag{78}$$

With this in mind, our Lagrangian functional $\mathcal{L}$ has as parameters the functions $u, u^\dagger, \lambda, \lambda^\dagger$ and m, where we treat each of the variables as if they were independent. This includes considering a variable and its adjoint as independent functions. The dependence between the variables will only manifest itself at the minimum, where the imposed constraints are satisfied. Generally, when the imposed constraints are satisfied, we refer to these as realizations of the state equations. u is a complex wavefield, λ is a complex adjoint variable, also known as a Lagrange multiplier, and m is a real model function. Let the wavefileds belong to a space $\mathcal{W}$ and the adjoint wavefileds to $\mathcal{W}^\dagger$.

$$\begin{aligned} \mathcal{L}(u, u^\dagger, \lambda, \lambda^\dagger, m) &= \phi(u; m) + \frac{1}{2} \langle \lambda, A(m)u - s \rangle + \frac{1}{2} \langle A(m)u - s, \lambda \rangle & (79) \\ &= \frac{1}{2} \langle Pu - d, Pu - d \rangle + \frac{1}{2} \langle \lambda, A(m)u - s \rangle + \frac{1}{2} \langle A(m)u - s, \lambda \rangle \,. & (80) \end{aligned}$$

Since to use the derivative definition in Brandwood (1983) we require a real valued Lagrangian, we have added the conjugate of the term $\langle \lambda, A(m)u - s \rangle$. We are seeking for a saddle point of the Lagrangian, and using equations (77), the following conditions are imposed

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \lambda^\dagger} = \nabla_{\lambda^\dagger} \mathcal{L} &= 0 & (81) \\ \frac{\partial \mathcal{L}}{\partial u^\dagger} = \nabla_{u^\dagger} \mathcal{L} &= 0 & (82) \end{aligned}$$

Using the definition of the derivative, conditions (81) - (82), leads to the state and adjoint

state equations:

$$A(m)u = s \tag{83}$$

$$A^\dagger \lambda = -P^\dagger(Pu - d) \tag{84}$$

$$\nabla_m \phi(u;m) = \Re\left\{ u^\dagger \left(\frac{\partial A}{\partial m}\right)^\dagger \lambda \right\}. \tag{85}$$

Now we impose the constraints that relate the different parameters. Let $\bar{u}, \bar{u}^\dagger, \bar{\lambda}, \bar{\lambda}^\dagger$ satisfy the constraints such that $\int_\Omega \bar{\lambda}^\dagger(A(m)\bar{u} - s) = 0$ and $\int_\Omega (A(m)\bar{u} - s)^\dagger \bar{\lambda} = 0$. To find the gradient, we need to derive the Lagrangian with respect to m, and evaluate where the variables satisfy the constraints. As we are imposing the constraints, the variables u and $u^\dagger$, depend on m. We thus start by perturbing the model,

$$\begin{aligned}
\mathcal{L}(\bar{u}, \bar{u^\dagger}, \bar{\lambda}, \bar{\lambda^\dagger}, m + \epsilon\zeta) - \mathcal{L}(\bar{u}, \bar{u^\dagger}, \bar{\lambda}, \bar{\lambda^\dagger}, m) &= \frac{\epsilon}{2}\left\langle P\frac{\partial \bar{u}}{\partial m}\zeta, P\bar{u} - d \right\rangle + \frac{\epsilon}{2}\left\langle P\bar{u} - d, P\frac{\partial \bar{u}}{\partial m}\right\rangle \\
&+ \frac{\epsilon}{2}\left\langle \bar{\lambda}, A\frac{\partial \bar{u}}{\partial m}\zeta \right\rangle + \frac{\epsilon}{2}\left\langle A\frac{\partial \bar{u}}{\partial m}\zeta, \lambda \right\rangle \\
&+ \frac{\epsilon}{2}\left\langle \bar{u^\dagger}\frac{\partial A}{\partial m}^\dagger \bar{\lambda}, \zeta \right\rangle + \frac{\epsilon}{2}\left\langle \zeta, \bar{u^\dagger}\frac{\partial A}{\partial m}^\dagger \bar{\lambda} \right\rangle + O(\epsilon^2)
\end{aligned}$$

Regrouping the first four terms, we have

$$\begin{aligned}
\mathcal{L}(\bar{u}, \bar{u^\dagger}, \bar{\lambda}, \bar{\lambda^\dagger}, m + \epsilon\zeta) - \mathcal{L}(\bar{u}, \bar{u^\dagger}, \bar{\lambda}, \bar{\lambda^\dagger}, m) &= \frac{\epsilon}{2}\left\langle \zeta, \frac{\partial \bar{u}}{\partial m}^\dagger \left(P^\dagger(P\bar{u} - d) + A^\dagger\bar{\lambda}\right) \right\rangle \\
&+ \frac{\epsilon}{2}\left\langle \frac{\partial \bar{u}}{\partial m}^\dagger \left(P^\dagger(P\bar{u} - d) + A^\dagger\bar{\lambda}\right), \zeta \right\rangle \\
&+ \frac{\epsilon}{2}\left\langle \bar{u^\dagger}\frac{\partial A}{\partial m}^\dagger \bar{\lambda}, \zeta \right\rangle + \frac{\epsilon}{2}\left\langle \zeta, \bar{u^\dagger}\frac{\partial A}{\partial m}^\dagger \bar{\lambda} \right\rangle + O(\epsilon^2).
\end{aligned}$$

And using the state equation (84) derived for the adjoint field λ, the first two terms are zero, which simplifies the Lagrangian perturbation to

$$\begin{aligned}
\mathcal{L}(\bar{u}, \bar{u^\dagger}, \bar{\lambda}, \bar{\lambda^\dagger}, m + \epsilon\zeta) - \mathcal{L}(\bar{u}, \bar{u^\dagger}, \bar{\lambda}, \bar{\lambda^\dagger}, m) &= \frac{\epsilon}{2}\left\langle \bar{\lambda}, \frac{\partial A}{\partial m}\bar{u}\zeta \right\rangle + \frac{\epsilon}{2}\left\langle \frac{\partial A}{\partial m}\bar{u}\zeta, \bar{\lambda} \right\rangle && (86) \\
&= \epsilon\Re\left\{ \left\langle \bar{u^\dagger}\frac{\partial A}{\partial m}^\dagger \bar{\lambda}, \zeta \right\rangle \right\} + O(\epsilon^2). && (87)
\end{aligned}$$

In general, for two complex functions x and y we have $x^\dagger y + xy^\dagger = 2\Re(xy)$. If $y \in \mathbb{R}$, $x^\dagger y + xy^\dagger = 2\Re(x)y$. Since the perturbation ζ is real, then

$$\langle \nabla_m \mathcal{L}, \zeta \rangle = \left\langle \Re\left\{ u^\dagger \frac{\partial A}{\partial m}^\dagger \lambda \right\}, \zeta \right\rangle, \tag{88}$$

leading to the gradient equation (85). Note that the real part in the gradient naturally appears, basically due to the fact that m is real and thus, for each term in the Lagrangian that we differentiate, we also differentiate its conjugate.

4 The gradient computation for the velocity-stress elastodynamic wave equations without attenuation in conservative form, with the adjoint-state method

In FWI, the gradient computation with the adjoint state method is generally implemented with a second order self-adjoint expression of the wave equation (Plessix, 2006) . We develop our inversion scheme on a first-order velocity-stress formulation which provides a great flexibility to recast the wave equation in pseudo-conservative form. We show how to take advantage of this formalism to develop the gradient of the misfit function with the adjoint-state method in a straightforward way. This work is presented in Castellanos et al. (2011).

4.1 Introduction

In this study, we present a self-adjoint formalism which makes the implementation of the adjoint state method straightforward for the velocity-stress elastodynamic wave equations. The adjoint state method (Lions, 1972; Chavent, 2009) is a well known technique in inverse problem theory with many geophysical applications (Plessix, 2006). Even though setting up the adjoint state problem requires additional work to solve for an adjoint variable which has no primary physical interest in the solution of the inverse problem, this method is appealing because the computation of the gradient with respect to a model parameter requires two evaluations of the partial differential equations. The alternative method, which consists in the explicit computation of the Fréchet derivatives, is expensive to compute, as it requires one forward modeling for each non redundant position of source and receiver (Shin et al., 2001b).

The gradient computation with the adjoint state method is generally implemented with a second order expression of the wave equation, which is self-adjoint (Plessix, 2006). To perform the inversion in a 3D elastic media we use the first-order velocity-stress elastodynamic wave equation to simplify the numerical implementation of the gradient. However, this formulation has the disadvantage that it is not self adjoint and an additional forward modeling operator needs to be implemented.

Our aim is to develop the gradient of the misfit function from the self-adjoint formulation of the velocity-stress elastodynamic wave equation to simplify the numerical implementation of the gradient. Indeed, we have shown that we can find a linear transformation to render the forward modeling operator self-adjoint. Moreover, under this transformation, the radiation pattern matrix in the kernel of the gradient (Pratt et al., 1998) is diagonal and, therefore easy to implement in a parallel environment.

The algorithm we developed allows to perform seismic modeling using the non-conservative form of the velocity-stress wave equation, which is suitable for conventional modeling schemes, and perform in the inversion using a self-adjoint first-order formulation of the visco-acoustic wave equation. We show that this simply amounts to adapting the source term of the non-conservative wave equation.

4.2 Forward Equations

We consider the 3D velocity-stress isotropic elastodynamic wave equations without attenuation for seismic wave modeling where the sources are either punctual forces $(s_{f_x}, s_{f_y}, s_{f_z})$ or external stresses $(s_{\sigma_{xx}}, s_{\sigma_{yy}}, s_{\sigma_{zz}}, s_{\sigma_{xy}}, s_{\sigma_{xz}}, s_{\sigma_{yz}})$ applied to elementary surfaces. These excita-

tions are denoted by the vector

$$\mathbf{s} = \left(s_{f_x}, s_{f_y}, s_{f_z}, s_{\sigma_{xx}}, s_{\sigma_{yy}}, s_{\sigma_{zz}}, s_{\sigma_{xy}}, s_{\sigma_{xz}}, s_{\sigma_{yz}}\right)^t.$$

These fields satisfy the elastodynamic equations

$$\begin{aligned}
\partial_t V_x &= b\left(\partial x \sigma_{xx} + \partial_y \sigma_{xy} + \partial_z \sigma_{xz} + f_x\right) \\
\partial_t V_y &= b\left(\partial x \sigma_{xy} + \partial_y \sigma_{yy} + \partial_z \sigma_{yz} + f_y\right) \\
\partial_t V_z &= b\left(\partial z \sigma_{xz} + \partial_z \sigma_{zy} + \partial_z \sigma_{zz} + fz\right) \\
\partial_t \sigma_{xx} &= (\lambda + \mu)\,\partial_x V_x + \lambda\left(\partial_y V y + \partial z V_z\right) + P_{xx} \\
\partial_t \sigma_{yy} &= (\lambda + \mu)\,\partial_y V_y + \lambda\left(\partial_x V x + \partial z V_z\right) + P_{yy} \\
\partial_t \sigma_{zz} &= (\lambda + \mu)\,\partial_z V_z + \lambda\left(\partial_x V x + \partial y V_y\right) + P_{zz} \\
\partial_t \sigma_{xy} &= \mu\left(\partial_y V_x + \partial_x V_y\right) + P_{xy} \\
\partial_t \sigma_{xz} &= \mu\left(\partial_z V_x + \partial_x V_z\right) + P_{xz} \\
\partial_t \sigma_{yz} &= \mu\left(\partial_z V_y + \partial_y V_z\right) + P_{yz},
\end{aligned}$$

which can be recast in compact form as

$$\partial_t \mathbf{u} = \mathbf{A}\mathbf{u} + \mathbf{s}, \tag{89}$$

where the vector $\mathbf{u}$ is given by

$$\mathbf{u} = \left(v_x, v_y, v_z, \sigma_{xx}, \sigma_{yy}, \sigma_{zz}, \sigma_{xy}, \sigma_{xz}, \sigma_{yz}\right)^t. \tag{90}$$

while the matrix A is defined by

$$A = \begin{pmatrix}
0 & 0 & 0 & b\partial_x & 0 & 0 & b\partial_y & b\partial_z & 0 \\
0 & 0 & 0 & 0 & b\partial_y & 0 & b\partial_x & 0 & b\partial_z \\
0 & 0 & 0 & 0 & 0 & b\partial_z & 0 & b\partial_x & b\partial_y \\
(\lambda+2\mu)\partial_x & \lambda\partial_y & \lambda\partial_z & 0 & 0 & 0 & 0 & 0 & 0 \\
\lambda\partial_x & (\lambda+2\mu)\partial_y & \lambda\partial_z & 0 & 0 & 0 & 0 & 0 & 0 \\
\lambda\partial_x & \lambda\partial_y & (\lambda+2\mu)\partial_z & 0 & 0 & 0 & 0 & 0 & 0 \\
\mu\partial_y & \mu\partial_x & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\mu\partial_z & 0 & \mu\partial_x & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & \mu\partial_z & \mu\partial_y & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix} \tag{91}$$

In isotropic media, one can apply a change of variables through a linear operator $\mathbf{T}$ to write the system in conservative form, where all the medium properties are on the left side of equations.

$$\Lambda \partial_t \mathbf{w} = \mathbf{A}'\mathbf{w} + \mathbf{s}', \tag{92}$$

where Λ is a diagonal matrix given by

$$\mathrm{diag}(\Lambda) = \left(\rho, \rho, \rho, \frac{1}{3\lambda + 2\mu}, \frac{1}{2\mu}, \frac{1}{2\mu}, \frac{1}{\mu}, \frac{1}{\mu}, \frac{1}{\mu}\right). \tag{93}$$

The change of variable operator $\mathbf{T}$ is a linear, orthonormal invertible matrix constructed from the eigenvectors of the isotropic elastic tensor. We postpone its precise formulation in order to focus on the introduction of the main objects. Now, the new components of the wavefield $\mathbf{w} = \mathbf{T}\mathbf{u}$ are given by

$$\begin{aligned}
\mathbf{w} &= \left(v_x, v_y, v_z, \frac{1}{\sqrt{3}}\mathrm{Tr}(\sigma), \frac{\sqrt{3}}{\sqrt{2}}(\sigma_{zz} - \frac{1}{3}\mathrm{Tr}(\sigma)),\right. \\
&\quad \left.\frac{1}{\sqrt{2}}(-\sigma_{xx} + \sigma_{yy}), \sigma_{xy}, \sigma_{xz}, \sigma_{yz}\right)^t,
\end{aligned} \tag{94}$$

where the velocity components are left unchanged, while the stress tensor has been modified. Note that $\mathrm{Tr}(\sigma)$ is the hydrostatic pressure. The new matrix, A', is given by

$$\mathbf{A}' = \Lambda T \mathbf{A} T^{-1}, \tag{95}$$

it is symmetrical and does not depend on the physical properties of the medium. Finally, the corresponding source term is

$$\mathbf{s}' = \Lambda T \mathbf{s}. \tag{96}$$

4.2.a *Lagrangian and adjoint variables in the time domain*

We define a misfit function

$$J(u;m) = \int_t \int_\Omega \left(Pu(x,t,m) - d(x,t)\right)^T \left(Pu(x,t,m) - d(x,t)\right) dxdt, \tag{97}$$

where P is a restriction operator on the receiver's positions for the components we consider in the misfit function. We sum up over time and over sources/receivers data. That is $P : \mathbb{R}^q \Rightarrow \mathbb{R}^p$, where q is the dimension of the state variable, and p is the dimension of the observation subspace. For example, if $u(x,t)$ has nine components, and we are calculating it on a 2D grid of size $N_x \times N_y$, then $q = 9 \times Nx \times Ny$, and $p = 9 \times N_R$, where N_R is the number of receivers. We want to minimize the function J with respect to model parameters $\mathbf{m} = (m_i)$.

In order to estimate this gradient with the adjoint method, we introduce a Lagrangian function L defined as

$$L(u,\lambda,m) = J(u;m) + \int_0^T \int_\Omega p^*(u,t) \left(\partial_t u(x,t,m) - A(m)u(x,t,m) - S(t)\right) dxdt.$$

where the field $p(x,t)$ is the adjoint field or Lagrange multiplier. Note that in this section we have chosen to call it p (instead of λ as before). This is to avoid confusion with the Lamèse parameter commonly denoted λ.

In an analogous way as we did in section 3, we will express this Lagrangian in terms of an appropriately chosen inner product. Indeed, for $f, g \in \mathcal{W}$, functions from $\omega \times [0,T]$ in $\mathcal{C}^9$, we define

$$\langle f \mid g \rangle = \int_0^T \int_\Omega f(t,x)^* g(t,x) dxdt.$$

Hence,

$$L(u,p,m) = J(u;m) + \langle p(x,t) \mid \left(\partial_t u(x,t,m) - Au(x,t,m) - S\right) \rangle \tag{98}$$

The saddle point must satisfy two conditions related to the Lagrangian L, which can be written as

$$\frac{\partial L}{\partial p} = 0 \tag{99}$$

$$\frac{\partial L}{\partial u} = 0 \tag{100}$$

We now evaluate each of these expressions. First, $\dfrac{\partial L}{\partial p} = 0$ leads to

$$\partial_t u - Au - S = 0 \tag{101}$$

This is the equation that needs to be satisfied by u, which we call the forward modelling. Next, we use

$$\frac{\partial L}{\partial u} = 0. \tag{102}$$

We can rewrite the Lagrangian as

$$L(u,p,m) = J(u;m) + \langle p|\left(\partial_t u - Au - s\right)\rangle \tag{103}$$

$$L(u,p,m) = J(u;m) + \int_\Omega \left[p^*(T)\cdot u(T) - p^*(0)u(0)\right]dx - \langle \partial_t p|u\rangle - \langle p|\left(Au+s\right)\rangle \tag{104}$$

We can find the derivative if we perturb the Lagrangian in a direction z

$$\frac{\partial L}{\partial u}(u,p,m)\cdot z = \lim_{\epsilon\to 0}\frac{L(u+\epsilon z,p,m)-L(u,p,m)}{\epsilon} = \langle \nabla_u L|z\rangle\,. \tag{105}$$

But since

$$L(u+\epsilon z,p,m)-L(u,p,m) = J(u+\epsilon z;m)-J(u;m)+\int_\Omega (p^*(T)\cdot(u+\epsilon z)(T))\,dx \tag{106}$$

$$- \langle \partial_t p|u+\epsilon z\rangle - \langle p|\left(A(u+\epsilon z)+S\right)\rangle \tag{107}$$

$$- \int_\Omega p^*(T)\cdot u(T)dx + \langle \partial_t p|u\rangle + \langle p|\left(Au+S\right)\rangle \tag{108}$$

$$= J(u+\epsilon z;m)-J(u;m)+\epsilon\int_\Omega p(T)z(T)dx \tag{109}$$

$$-\epsilon\langle \partial_t p|z\rangle - \epsilon\langle p|Az\rangle \tag{110}$$

Dividing by ϵ, and using definition (105),

$$\langle \nabla_u L|z\rangle = \left\langle \nabla_u J - \partial_t p - A^\dagger p|z\right\rangle + \int_\Omega p(T)z(T)dx. \tag{111}$$

Since $\nabla_u L = 0$, for all directions z, we must impose $p(T)=0$. We obtain the adjoint state equation,

$$\partial_t p + A^\dagger p = \nabla_u J \tag{112}$$

The last equation we will find is related to the derivative of the Lagrangian with respect to the parameters. Again, we perturb the parameter in some direction

$$L(u,p,m+\epsilon z)-L(u,p,m) = \langle p(x,t)|\partial_t u - A(m+\epsilon z)u - s\rangle \tag{113}$$

$$- \langle p(x,t)|\partial_t u - A(m)u - s\rangle \tag{114}$$

Expanding to first order

$$A(m+\epsilon z)\approx A(m)+\epsilon\frac{\partial A}{\partial m}z + O(\epsilon^2) \tag{115}$$

$$L(u,m+\epsilon z,p)-L(u,m,p) = -\epsilon\left\langle p(x,t)|\frac{\partial A}{\partial m}zu(x,t)\right\rangle \tag{116}$$

$$\tag{117}$$

$$\langle \nabla_m L|z\rangle = \left\langle u^\dagger(x,t)\frac{\partial A}{\partial m}^\dagger p(x,t)|z\right\rangle \tag{118}$$

But, when the restrictions are satisfied, we have

$$\nabla_m L = \nabla_m J$$

Therefore,

$$\nabla_m J = u^\dagger(x,t)\frac{\partial A^\dagger}{\partial m}p(x,t) \tag{119}$$

Equations (101), (112) and (119) are the three equations that need to be solved, in this order, to find the gradient.

4.3 The Adjoint State Equations

In order to solve the adjoint state equations (112) , we need the adjoint operator $A^\dagger$. Let p be the adjoint vector,

$$p = \left(\tilde{V}_x, \tilde{V}_y, \tilde{V}_z, \tilde{\sigma}_{xx}, \tilde{\sigma}_{yy}, \tilde{\sigma}_{zz}, \tilde{\sigma}_{xy}, \tilde{\sigma}_{xz}, \tilde{\sigma}_{yz}\right).$$

To obtain the adjoint of operator (91), the transpose of the matrix must be taken and the adjoint of each component must be found. All the terms of the forward modelling operator (91), have a similar structure. Following the procedure described in the example of section 3, we can see that, for example

$$\langle f, b\partial_x g\rangle = -\langle \partial_x(bf), g\rangle$$

(assuming zero boundary conditions at the limits).

Using this procedure, the adjoint operator of (91) is given by

$$A^\dagger = -\begin{pmatrix} 0 & 0 & 0 & \partial_x(\lambda+2\mu\cdot) & \partial_x(\lambda\cdot) & \partial_x(\lambda\cdot) & \partial_y(\mu\cdot) & \partial_z(\mu\cdot) & 0 \\ 0 & 0 & 0 & \partial_y(\lambda\cdot) & \partial_y(\lambda+2\mu\cdot) & \partial_y(\lambda\cdot) & \partial_x(\mu\cdot) & 0 & \partial_z(\mu\cdot) \\ 0 & 0 & 0 & \partial_z(\lambda\cdot) & \partial_z(\lambda\cdot) & \partial_z(\lambda+2\mu\cdot) & 0 & \partial_x(\mu\cdot) & \partial_y(\mu\cdot) \\ \partial_x(b\cdot) & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \partial_y(b\cdot) & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \partial_z(b\cdot) & 0 & 0 & 0 & 0 & 0 & 0 \\ \partial_y(b\cdot) & \partial_x(b\cdot) & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \partial_z(b\cdot) & 0 & \partial_x(b\cdot) & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \partial_z(b\cdot) & \partial_y(b\cdot) & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{120}$$

With this expression we can now solve the adjoint system (150). Since $A \neq A^\dagger$, we can see that the forward and backward modelling equations are different, and now the equations include spatial derivatives of the physical parameters.

4.4 The symmetric Pseudo-Conservative Form

We can try an alternative approach to find the adjoint state variable, which consists in first finding a conservative and symmetric form of the forward modelling equations, as mentioned in Section 4.2. We will see that this approach is valid, as long as $\mu \neq 0$.

Let us rewrite the system of equations of the forward problem (89), as

$$\begin{aligned} \partial_t v &= \sum_{\alpha=\{x,y,z\}} M_\alpha \partial_\alpha \sigma + S_f \\ \partial_t \sigma &= \sum_{\alpha=\{x,y,z\}} N_\alpha \partial_\alpha v + S_P \end{aligned} \tag{121}$$

Where,

$$M_x = \begin{pmatrix} \frac{1}{\rho} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{\rho} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{\rho} & 0 \end{pmatrix}$$

$$M_y = \begin{pmatrix} 0 & 0 & 0 & \frac{1}{\rho} & 0 & 0 \\ 0 & \frac{1}{\rho} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{\rho} \end{pmatrix}$$

$$M_z = \begin{pmatrix} 0 & 0 & 0 & 0 & \frac{1}{\rho} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{\rho} \\ 0 & 0 & \frac{1}{\rho} & 0 & 0 & 0 \end{pmatrix}$$

$$N_x = \begin{pmatrix} \lambda+2\mu & 0 & 0 \\ \lambda & 0 & 0 \\ \lambda & 0 & 0 \\ 0 & \mu & 0 \\ 0 & 0 & \mu \\ 0 & 0 & 0 \end{pmatrix}, \quad N_y = \begin{pmatrix} 0 & \lambda & 0 \\ 0 & \lambda+2\mu & 0 \\ 0 & \lambda & 0 \\ \mu & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \mu \end{pmatrix}, \quad N_z = \begin{pmatrix} 0 & 0 & \lambda \\ 0 & 0 & \lambda \\ 0 & 0 & \lambda+2\mu \\ 0 & 0 & 0 \\ \mu & 0 & 0 \\ 0 & \mu & 0 \end{pmatrix}.$$

In order to arrive to a conservative formulation, we do the change of variable

$$\mathbf{w} = \mathbf{T}\ \mathbf{u},$$

where

$$\mathbf{T} = \begin{pmatrix} \mathbb{I}_{3\times 3} & 0 \\ 0 & R \end{pmatrix}. \tag{122}$$

We choose $\mathbf{T}$ to be the linear, orthonormal, invertible matrix where the columns of R are the eigenvectors of the isotropic elastic tensor $\mathbf{C}$,

$$\mathbf{C} = \begin{pmatrix} \lambda+2\mu & \lambda & \lambda & 0 & 0 & 0 \\ \lambda & \lambda+2\mu & \lambda & 0 & 0 & 0 \\ \lambda & \lambda & \lambda+2\mu & 0 & 0 & 0 \\ 0 & 0 & 0 & 2\mu & 0 & 0 \\ 0 & 0 & 0 & 0 & 2\mu & 0 \\ 0 & 0 & 0 & 0 & 0 & 2\mu \end{pmatrix}. \tag{123}$$

The transformation matrix $\mathbf{T}$ is therefore given by:

$$\mathbf{T} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & 0 & 0 & 0 \\ 0 & 0 & 0 & -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & 0 & 0 & 0 \\ 0 & 0 & 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}. \tag{124}$$

As we can see, the velocity components remain unchanged. For notation, we will use

$$\tilde{\sigma} = R\sigma \tag{125}$$

Multiplying by the identity matrix where the change of variable needs to be done,

$$\begin{aligned} \partial_t v &= \sum_{\alpha=\{x,y,z\}} M_\alpha \partial_\alpha R^{-1} R\sigma + S_f \\ R^{-1}\partial_t R\sigma &= \sum_{\alpha=\{x,y,z\}} N_\alpha \partial_\alpha v + S_P, \end{aligned}$$

Using the notation (125), leads to the set of equations

$$\begin{aligned} \partial_t v &= \sum_{\alpha=\{x,y,z\}} \partial_\alpha \tilde{M}_\alpha \tilde{\sigma} + S_f \\ \partial_t \tilde{\sigma} &= \sum_{\alpha=\{x,y,z\}} \partial_\alpha \tilde{N}_\alpha v + RS_P, \end{aligned} \tag{126}$$

where

$$\begin{aligned} \tilde{M}_\alpha &= M_\alpha R^{-1} \\ \tilde{N}_\alpha &= RN_\alpha. \end{aligned}$$

The advantage of this variable change is that it is possible to take as common factors, matrices Λ_1 and Λ_2 which contain the physical parameters:

$$\begin{aligned} \sum \partial_\alpha \tilde{M}_\alpha &= \Lambda_1^{-1} \sum \partial_\alpha M'_\alpha \\ \sum \partial_\alpha \tilde{N}_\alpha &= \Lambda_2^{-1} \sum \partial_\alpha N'_\alpha. \end{aligned}$$

with

$$\begin{aligned} \Lambda_1 &= \mathrm{diag}\,(\rho, \rho, \rho) \\ \Lambda_2 &= \mathrm{diag}\left(\frac{1}{3\lambda + 2\mu}, \frac{1}{2\mu}, \frac{1}{2\mu}, \frac{1}{\mu}, \frac{1}{\mu}, \frac{1}{\mu}\right) \end{aligned}$$

and where M'_α and N'_α no longer contain the physical parameters.

Finally, the pseudo-conservative system has the form

$$\begin{aligned}\Lambda_1\partial_t v &= \sum_{\alpha=\{x,y,z\}} \partial_\alpha M'_\alpha \tilde{\sigma} + \Lambda_1 S_f \\ \Lambda_2\partial_t \tilde{\sigma} &= \sum_{\alpha=\{x,y,z\}} \partial_\alpha N'_\alpha v + \Lambda_2 R S_P,\end{aligned} \tag{127}$$

Let

$$\begin{aligned}\sum \partial_\alpha M'_\alpha &= M'' \\ \sum \partial_\alpha N'_\alpha &= N'',\end{aligned}$$

Where we have

$$N'' = \begin{pmatrix} \frac{1}{\sqrt{3}}\partial_x & \frac{1}{\sqrt{3}}\partial_y & \frac{1}{\sqrt{3}}\partial_z \\ -\frac{1}{\sqrt{6}}\partial_x & -\frac{1}{\sqrt{6}}\partial_y & \frac{\sqrt{2}}{\sqrt{3}}\partial_z \\ -\frac{1}{\sqrt{2}}\partial_x & \frac{1}{\sqrt{2}}\partial_y & 0 \\ \partial_y & \partial_x & 0 \\ \partial_z & 0 & \partial_x \\ 0 & \partial_z & \partial_y \end{pmatrix}, \tag{128}$$

and $M''^T = N''$. The pseudo conservative formulation can be written, in condensed form, as:

$$\Lambda\partial_t \mathbf{w} = \mathbf{A}'\mathbf{w} + \mathbf{s}' \tag{129}$$

$$\Lambda = \begin{pmatrix} \Lambda_1 & 0_{3\times3} \\ 0_{6\times3} & \Lambda_2 \end{pmatrix}, \quad \mathbf{A}' = \begin{pmatrix} 0_{3\times3} & \mathbf{M}'' \\ \mathbf{N}'' & 0_{6\times6} \end{pmatrix}, \quad \mathbf{A}' = \mathbf{\Lambda T A T^{-1}} \quad \mathbf{s}' = \mathbf{\Lambda T s} \tag{130}$$

We have obtained a system where $\mathbf{A} = \mathbf{A^T}$ and $\mathbf{A}^\dagger = -\mathbf{A^T} = -\mathbf{A}$.
Starting from the conservative system (129), we can write the corresponding adjoint system.

4.4.a *Adjoint State Equations for the Conservative Formulation*

Let us consider the misfit function

$$J(\mathbf{u};\mathbf{m}) = \frac{1}{2}\left\langle \mathbf{Pu}(m) - \mathbf{d} | \mathbf{Pu}(m) - \mathbf{d} \right\rangle, \tag{131}$$

where $u(x,t) \in W$, $W = \{f : \mathbb{R}^3 \times [0,T] \longrightarrow \mathbb{R}\}$. Here we have assumed only one source, but this can be easily generalized by summation over source positions. The elements of W that satisfy the state (wave) equation are called realizations of the state equation. As before the operator P is a restriction operator on the receivers positions for the components we consider in the misfit function and we want to minimize the function J with respect to model parameters $\mathbf{m} = (m_i) \in M$. At the (global) minimum of J, we have

$$\frac{\partial J}{\partial m_i} = 0. \tag{132}$$

The cost function (131) can be recast into another formulation, starting from the general cost function given by Plessix:

$$
\begin{aligned}
J(u;m) &= \frac{1}{2}\sum_r \int_0^T \left(\mathbf{P}\mathbf{u_r}(t,m) - \mathbf{d}_r(t)\right)^2 dt && (133)\\
&= \frac{1}{2}\sum_r \int_0^T \int_\Omega \left(\mathbf{P}\mathbf{u}(x,t,m) - \mathbf{d}(x,t)\right)^2 \delta(x-x_r)dxdt && (134)\\
&= \frac{1}{2}\sum_r \left\langle \mathbf{P}\mathbf{u}(m) - \mathbf{d}\right| \left(\mathbf{P}\mathbf{u}(m) - \mathbf{d}\right)\delta(x-x_r)\rangle && (135)
\end{aligned}
$$

The minimization of the misfit function is done under the constraint (92) and $\mathbf{w} = \mathbf{T}\mathbf{u}$. One can recast the optimization problem with constraints into an unconstrained problem by introducing a Lagrangian function $L : W \times W \times W^* \times W^* \times M \to R$, defined as

$$
\begin{aligned}
L(\mathbf{u},\mathbf{w},\mathbf{p}_1,\mathbf{p}_2,\mathbf{m}) &= J(\mathbf{u};m) + \left\langle \mathbf{p}_1 \mid \left(\Lambda\partial_t\mathbf{w} - \mathbf{A}'\mathbf{w} - \mathbf{s}'\right)\right\rangle \\
&+ \left\langle \mathbf{p}_2 \mid \mathbf{w} - \mathbf{T}\mathbf{u}\right\rangle, && (136)
\end{aligned}
$$

where the fields $p_i(x,t)$ are the adjoint fields belonging to the dual space of W, *i.e.* $w, u \in W$ and $p_1, p_2 \in W^*$, and where $\langle f(x,t)|g(x,t)\rangle_W = \int_\Omega \int_0^T f^*(x,t)g(x,t)dtdx$. The adjoint fields p_i can also be referred to as Lagrange multipliers. In the construction of the Lagrangian, the fields u, w, p_1, p_2 and m, are assumed independent. The dependence between them will be found later at the minimum. At the minimum, where the constraints are satisfied

$$
L(\mathbf{u},\mathbf{w},\mathbf{p}_1,\mathbf{p}_2,\mathbf{m}) = J(\mathbf{u};\mathbf{m}) \qquad (137)
$$

It is equivalent to solve the minimization problem of a misfit function under constraint, or to find a saddle point of the Lagrangian, which can be characterised by setting the partial derivatives of the Lagrangian equal to zero.

4.4.b *State equations*

Zeroing the derivative of the Lagrangian with respect to the Lagrange multipliers leads to the state equation:

$$
\begin{aligned}
\partial L/\partial\mathbf{p}_1 &= 0 \\
\Lambda\partial_t\mathbf{w}(x,t) - \mathbf{A}'\mathbf{w}(x,t) &= \mathbf{s}'(t), && (138)
\end{aligned}
$$

which is exactly the forward problem equation in the conservative form (equation 92). The other restriction is found by

$$
\begin{aligned}
\partial L/\partial\mathbf{p}_2 &= 0 \\
\mathbf{w} &= \mathbf{T}\mathbf{u}, && (139)
\end{aligned}
$$

that corresponds to the change of variable required to transform the wave equation in pseudo-conservative form.

4.4.c *Adjoint state equations*

The relation $\partial L/\partial w = 0$ requires a rewriting of the Lagrangian using derivation by parts which gives the relation

$$
\begin{aligned}
L(\mathbf{u},\mathbf{w},\mathbf{p}_1,\mathbf{p}_2,m) = J(\mathbf{u},\mathbf{m}) &+ \int_\Omega \left(\mathbf{p}_1(T)\Lambda\mathbf{w}(T) - \mathbf{p}_1(0)\Lambda\mathbf{w}_0\right)dx \\
&- \langle\partial_t\mathbf{p}_1|\Lambda\mathbf{w}\rangle - \left\langle\mathbf{p}_1\right|\left(\mathbf{A}'\mathbf{w} + \mathbf{s}'(t)\right)\rangle + \left\langle\mathbf{p}_2 \mid \mathbf{w} - \mathbf{T}\mathbf{u}\right\rangle, \qquad (140)
\end{aligned}
$$

where we have used the initial condition $\mathbf{w}(0) = \mathbf{w}_0$. We can find the derivative if we perturb the variable $\mathbf{w}$ in the Lagrangian in an arbitrary direction $\mathbf{z}$, giving us the following definition

$$\frac{\partial L}{\partial \mathbf{w}} \cdot \mathbf{z} = \lim_{\epsilon \to 0} \frac{L(\mathbf{w} + \epsilon \mathbf{z}) - L(\mathbf{w})}{\epsilon}, \tag{141}$$

leading to the expression

$$\begin{aligned} L(\mathbf{w} + \epsilon \mathbf{z}) - L(\mathbf{w}) &= \epsilon \int_\Omega \mathbf{p}(T) \Lambda \mathbf{z}(T) dx - \epsilon \left\langle \partial_t \mathbf{p}_1 | \Lambda \ \mathbf{z} \right\rangle \\ &- \epsilon \left\langle \mathbf{p}_1 | \mathbf{A}' \ \mathbf{z} \right\rangle + \epsilon \left\langle \mathbf{p}_2 | \ \mathbf{z} \right\rangle . \end{aligned} \tag{142}$$

Following the definition (141), one can get the expression

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{w}} \cdot \mathbf{z} &= -\left\langle \Lambda^\dagger \partial_t \mathbf{p}_1 | \mathbf{z} \right\rangle - \left\langle \mathbf{A}'^\dagger \mathbf{p}_1 | \mathbf{z} \right\rangle \\ &+ \int_\Omega \mathbf{p}_1(T) \mathbf{z}(T) dx + \left\langle \mathbf{p}_2 | \ \mathbf{z} \right\rangle . \end{aligned} \tag{143}$$

If we impose that this derivative must be zero, whatever the value of $\mathbf{z}$, one must consider that $\mathbf{p}_1(T) = 0$, from which we infer the first adjoint state equation for $\mathbf{p}_1$,

$$\Lambda^\dagger \partial_t \mathbf{p}_1 + \mathbf{A}'^\dagger \mathbf{p}_1 = \mathbf{p}_2. \tag{144}$$

Since the diagonal matrix Λ is symmetric and real and $\mathbf{A}'^\dagger = -\mathbf{A}'$, the equation (144) can be simplified into

$$\Lambda \partial_t \mathbf{p}_1 - \mathbf{A}' \mathbf{p}_1 = \mathbf{p}_2. \tag{145}$$

The first-adjoint state equation shows that the field $\mathbf{p}_1$ satisfies the conservative wave equation, where the source term is the adjoint-state variable $\mathbf{p}_2$. From the third condition $\partial L / \partial \mathbf{u} = 0$, we find the relation

$$\frac{\partial L}{\partial \mathbf{u}} \cdot \mathbf{z} = \frac{\partial h}{\partial \mathbf{u}} \cdot \mathbf{z} - \left\langle \mathbf{p}_2 | \mathbf{T} \mathbf{z} \right\rangle . \tag{146}$$

Using the definition of the misfit function one can write,

$$\frac{\partial h}{\partial \mathbf{u}}(\mathbf{u}) = \sum_r \mathbf{P}^\mathbf{T} \left(\mathbf{P}\mathbf{u}(\mathbf{x}, \mathbf{t}) - \mathbf{d}(\mathbf{x}, \mathbf{t}) \right) \delta(x - x_r) \tag{147}$$

$$= \sum_r \mathbf{P}^\mathbf{T} \left(\mathbf{P}\mathbf{u}(\mathbf{x_r}, \mathbf{t}) - \mathbf{d}(\mathbf{x_r}, \mathbf{t}) \right) . \tag{148}$$

By using the minimality condition in (146):

$$\mathbf{P}^t (\mathbf{P}\mathbf{u} - \mathbf{d}) = \mathbf{T}^t \mathbf{p}_2. \tag{149}$$

Using this value of the field $\mathbf{p}_2$ in the equation (145), we obtain the expression:

$$\Lambda \partial_t \mathbf{p}_1 - \mathbf{A}' \mathbf{p}_1 = \mathbf{T}^{t^{-1}} \mathbf{P}^t (\mathbf{P}\mathbf{u} - \mathbf{d}), \tag{150}$$

which are the partial differential equations of the adjoint field $\mathbf{p}_1$. Using the relation (95), we transform back this equation into a non-conservative form as

$$\partial_t \mathbf{q}_1 - \mathbf{A}\mathbf{q}_1 = (\mathbf{T}^t \Lambda \mathbf{T})^{-1} \mathbf{P}^t (\mathbf{P}\mathbf{u} - \mathbf{d}), \tag{151}$$

with the introduction of the new field $\mathbf{q}_1$ defined by the relation $\mathbf{p}_1 = \mathbf{T}\mathbf{q}_1$. This equation allows to compute this new adjoint wavefield for the non-conservative system with a specific source term. Since we imposed $\mathbf{q}_1(T) = 0$, we have to solve the problem from a final time to the initial time considering the RHS has the excitation.

4.4.d *Gradient of the misfit function*

The last condition deals with the derivative of the Lagrangian with respect to model parameters $\mathbf{m}$. Let us first rewrite the Lagrangian as

$$\begin{aligned} L(\mathbf{u},\mathbf{w},\mathbf{p}_1,\mathbf{p}_2,\mathbf{m}) &= J(\mathbf{u};\mathbf{m}) + \left\langle \mathbf{p}_1 \mid \left(\Lambda\left(\partial_t\mathbf{w} - \mathbf{T}\mathbf{s}\right) - \mathbf{A}'\mathbf{w}\right)\right\rangle \\ &+ \left\langle \mathbf{p}_2 \mid \mathbf{w} - \mathbf{T}\mathbf{u}\right\rangle, \end{aligned} \tag{152}$$

where we have used the relation (130) to write explicitly the dependence of the source on the parameters. We perturb the model in a certain direction following the same procedure as before. We end up with the relation

$$L(\mathbf{m}+\epsilon\mathbf{z}) - L(\mathbf{m}) = \epsilon\left\langle \mathbf{p}_1(t)|\frac{\partial\Lambda}{\partial m_i}\mathbf{z}\left(\partial_t\mathbf{w} - \mathbf{T}\mathbf{s}\right)\right\rangle. \tag{153}$$

Dividing by the small parameter gives the expression of the gradient with respect to the parameter m_i at the position x_i as

$$\frac{\partial L}{\partial m_i}(\mathbf{u},\mathbf{m}) = -\left(\partial_t\mathbf{w} - \mathbf{T}\mathbf{s}\right)^t\left(\frac{\partial\Lambda}{\partial m_i}\right)^t\mathbf{p}_1, \tag{154}$$

where the summation is performed on W, namely, over time and space. Recalling condition (137), at the minimum where the constraints are satisfied

$$\frac{\partial L}{\partial m_i}(\mathbf{u},\mathbf{m}) = \frac{\partial J}{\partial m_i}(\mathbf{u};\mathbf{m}).$$

Eq. (154) can be expressed in the final non-conservative expression

$$\frac{\partial J}{\partial m_i}(\mathbf{u};\mathbf{m}) = -\left(\partial_t\mathbf{u} - \mathbf{s}\right)^t\mathbf{T}^t\left(\frac{\partial\Lambda}{\partial m_i}\right)^t\mathbf{T}\mathbf{q}_1. \tag{155}$$

This gradient expression is the one we could implement numerically when considering a time domain formulation.

The gradient will be found using the same procedure in the frequency domain. We directly jump into the final expression of the gradient as

$$\frac{\partial J}{\partial m_i}(\mathbf{u};\mathbf{m}) = \Re\left(\sum_\omega\left(i\omega\mathbf{u} + \mathbf{s}\right)^\dagger\mathbf{T}^t\left(\frac{\partial\Lambda}{\partial m_i}\right)^\dagger\mathbf{T}\mathbf{q}_1\right), \tag{156}$$

which is the gradient we estimate in the inversion in the frequency domain. Let us remark that (a) the influence of the source term is concentrated at the source, that (b) the gradient is valid everywhere even if the medium parameter μ goes or is equal to zero and that (c) the gradient is a pure local expression easing the numerical implementation.

4.4.e *Algorithm*

In summary, the workflow will be as following:

1. Compute the incident wavefield $\mathbf{u}$ with the equation (89).
2. Compute the data residuals: $\mathbf{P}\mathbf{u} - \mathbf{d}$.
3. Compute the adjoint wavefield $\mathbf{q}_1$ from the non-conservative wave equation with the equation (151).
4. Compute the gradient of the misfit function with the equation (156).

5 The Hessian : computation with the adjoint state method

We wish to compute the Hessian

$$H = \nabla^2_{m_i,m_j}\phi(u;m), \tag{157}$$

using the adjoint state method. This development for FWI is done in Métivier et al. (2013b, 2014). In this section, we return to the frequency domain equations. Previously, we calculated the directional derivative of ϕ with respect to m, in an arbitrary direction v, as $\dfrac{\partial\phi}{\partial m_i}\cdot v$. Following the same procedure used to compute the gradient from the misfit function, here we assume we know the gradient and perturb it to find the Hessian. That is, previously we found

$$\frac{\partial\phi(u)}{\partial m}\cdot v \quad = \quad g_v = \langle g, v\rangle\,, \tag{158}$$

and now we perturb g_v,

$$\frac{\partial g_v}{\partial m}\cdot\zeta \quad = \quad \frac{\partial}{\partial m}\langle g, v\rangle\cdot\zeta = \frac{\partial g^\dagger v}{\partial m}\cdot\zeta = \left\langle H^\dagger v, \zeta\right\rangle = \langle\zeta, Hv\rangle\,, \tag{159}$$

which allows us to find H, and Hv, which is the matrix vector product we are also interested in. Notice that all of the following expressions are equivalent:

$$\left\langle H^\dagger v, \zeta\right\rangle = \langle v, H\zeta\rangle = \langle\zeta, Hv\rangle = \zeta^\dagger Hv = v^\dagger H\zeta. \tag{160}$$

We shall now start by defining the Lagrangian as a real functional,

$$\begin{aligned}
\mathcal{L}(u,u^\dagger,\lambda,\lambda^\dagger,\mu_1,\mu_1^\dagger,\mu_2,\mu_2^\dagger,m) &= \frac{1}{2}\left\langle u^\dagger\left(\frac{\partial A}{\partial m}\right)^\dagger\lambda, v\right\rangle + \frac{1}{2}\left\langle v, u^\dagger\left(\frac{\partial A}{\partial m}\right)^\dagger\lambda\right\rangle && (161)\\
&+ \frac{1}{2}\left\langle\mu_1, A^\dagger\lambda + P^\dagger(Pu-d)\right\rangle + \frac{1}{2}\left\langle A^\dagger\lambda + P^\dagger(Pu-d), \mu_1\right\rangle && (162)\\
&+ \frac{1}{2}\langle\mu_2, Au-s\rangle + \frac{1}{2}\langle Au-s, \mu_2\rangle\,, && (163)
\end{aligned}$$

Where v is a real direction in which the gradient was found. At the saddle point of $\mathcal{L}$ the following conditions are satisfied,

$$\begin{aligned}
\nabla_{\mu_1^\dagger}\mathcal{L} &= 0 && (164)\\
\nabla_{\mu_2^\dagger}\mathcal{L} &= 0 && (165)\\
\nabla_{\lambda^\dagger}\mathcal{L} &= 0 && (166)\\
\nabla_{u^\dagger}\mathcal{L} &= 0 && (167)\\
\nabla_{m}\mathcal{L} &= 0 && (168)
\end{aligned}$$

Developing in the same order, the state equations are

$$\begin{aligned}
A^\dagger\lambda &= -P^\dagger(Pu-d) && (169)\\
Au &= s && (170)\\
A\mu_1^j &= \frac{\partial A}{\partial m_j}u && (171)\\
A^\dagger\mu_2^j &= -P^\dagger P\mu_1^j - \frac{\partial A}{\partial m_j}^\dagger\lambda && (172)\\
H_{ij} &= \Re\left\{u^\dagger\left(\frac{\partial^2 A}{\partial m_i m_j}\right)^\dagger\lambda + \lambda^\dagger\frac{\partial A}{\partial m_i}\mu_1^j + u^\dagger\left(\frac{\partial A}{\partial m_i}\right)^\dagger\mu_2^j\right\}. && (173)
\end{aligned}$$

We show how to develop $\nabla_{u^\dagger}\mathcal{L}$ (equation (167)) as an example, and the rest follow in a similar fashion. We begin by rewriting the Lagrangian so as to make $\langle u|$ appear explicitly. Note that the second term in the Lagrangian can be rewritten as,

$$\left\langle v, u^\dagger \left(\frac{\partial A}{\partial m}\right)^\dagger \lambda \right\rangle = \left\langle \lambda^\dagger \left(\frac{\partial A}{\partial m}\right) u, v^\dagger \right\rangle \tag{174}$$

$$= \left\langle \lambda^\dagger \left(\frac{\partial A}{\partial m}\right) u, v \right\rangle, \tag{175}$$

where we used the fact that $v = v^\dagger$ because v is real. The Lagrangian can thus be equivalently written as

$$\mathcal{L}(u, u^\dagger, \lambda, \lambda^\dagger, \mu_1, \mu_1^\dagger, \mu_2, \mu_2^\dagger, m) = \left\langle u^\dagger \left(\frac{\partial A}{\partial m}\right)^\dagger \lambda, v \right\rangle + \left\langle \lambda^\dagger \left(\frac{\partial A}{\partial m}\right) u, v \right\rangle \tag{176}$$

$$+ \left\langle \mu_1, A^\dagger\lambda + P^\dagger(Pu - d) \right\rangle + \left\langle A^\dagger\lambda + P^\dagger(Pu - d), \mu_1 \right\rangle \tag{177}$$

$$+ \langle \mu_2, Au - s \rangle + \langle Au - s, \mu_2 \rangle, \tag{178}$$

Perturbing $u^\dagger$, we obtain,

$$\mathcal{L}(u, u^\dagger + \epsilon\zeta, \lambda, \lambda^\dagger, \mu_1, \mu_1^\dagger, \mu_2, \mu_2^\dagger, m) - \mathcal{L}(u, u^\dagger, \lambda, \lambda^\dagger, \mu_1, \mu_1^\dagger, \mu_2, \mu_2^\dagger, m) =$$

$$\begin{aligned} & \frac{\epsilon}{2}\left\langle \lambda^\dagger \left(\frac{\partial A}{\partial m}\right)\zeta, v \right\rangle + \frac{\epsilon}{2}\left\langle P^\dagger P\zeta, \mu_1 \right\rangle \\ + \; & \frac{\epsilon}{2}\langle A\zeta, \mu_2 \rangle + O(\epsilon^2) \\ = \; & \frac{\epsilon}{2}\left\langle \zeta, \left(\frac{\partial A}{\partial m}\right)^\dagger \lambda v + P^\dagger P\mu_1 + A^\dagger\mu_2 \right\rangle + O(\epsilon^2). \end{aligned}$$

Diving by ϵ, taking the limit when it goes to zero, and using (160) to identify the operator that is perturbed in ζ and v ,we obtain;

$$\frac{\partial A}{\partial m}^\dagger \lambda + P^\dagger P\mu_1 + A^\dagger\mu_2 = 0, \tag{179}$$

which is equation (167). For the Hessian equation (173), the real part for each of the terms will naturally appear, in a similar way and for the same reason that the real part also appears in the gradient expression.

Note that when the restrictions for u and λ are satisfied :

$$\mathcal{L}(\bar{u}, \bar{u}^\dagger, \bar{\lambda}, \bar{\lambda}^\dagger, \mu_1, \mu_1^\dagger, \mu_2, \mu_2^\dagger, m) = \left\langle \bar{u}^\dagger \left(\frac{\partial A}{\partial m}\right)^\dagger \bar{\lambda}, v \right\rangle + \left\langle v, \bar{u}^\dagger \left(\frac{\partial A}{\partial m}\right)^\dagger \bar{\lambda} \right\rangle \tag{180}$$

$$= \left\langle 2\Re\left\{ \bar{u}^\dagger \left(\frac{\partial A}{\partial m}\right)^\dagger \bar{\lambda} \right\}, v \right\rangle, \tag{181}$$

because v is real. One may recognize that on the left hand side is the gradient, $g = \Re\left\{ \bar{u}^\dagger \left(\frac{\partial A}{\partial m}\right)^\dagger \bar{\lambda} \right\}$.

Therefore,

$$\nabla_m \mathcal{L}(\bar{u}, \bar{u}^\dagger, \bar{\lambda}, \bar{\lambda}^\dagger, \mu_1, \mu_1^\dagger, \mu_2, \mu_2^\dagger, m) = \nabla_m(g \cdot v) = Hv \tag{182}$$

If the quantity of interest is Hv, like it is in our case, the corresponding equations are

$$A\mu_1 = \frac{\partial A}{\partial m}uv = \sum_{j=1}^{N}\frac{\partial A}{\partial m_j}uv_j \tag{183}$$

$$A^\dagger\mu_2 = -P^\dagger P\mu_1 - \frac{\partial A}{\partial m}^\dagger \lambda v = -P^\dagger P\mu_1 - \sum_{j=1}^{N}\frac{\partial A}{\partial m_j}^\dagger \lambda v_j \tag{184}$$

$$(Hv)_i = u^\dagger\left(\frac{\partial^2 A}{\partial m_i m}\right)^\dagger \lambda v + \lambda^\dagger\frac{\partial A}{\partial m_i}\mu_1 + u^\dagger\left(\frac{\partial A}{\partial m_i}\right)^\dagger\mu_2. \tag{185}$$

Therefore to compute the Hessian vector product, it is necessary to solve two additional direct problems : one for μ_1 and another one for μ_2. The variable μ_1 represents the diffracted wavefield, $\mu_1 = (\partial u/\partial m)$ where the virtual source is the diffraction pattern times the direct wavefield, weighted by the direction v. Observing equation (173), we see that the terms in the Hessian involve cross correlations between (u,λ), (λ,μ_1) and (u,μ_2), weighted by a matrix that represents the radiation pattern.

If we wish to recover the Gauss - Newton approximation of the Hessian $H_{GN} = \left(\frac{\partial u}{\partial m}\right)^\dagger\left(\frac{\partial u}{\partial m}\right)$, we should only keep in the Hessian the cross correlation between the diffracted wavefields, and leave out any second order interactions. For this reason, we do not take into account the term $\left(\frac{\partial^2 A}{\partial m_i m_j}\right)$ in the Hessian because it represents a double refracted wavefield, or the terms that cross correlated the backpropagated wavefield λ with the diffracted wavefield μ_1. This eliminates completely the first two terms of equation (173), and the last term is the cross correlation of (u,μ_2). The source terms for μ_2 are the diffracted wavefield μ_1 and the back propagated wavefield λ. Since we are not interested in the cross correlation (u,λ), we eliminate the last term of equation (173), which leaves us with the following reduced system of equations

$$A\mu_1 = \frac{\partial A}{\partial m}uv \tag{186}$$

$$A^\dagger\mu_3 = -P^\dagger P\mu_1 \tag{187}$$

$$H_{GN}v = u^\dagger\left(\frac{\partial A}{\partial m}\right)^\dagger\mu_3. \tag{188}$$

Once again we need to solve two additional direct problems : one for μ_1 and one for μ_3. Notice that if we solve for μ_3 and replace it back into equation , we obtain

$$H_{GN} = -u^\dagger\left(\frac{\partial A}{\partial m}\right)^\dagger A^{-1\dagger}\left(P^\dagger P\mu_1\right) \tag{189}$$

$$H_{GN} = -\left(A^{-1}\left(\frac{\partial A}{\partial m}\right)u\right)^\dagger\left(P^\dagger P\frac{\partial u}{\partial m}\right) \tag{190}$$

$$H_{GN} = -\left(\frac{\partial u}{\partial m}\right)^\dagger P^\dagger P\left(\frac{\partial u}{\partial m}\right), \tag{191}$$

where we can recover the original expression for the Gauss Newton approximation.

Like we mentioned, through this approach, two additional direct problems need to be solved. Moreover, if $A = A^\dagger$, only one forward modeling tool is necessary, and the only term that changes is the source term. However, if another forward modelling tool is implemented, it can be possible

to reduce the number of direct problems in the Gauss Newton approximation, to only one. Once again, we start with the general expression for the Gauss Newton approximation of the Hessian

$$
\begin{aligned}
H_{GN}v &= -\left(\frac{\partial u}{\partial m}\right)^{\dagger} P^{\dagger}P\left(\frac{\partial u}{\partial m}\right)v && (192)\\
&= -u^{\dagger}\left(\frac{\partial A}{\partial m}\right)^{\dagger} A^{-1\dagger}P^{\dagger}P\mu_1 && (193)\\
&= -u^{\dagger}\left(\frac{\partial A}{\partial m}\right)^{\dagger}\mu_4 && (194)\\
& && (195)
\end{aligned}
$$

where

$$
\begin{aligned}
A^{\dagger}\mu_4 &= P^{\dagger}P\mu_1 && (196)\\
A^{\dagger}\mu_4 &= A^{-1}\left(P^{\dagger}P\frac{\partial A}{\partial m}uv\right) && (197)\\
AA^{\dagger}\mu_4 &= P^{\dagger}P\left(\frac{\partial A}{\partial m}uv\right), && (198)
\end{aligned}
$$

Therefore, if we construct the self adjoint operation $AA^{\dagger}$ (which requires more access to memory), we are able to solve only one direct problem (equation (198)) for μ_4 and replace it back in (194) to obtain the Hessian. We summarize,

$$
\begin{aligned}
AA^{\dagger}\mu_4 &= P^{\dagger}P\sum_{j=1}^{N}\left(\frac{\partial A}{\partial m_j}\right)uv_j && (199)\\
(H_{GN}v)_i &= -u^{\dagger}\left(\frac{\partial A}{\partial m_i}\right)^{\dagger}\mu_4. && (200)\\
& && (201)
\end{aligned}
$$

6 Source encoding with Dirac notation

We will now work in discrete form and rewrite the misfit function and gradient in an equivalent way using the Dirac notation, that will later simplify the extension to source encoding. We first discretize the domain Ω in N points and approximate the second-order derivative with a second-order scheme to solve the direct problem (3.4) numerically in *2D* in the frequency domain,. In discrete form, A is a $N \times N$ sparse matrix with 3 bands, where the bandwidth depends on the order of the numerical scheme to discretize the spatial derivatives. u is a column vector of dimension $N \times 1$, and each source term is a vector of dimension $N \times 1$.

6.1 Misfit function

Let $\mathbf{U} \in \mathbb{R}^{N\times N_S}$, be a matrix containing the computed wavefields in N points of a discretized grid due to N_S sources. The projection operator $\mathbf{P}$ restricts the solution of the wavefield only on the receiver positions. Assuming the wavefields for all sources are projected on the same set of receivers, $\mathbf{P} \in \mathbb{R}^{N\times N_R}$ and $\mathbf{PU} \in \mathbb{R}^{N_R\times N_s}$. Let $\mathbf{D} \in \mathbb{R}^{N_R\times N_s}$ represent the observed data matrix, where each column corresponds to the observed data of a different source. The residuals can be written in matrix form $\mathbf{R} = \mathbf{PU} - \mathbf{D}$, $\mathbf{R} \in \mathbb{R}^{N_R\times N_s}$, where there are N_s columns and

each column represents the residuals due to one source at all receiver positions. Using the Dirac notation, it is possible to rewrite the misfit function as :

$$\phi\left(\mathbf{U}\right)=\sum_{i=1}^{N_s}\left\langle \mathbf{e}_i|\mathbf{R}^\dagger\mathbf{R}|\mathbf{e}_i\right\rangle, \tag{202}$$

where $|\mathbf{e}_i\rangle = (0,0,1,\ldots,0)^T \in \mathbb{R}^{N_S}$ is a unit column vector with a non zero value in the i position. The operation $\mathbf{R}\,|\mathbf{e}_i\rangle$ projects the matrix $\mathbf{R}$ on the basis vector $|\mathbf{e}_i\rangle$. Since this is a canonical basis, this operation extracts one column of the matrix $\mathbf{R}$ corresponding to the wavefield generated by source source i at all receiver positions. Similarly, the adjoint operations can be defined. Note that $(|\mathbf{e}_i\rangle)^\dagger = \langle\mathbf{e}_i| = (0,0,1,\ldots,0) \in \mathbb{R}^{N_S}$ is a unitary row vector with a non zero value in position i, and $\langle\mathbf{e}_j|\mathbf{e}_i\rangle = \delta_{i,j}$. In an analogous way, $(\mathbf{R}\,|\mathbf{e}_i\rangle)^\dagger = \langle\mathbf{e}_i|\,\mathbf{R}^\dagger$, which extracts a row vector of $\mathbf{R}^\dagger$ of dimension $1\times N_r$, corresponding to the conjugate of the residuals generated by source i.

For any matrix $\mathbf{B}\in\mathbb{R}^{N_s\times N_s}$, the following property holds,

$$\sum_{i=1}^{N_s}\langle\mathbf{e}_i|\mathbf{B}|\mathbf{e}_i\rangle = \mathrm{Tr}\left(\mathbf{B}\sum_{i=1}^{N_s}|\mathbf{e}_i\rangle\langle\mathbf{e}_i|\right). \tag{203}$$

Because we are working in the canonical basis

$$\sum_{i=1}^{N_s}|\mathbf{e}_i\rangle\langle\mathbf{e}_i| = \mathbb{I}, \tag{204}$$

and using (203) the misfit function is simply

$$\phi\left(\mathbf{U}\right)=\mathrm{Tr}\left(\mathbf{R}^\dagger\mathbf{R}\right). \tag{205}$$

Source encoding can also be considered as a change of basis on which the wavefields are represented. Originally they were expressed on the canonical basis $|\mathbf{e}_i\rangle$, and when the sources are encoded we now project on $|\gamma_k\rangle = \sum_{i=1}^{N_s}\alpha_i^k\,|\mathbf{e}_i\rangle$. The misfit function can thus be rewritten a in the same way as before,

$$\tilde{\phi}\left(\tilde{\mathbf{U}}\right) = \sum_{k=1}^{K}\mathrm{Tr}\left\{\mathbf{R}^\dagger\mathbf{R}\,|\gamma_k\rangle\langle\gamma_k|\right\} = \mathrm{Tr}\left\{\mathbf{R}^\dagger\mathbf{R}\sum_{k=1}^{K}|\gamma_k\rangle\langle\gamma_k|\right\} \tag{206}$$

However, this time the projection is not on the identity but,

$$|\gamma_k\rangle\langle\gamma_k| = \left(\sum_{i=1}^{N_s}\alpha_i^k\,|\mathbf{e}_i\rangle\right)\left(\sum_{j=i}^{N_s}\alpha_j^{*k}\,\langle\mathbf{e}_j|\right) = \sum_{i=1}^{N_s}\alpha_i^k\alpha_i^{*k}\,|\mathbf{e}_i\rangle\langle\mathbf{e}_i| + \sum_{i=1}^{N_s}\sum_{j\neq i}^{N_s}\alpha_i^k\alpha_j^{*k}\,|\mathbf{e}_i\rangle\langle\mathbf{e}_j| \tag{207}$$

This projection operator has the form

$$|\gamma_k\rangle\langle\gamma_k| = \begin{pmatrix}\alpha_1^k\alpha^{*k}{}_1 & 0 & \ldots & 0\\ 0 & \alpha_2^k\alpha^{*k}{}_2 & \ldots & 0\\ \vdots & \vdots & \vdots & \vdots\\ 0 & 0 & \ldots & \alpha_{N_s}^k\alpha^{*k}{}_{N_s}\end{pmatrix} + \begin{pmatrix}0 & \alpha_1^k\alpha^{*k}{}_2 & \ldots & \alpha_1^k\alpha^{*k}{}_{N_s}\\ \alpha_2^k\alpha^{*k}{}_1 & 0 & \ldots & \alpha_2^k\alpha^{*k}{}_{N_s}\\ \vdots & \vdots & \vdots & \vdots\\ \alpha_{N_s}^k\alpha^{*k}{}_1 & \alpha_{N_s}^k\alpha^{*k}{}_2 & \ldots & 0\end{pmatrix} = \mathbf{\Gamma_D}+\mathbf{\Gamma_O}. \tag{208}$$

Going back to the expression of the misfit function we have,

$$\tilde{\phi}\left(\tilde{\mathbf{U}}\right) = \mathrm{Tr}\left\{\mathbf{R}^\dagger\mathbf{R}\boldsymbol{\Gamma}_\mathbf{D}\right\} + \mathrm{Tr}\left\{\mathbf{R}^\dagger\mathbf{R}\boldsymbol{\Gamma}_\mathbf{O}\right\} \tag{209}$$

Because $\boldsymbol{\Gamma}_\mathbf{D}$ is a diagonal matrix, $\mathrm{Tr}\left\{\mathbf{R}^\dagger\mathbf{R}\boldsymbol{\Gamma}_\mathbf{D}\right\}$ corresponds to the cross correlations of the residuals of the same source i, weighted by some random coefficients $\alpha_i^k\alpha^{*k}_i$. On the other hand, $\boldsymbol{\Gamma}_\mathbf{O}$ has non zero terms on all the non diagonal entries thus the term $\mathrm{Tr}\left\{\mathbf{R}^\dagger\mathbf{R}\boldsymbol{\Gamma}_\mathbf{O}\right\}$ of the misfit function will correspond to the cross correlation of residuals generated by different sources i, j $(i \neq j)$; weighted by the random coefficient $\alpha_i^k\alpha^{*k}_j$. Physically, the cross correlations of residuals generated by different sources, known as crosstalk noise, have no meaning in the misfit function, and we do not wish to take them into account. As we can see, the crosstalk arises naturally when cross correlating the residuals of the super sources, and the only way to control it is through the random coefficients α_i. We therefore impose

$$\mathbb{E}[\boldsymbol{\Gamma}_\mathbf{D}] = \mathbb{I} \tag{210}$$

$$\mathbb{E}[\boldsymbol{\Gamma}_\mathbf{O}] = 0. \tag{211}$$

This means, for each random vector $|\gamma_k\rangle$, two of its components i,j should have an average correlation that satisfies ,

$$\mathbb{E}\left[\langle\gamma_i|\gamma_j\rangle\right] = \mathbb{E}[\alpha_i^k\alpha^{*k}_j] = \delta_{i,j}. \tag{212}$$

In other words, we wish that the expected value of the projection matrix $|\gamma_k\rangle\langle\gamma_k|$, to be equal to the identity as was the case for the canonical basis, $\mathbb{E}[|\gamma_k\rangle\langle\gamma_k|] = \mathrm{Cov}(\gamma_k) = \mathbb{I}$. If the random coefficients satisfy the statistical properties (210) and (211), the expected value of the encoded misfit function is equal to the original misfit,

$$\mathbb{E}\left[\tilde{\phi}\left(\tilde{\mathbf{U}}\right)\right] = \mathrm{Tr}\left\{\mathbf{R}^\dagger\mathbf{R}\mathbb{E}\left[\boldsymbol{\Gamma}_\mathbf{D}\right]\right\} + \mathrm{Tr}\left\{\mathbf{R}^\dagger\mathbf{R}\mathbb{E}\left[\boldsymbol{\Gamma}_\mathbf{O}\right]\right\} = \mathrm{Tr}\left\{\mathbf{R}^\dagger\mathbf{R}\right\} = \phi\left(\mathbf{U}\right) \tag{213}$$

We may of course also write the misfit in the conventional form without the trace operation as

$$\tilde{\phi}\left(\mathbf{U}\right) = \sum_{k=1}^{K}\sum_{i=1}^{N_s}\alpha_i^k\alpha_i^{k^*}\left(\mathbf{R}^\dagger\mathbf{R}\right)_i^i + \sum_{k=1}^{K}\sum_{i=1}^{N_s}\sum_{j=1}^{N_s}\alpha_i^k\alpha_j^{k^*}\left(\mathbf{R}^\dagger\mathbf{R}\right)_i^j, \tag{214}$$

where $\left(\mathbf{R}^\dagger\mathbf{R}\right)_i^j$ represents the the element corresponding to the column i and row j.

6.2 Gradient

The gradient in equation (3.8) can be rewritten and expressed in any desired basis. Let $\boldsymbol{\Lambda} \in \mathbb{R}^{N_R\times N_s}$ be a matrix that contains in each column the back propagated wavefield λ for each individual source. In the canonical basis the gradient can be written as

$$\nabla_{m_p}\phi(u;m) = \mathcal{R}\left\{\mathrm{Tr}\left(\mathbf{U}^\dagger\left(\frac{\partial\mathbf{A}}{\partial m_p}\right)^\dagger\boldsymbol{\Lambda}\right)\right\}. \tag{215}$$

Similarly, when the data and sources are encoded, we project on the $|\gamma\rangle$ basis and the corresponding gradient expression is

$$\nabla_m\tilde{\phi}(u;m) = \mathcal{R}\left\{\mathrm{Tr}\left(\mathbf{U}^\dagger\left(\frac{\partial\mathbf{A}}{\partial m}\right)^\dagger\boldsymbol{\Lambda}\sum_{k=1}^{K}|\gamma_k\rangle\langle\gamma_k|\right)\right\} \tag{216}$$

$$= \mathcal{R}\left\{\mathrm{Tr}\left(\mathbf{U}^\dagger\left(\frac{\partial\mathbf{A}}{\partial m}\right)^\dagger\boldsymbol{\Lambda}\sum_{k=1}^{K}\left(\boldsymbol{\Gamma}_\mathbf{D}{}^k + \boldsymbol{\Gamma}_\mathbf{O}{}^k\right)\right)\right\}. \tag{217}$$

Using (210) and (211) once more, we see that

$$\mathbb{E}[\nabla_m \tilde{\phi}(u;m)] = \nabla_m \phi(u;m) \tag{218}$$

Adopting a steepest descent algorithm, the update of the model parameter at point p during iteration t,

$$m_p^{n+1} = m_p^0 + \mathcal{R}\left\{\sum_{l=0}^{n} \alpha^l \,\mathrm{Tr}\left(\mathbf{U^l}^\dagger \left(\frac{\partial \mathbf{A}}{\partial m_p}\right)^\dagger_{m=m^l} \mathbf{\Lambda^l} \sum_{k=1}^{K} {\mathbf{\Gamma_D}}^{k^l}\right)\right\} \tag{219}$$

$$+ \mathcal{R}\left\{\sum_{l=0}^{n} \alpha^l \,\mathrm{Tr}\left(\mathbf{U^l}^\dagger \left(\frac{\partial \mathbf{A}}{\partial m_p}\right)^\dagger_{m=m^l} \mathbf{\Lambda^l} \sum_{k=1}^{K} {\mathbf{\Gamma_O}}^{k^l}\right)\right\}. \tag{220}$$

The second term in the gradient expression containing the matrix $\mathbf{\Gamma_O}$ contains the unwanted term in the gradient that corresponds to the cross correlations between residuals from different sources. For the sake of insight, assume for one moment we only have one super source ($K = 1$), and that the original gradient is not changing throughout iterations (independent of index l),

$$m_p^{n+1} \approx m_p^0 + \mathcal{R}\left\{\alpha \,\mathrm{Tr}\left(\mathbf{U}^\dagger \left(\frac{\partial \mathbf{A}}{\partial m_p}\right)^\dagger_{m=m^l} \mathbf{\Lambda} \left[\sum_{l=0}^{n} {\mathbf{\Gamma_D}}^{l}\right]\right)\right\} \tag{221}$$

$$+ \mathcal{R}\left\{\alpha \,\mathrm{Tr}\left(\mathbf{U}^\dagger \left(\frac{\partial \mathbf{A}}{\partial m_p}\right)^\dagger_{m=m^l} \mathbf{\Lambda} \left[\sum_{l=0}^{n} {\mathbf{\Gamma_O}}^{l}\right]\right)\right\}. \tag{222}$$

We consider K to be small, so therefore it is the sum over the iterations l that will contribute in making the sample mean tend to the distribution mean in equations (210) and (211), and effectively make the second term tend to zero. Thus the importance of performing many iterations, and changing the random variables as frequently as possible.

For the truncated Newton methods, let $\mathbf{M_1} \in \mathbb{R}^{N_r \times N_s}$ be a matrix where each column represents the wavefield μ_1 for each individual source, and $\mathbf{M_2} \in \mathbb{R}^{N_r \times N_s}$ a matrix whose columns are the wavefields μ_2. One term in the Hessian in discrete vector notation is equal to,

$$\tilde{H}_{p,q} = \mathcal{R}\left\{\mathrm{Tr}\left\{\left(\mathbf{U}^\dagger \left(\frac{\partial^2 A}{\partial m_p m_q}\right)^\dagger \mathbf{\Lambda} + \mathbf{\Lambda}^\dagger \left(\frac{\partial A}{\partial m_p}\right) \mathbf{M_1^q} + \mathbf{U}^\dagger \left(\frac{\partial A}{\partial m_p}\right)^\dagger \mathbf{M_2^q}\right) \sum_{k=1}^{K} {\mathbf{\Gamma_D}}^{k}\right\}\right\} \tag{223}$$

$$+ \mathcal{R}\left\{\mathrm{Tr}\left\{\left(\mathbf{U}^\dagger \left(\frac{\partial^2 A}{\partial m_p m_q}\right)^\dagger \mathbf{\Lambda} + \mathbf{\Lambda}^\dagger \left(\frac{\partial A}{\partial m_p}\right) \mathbf{M_1^q} + \mathbf{U}^\dagger \left(\frac{\partial A}{\partial m_p}\right)^\dagger \mathbf{M_2^q}\right) \sum_{k=1}^{K} {\mathbf{\Gamma_O}}^{k}\right\}\right\} \tag{224}$$

where the second term represents the crosstalk noise that should average to zero throughout iterations.

Bibliography

Abubakar, A., van den Berg, P. M., and Habashy, T. M. (2004). A Multiplicative Regularization approach for Deblurring Problems. *IEEE Transactions on Image Processing*, 13:1524–1532.

Abubakar, A., van den Berg, P. M., and Mallorqui, J. J. (2002). Imaging of biomedical data using a multiplicative regularized contrast source inversion method. *IEEE Transactions on Microwave and Techniques*, 50:1761–1770.

Agudelo, W. (2005). Imagerie sismique quantitative de la marge convergent d'equator-colombie. Université Paris 6.

Almomin, A. and Biondi, B. (2012). Tomographic full waveform inversion: Practical and computationally feasible approach. *SEG Technical Program Expanded Abstracts 2012*, pages 1–5.

Amestoy, P., Duff, I. S., and L'Excellent, J. Y. (2000). Multifrontal parallel distributed symmetric and unsymmetric solvers. *Computer Methods in Applied Mechanics and Engineering*, 184(2-4):501–520.

Ammari, H., Bretin, E., Garnier, J., and Wahab, A. (2011). Time reversal in attenuating acoustic media. *Contemporary Mathematics*, 548:151–163.

Anagaw, A. Y. and Sacchi, M. D. (2012). Edge-preserving seismic imaging using the total variation method. *Journal of Geophysics and Engineering*, 9(2):138.

Andrei, N. (2007). A scaled BFGS preconditioned conjugate gradient algorithm for unconstrained optimization. *Applied Mathematics Letters*, 20:645–650.

Asnaashari, A., Brossier, R., Garambois, S., Audebert, F., Thore, P., and Virieux, J. (2013). Regularized seismic full waveform inversion with prior model information. *Geophysics*, 78(2):R25–R36.

Aster, R. C., Borchers, B., and Thurber, C. H. (2005). *Parameter Estimation and Inverse Probelms*. Elsevier Academic Press.

Bansal, R., Routh, P., Krebs, J., Lee, S., Baumstein, A., Anderson, J., Downey, N., Lazaratos, S., Lu, R., and Saldarriaga, S. (2013). Full wavefield inversion of ocean bottom node data. In *EAGE Technical Program Expanded Abstracts 2013*, page We1104.

Baumstein, A., Ross, W., and Lee, S. (2011). Simultaneous source elastic inversion of surface waves. In *Expanded Abstracts*, page C040. European Association of Geoscientists and Engineers.

Beale, E. (1972). A derivation of conjugate gradients. *Numerical Methods for non-linear optimization*, pages 39–43.

Beasley, C. J. (2008). A new look at marine simultaneous sources. *The Leading Edge*, 27(7):914–917.

Ben Hadj Ali, H., Operto, S., and Virieux, J. (2011). An efficient frequency-domain full waveform inversion method using simultaneous encoded sources. *Geophysics*, 76(4):R109.

Bérenger, J.-P. (1994). A perfectly matched layer for absorption of electromagnetic waves. *Journal of Computational Physics*, 114:185–200.

Berkhout, A. J. (2008). Changing the mindest in seismic data acquisition. *The Leading Edge*, 27(7):924–938.

Bertalmio, M., Vese, L., Sapiro, G., and Osher, S. (2003). Simultaneous structure and texture image inpainting. *Image Processing, IEEE Transactions on*, 12(8):882–889.

Beylkin, G. (1987). *Mathematical theory for seismic migration and spatial resolution*, pages 291–304. Blackwell scientific publications (Oxford).

Biondi, B. and Almomin, A. (2013). Tomographic full waveform inversion (TFWI) by combining FWI and wave-equation migration velocity analysis. *The Leading Edge*, September, special section: full waveform inversion:1074–1080.

Bottou, L. (1991). Stochastic gradient learning in neural networks. In *Proceedings of Neuro-Nîmes*, volume 91.

Bottou, L. and Bousquet, O. (2011). The tradeoffs of large-scale learning. *Optimization for Machine Learning*, page 351.

Bottou, L. and Le Cun, Y. (2005). On-line learning for very large data sets. *Applied Stochastic Models in Business and Industry*, 21(2):137–151.

Boyd, S. P. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Brandwood, D. (1983). A complex gradient operator and its application in adaptive array theory. *Microwaves, Optics and Antennas, IEE Proceedings H*, 130(1):11–16.

Brossier, R. (2011). Two-dimensional frequency-domain visco-elastic full waveform inversion: Parallel algorithms, optimization and performance. *Computers & Geosciences*, 37(4):444 – 455.

Brossier, R., Métivier, L., Operto, S., Ribodetti, A., and Virieux, J. (2013a). VTI acoustic equations: a first-order symmetrical PDE. Technical Report Technical report n° 50, SEISCOPE project.

Brossier, R., Operto, S., and Virieux, J. (2009a). Seismic imaging of complex onshore structures by 2D elastic frequency-domain full-waveform inversion. *Geophysics*, 74(6):WCC105–WCC118.

Brossier, R., Operto, S., and Virieux, J. (2009b). Two-dimensional seismic imaging of the Valhall model from synthetic OBC data by frequency-domain elastic full-waveform inversion. *SEG Technical Program Expanded Abstracts*, 28(1):2293–2297.

Brossier, R., Operto, S., and Virieux, J. (2010). Which data residual norm for robust elastic frequency-domain full waveform inversion? *Geophysics*, 75(3):R37–R46.

Brossier, R., Operto, S., and Virieux, J. (2013b). Velocity model building from seismic reflection data by full waveform inversion. *Geophysical Prospecting*, submitted.

Bunks, C., Salek, F. M., Zaleski, S., and Chavent, G. (1995). Multiscale seismic waveform inversion. *Geophysics*, 60(5):1457–1473.

Burstedde, C. and Ghattas, O. (2009). Algorithmic strategies for full waveform inversion: 1d experiments. *Geophysics*, 74(6):WCC37–WCC46.

Byrd, R. H., Lu, P., and Nocedal, J. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific and Statistical Computing*, 16:1190–1208.

Campillo, M. and Paul, A. (2003). Long Range correlations in the diffuse seismic coda. *Science 299, 547*.

Candes, E. and Romberg, J. (2005). l_1-MAGIC: Recovery of sparse signals via convex programming,. *Technical Report*.

Caselles, V., Chambolle, A., and Novaga, M. (2011). Total variation in imaging. In *Handbook of Mathematical Methods in Imaging*, pages 1016 – 1057. Springer.

Castellanos, C., Etienne, V., Hu, G., Operto, S., Brossier, R., and Virieux, J. (2011). Algorithmic and methodological developments towards full waveform inversion in 3d elastic media. *SEG Technical Program Expanded Abstracts*, 30:2793–2798.

Castellanos, C., Métivier, L., Operto, S., Brossier, R., and Virieux, J. (2013). Fast full waveform inversion with source encoding and second-order optimization methods. *Geophysical Journal International*, submitted.

Chambolle, A. (2004). An algorithm for total variation minimization and applications. *Journal of Mathematical Imaging and Vision*, 20:89–97.

Chambolle, A., Levine, S. E., and Lucier, B. J. (2011). An upwind finite-difference method for total variation-based image smoothing. *SIAM Journal on Imaging Sciences*, 4(1):277–299.

Chambolle, A. and Lions, P.-L. (1997). Image recovery via total variation minimization and related problems. *Numer. Math*, 76:167–188.

Chan, T. and Shen, J. (2005). *Image processing and analysis: variational, PDE, wavelet, and stochastic methods*. Siam.

Chan, T. F., Osher, S., and Shen, J. (2001). The digital tv filter and nonlinear denoising. *Image Processing, IEEE Transactions on*, 10(2):231–241.

Chavent, G. (2009). *Nonlinear least squares for inverse problems*. Springer Dordrecht Heidelberg London New York.

Chen, J. and Schuster, G. T. (1999). Resolution limits of migrated images. *Geophysics*, 64(4):1046–1053.

Choi and Alkhalifah, T. (2012). Multi-source waveform inversion of marine streamer data using the normalized wavefield. In *Expanded Abstracts*. EAGE.

Claerbout, J. (1985). *Imaging the Earth's interior*. Blackwell Scientific Publication.

Crase, E., Pica, A., Noble, M., McDonald, J., and Tarantola, A. (1990). Robust elastic non-linear waveform inversion: application to real data. *Geophysics*, 55:527–538.

Crase, E., Wideman, C., Noble, M., and Tarantola, A. (1992). Nonlinear elastic inversion of land seismic reflection data. *Journal of Geophysical Research*, 97:4685–4705.

Cupillard, P., Stehly, L., and Romanowicz, B. (2011). The one-bit noise correlation: A theory based on the concepts of coherent and incoherent noise. *Geophysical Journal International*, 184(3):1397–1414.

Dai, W., Huang, Y., and Schuster, G. T. (2013). Least-squares reverse time migration of marine data with frequency-selection encoding. *Geophysics*, 78(4):S233–S242.

Dai, W. and Schuster, G. T. (2013). Plane-wave least-squares reverse time migration. *Geophysics*, 78(4):S165–S177.

Darbon, J. and Sigelle, M. (2005). A fast and exact algorithm for total variation minimization. In *Pattern recognition and image analysis*, pages 351–359. Springer.

De Hoop, M., Fedrizzi, E., Garnier, J., Solna, K., et al. (2012). Imaging with noise blending. *Multi-Scale and High-Contrast PDE: From Modelling, to Mathematical Analysis, to Inversion*, pages 105–124.

Derode, A., Larose, E., Tanter, M., De Rosny, J., Tourin, A., Campillo, M., and Fink, M. (2003). Recovering the green's function from field-field correlations in an open scattering medium (l). *The Journal of the Acoustical Society of America*, 113:2973.

Devaney, A. (1984). Geophysical diffraction tomography. *Geoscience and Remote Sensing, IEEE Transactions on*, (1):3–13.

Devaney, A. J. and Zhang, D. (1991). Geophysical diffraction tomography in a layered background. *Wave motion*, 14:243–265.

Djikpéssé, H. A. and Tarantola, A. (1999). Multiparameter l_1 norm waveform fitting: Interpretation of gulf of mexico reflection seismograms. *Geophysics*, 64(4):1023–1035.

Engl, H. W., Hanke, M., and Neubauer, A. (2000). *Regularization of Inverse Problems*, volume 35. Kluwer Academic Publishers.

Erlangga, Y. A. (2005). *A robust and efficient iterative method for the numerical solution of the Helmholtz equation*. PhD thesis, Delft University of Technology.

Erlangga, Y. A. and Nabben, R. (2008). Deflation and balancing preconditioners for Krylov subspace methods applied to nonsymmetrix matrices. *SIAM Journal on Matrix Analysis and Applications*, 30(2):684–699.

Fink, M. (1993). Time-reversal mirrors. *Journal of Physics D: Applied Physics*, 26(9):1333–1350.

Fink, M. (2008). Time-reversal waves and super resolution. In *Journal of Physics: Conference Series*, volume 124, page 012004. IOP Publishing.

Forgues, E. and Lambaré, G. (1997). Parameterization study for acoustic and elastic ray+born inversion. *Journal of Seismic Exploration*, 6:253–278.

Freudenreich, Y. and Singh, S. (2000). Full waveform inversion for seismic data - frequency versus time domain. In *EAGE Technical Program Expanded Abstracts 2000*, page C54.

Friedlander, M. P. and Schmidt, M. (2012). Hybrid deterministic-stochastic methods for data fitting. *To appear in SIAM Journal on Scientific Computing*.

Gao, F., Atle, A., and Williamson, P. (2010). Full waveform inversion using deterministic source encoding. *SEG Technical Program Expanded Abstracts*, 29(1):1013–1017.

Garnier, J. and Papanicolaou, G. (2009). Passive sensor imaging using cross correlations of noisy signals in a scattering medium. *SIAM Journal on Imaging Sciences*, 2(2):396–437.

Gauthier, O., Virieux, J., and Tarantola, A. (1986). Two-dimensional nonlinear inversion of seismic waveform : numerical results. *Geophysics*, 51:1387–1403.

George, A. and Liu, J. W. (1981). *Computer solution of large sparse positive definite systems*. Prentice-Hall, Inc.

Gholami, Y., Brossier, R., Operto, S., Prieux, V., Ribodetti, A., and Virieux, J. (2013a). Which parametrization is suitable for acoustic VTI full waveform inversion? - Part 2: application to Valhall. *Geophysics*, 78(2):R107–R124.

Gholami, Y., Brossier, R., Operto, S., Ribodetti, A., and Virieux, J. (2013b). Which parametrization is suitable for acoustic VTI full waveform inversion? - Part 1: sensitivity and trade-off analysis. *Geophysics*, 78(2):R81–R105.

Gilbert, J. C. (1997). On the regularization of the Wolfe conditions in reduced quasi Newton methods for equality constrained optimization. *SIAM J. Optim*, 7(3):780–813.

Gill, P. and Leonard, M. W. (2003). Limited memory reduced Hessian methods for large scale unconstrained optimization. *SIAM J. Optim*, 12:380–401.

Gill, P. E. and Murray, W. (1979). Conjugate-gradient methods for large-scale nonlinear optimization. Technical report, DTIC Document.

Godwin, J. and Sava, P. (2013). A comparison of shot-encoding schemes for wave-equation migration. *Geophysical Prospecting*.

Goldstein, T. and Osher, S. (2009). The split bregman method for l1-regularized problems. *SIAM Journal on Imaging Sciences*, 2(2):323–343.

Guitton, A. (2012). Blocky regularization schemes for full waveform inversion. *Geophysical Prospecting*, 60(5):870–884.

Guitton, A. and Symes, W. W. (2003). Robust inversion of seismic data using the Huber norm. *Geophysics*, 68(4):1310–1319.

Ha, T., Chung, W., and Shin, C. (2009). Waveform inversion using a back-propagation algorithm and a Huber function norm. *Geophysics*, 74(3):R15–R24.

Habashy, T. M., Abubakar, A., Pan, G., and Belani, A. (2011). Source-receiver compression scheme for full-waveform seismic inversion. *Geophysics*, 76(4):R95–R108.

Hale, D. (2013). Dynamic warping of seismic images. *Geophysics*, 78(2):S105–S115.

Hansen, C. (1998). *Rank-defficient and discrete ill-posed problems - Numerical aspects of linear inversion*. Society for Industrial and Applied Mathematics - Mathematical modeling and computation.

Herrmann, F. J. and Li, X. (2012). Efficient least-squares imaging with sparsity promotion and compressive sensing. *Geophysical prospecting*, 60(4):696–712.

Hicks, G. J. (2002). Arbitrary source and receiver positioning in finite-difference schemes using kaiser windowed sinc functions. *Geophysics*, 67:156–166.

Hou, S., Solna, K., and Zhao, H. (2006). A direct imaging algorithm for extended targets. *Inverse Problems*, 22(4):1151.

Huang, Y. and Schuster, G. T. (2012). Multisource least-squares migration of marine streamer and land data with frequency-division encoding. *Geophysical Prospecting*, 60(4):663–680.

Hustedt, B., Operto, S., and Virieux, J. (2004). Mixed-grid and staggered-grid finite difference methods for frequency domain acoustic wave modelling. *Geophysical Journal International*, 157:1269–1296.

Jannane, M., Beydoun, W., Crase, E., Cao, D., Koren, Z., Landa, E., Mendes, M., Pica, A., Noble, M., Roeth, G., Singh, S., Snieder, R., Tarantola, A., and Trezeguet, D. (1989). Wavelengths of Earth structures that can be resolved from seismic reflection data. *Geophysics*, 54(7):906–910.

Kaltenbacher, B., Neubauer, A., and Scherzer, O. (2008). *Iterative Regularization Methods for Nonlinear Problems*. de Gruyter, Berlin, New York.

Kelley, C. (1999). *Iterative Methods for Optimization*. SIAM.

Krebs, J., Anderson, J., Hinkley, D., Neelamani, R., Lee, S., Baumstein, A., and Lacasse, M. D. (2009). Fast full-wavefield seismic inversion using encoded sources. *Geophysics*, 74(6):WCC105–WCC116.

Lailly, P. (1983). The seismic inverse problem as a sequence of before stack migrations. In Bednar, R. and Weglein, editors, *Conference on Inverse Scattering, Theory and application, Society for Industrial and Applied Mathematics, Philadelphia*, pages 206–220.

Lailly, P. (1984). The seismic inverse problem as a sequence of before stack migrations. In Bednar, R. and Weglein, editors, *Conference on Inverse Scattering, SIAM, Philadelphia*, pages 206–220. Soc. Ind. appl. Math.

Lambaré, G., Operto, S., Podvin, P., Thierry, P., and Noble, M. (2003). 3-D ray+Born migration/inversion - part 1: theory. *Geophysics*, 68:1348–1356.

Lambaré, G., Virieux, J., Madariaga, R., and Jin, S. (1992). Iterative asymptotic inversion in the acoustic approximation. *Geophysics*, 57:1138–1154.

Larmat, C., Montagner, J. P., Fink, M., and Capdeville, Y. (2006). Time-reversal imaging of seimic sources and appplication to the great Sumatra earthquake. *Geophysical Research Letters*, 33:L19312.

Laruelle, S. and Pagés, G. (2012). Stochastic approximation with averaging innovation applied to finance. *Monte Carlo Methods Appl.*, 18(1):1–51.

Lions, J. (1972). *Nonhomogeneous boundary value problems and applications*. Springer Verlag, Berlin.

Lions, J. L. (1968). *Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles*. Dunod, Paris.

Liu, D. C. and Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45:503–528.

Liu, F., Zhang, G., Morton, S. A., and Leveille, J. P. (2011). An effective imaging condition for reverse-time migration using wavefield decomposition. *Geophysics*, 76(1):S29–S39.

Loris, I., Douma, H., Nolet, G., Daubechies, I., and Regone, C. (2010). Nonlinear regularization techniques for seismic tomography. *Journal of Computational Physics*, 229:890–905.

Luo, Y. and Schuster, G. T. (1991). Wave-equation traveltime inversion. *Geophysics*, 56(5):645–653.

Ma, Y., Hale, D., Meng, Z. J., and Gong, B. (2010). Full waveform inversion with image-guided gradient. *SEG Technical Program Expanded Abstracts*, 29(1):1003–1007.

Menke, W. (1984). *Geophysical Data Analysis: Discrete Inverse Theory*. Academic Press, Inc., Orlando, USA.

Mercerat, E. and Nolet, G. (2012). Comparison of ray-and adjoint-based sensitivity kernels for body-wave seismic tomography. *Geophysical Research Letters*, 39(12).

Mercerat, E. D. and Nolet, G. (2013). On the linearity of cross-correlation delay times in finite-frequency tomography. *Geophysical Journal International*, 192(2):681–687.

Métivier, L., Bretaudeau, F., Brossier, R., Operto, S., and Virieux, J. (2014). Full waveform inversion and the truncated newton method: quantitative imaging of complex subsurface structures. *Geophysical Prospecting*, In press.

Métivier, L., Brossier, R., Virieux, J., and Operto, S. (2013a). Full waveform inversion and the truncated newton method. *SIAM Journal On Scientific Computing*, 35(2):B401–B437.

Métivier, L., Brossier, R., Virieux, J., and Operto, S. (2013b). Full waveform inversion and the truncated newton method. *SIAM Journal On Scientific Computing*, 35(2):B401–B437.

Miller, D., Oristaglio, M., and Beylkin, G. (1987). A new slant on seismic imaging: Migration and integral geometry. *Geophysics*, 52(7):943–964.

Montagner, J.-P., Larmatand, C., Capdeville, Y., Fink, M., Phung, H., and Romanowicz, B. (2012). Time-reversal method and cross-correlation techniques by normal mode theory: a three point problem. *Geophysical Journal International*, (191):637–652.

Montelli, R., Nolet, G., Dahlen, F. A., Masters, G., Engdahl, E. R., and Hung, S. H. (2004). Finite-frequency tomography reveals a variety of plumes in the mantle. *Science*, 303:338–343.

Mora, P. R. (1987). Nonlinear two-dimensional elastic inversion of multi-offset seismic data. *Geophysics*, 52:1211–1228.

Mora, P. R. (1988). Elastic wavefield inversion of reflection and transmission data. *Geophysics*, 53:750–759.

Mora, P. R. (1989). Inversion = migration + tomography. *Geophysics*, 54(12):1575–1586.

Neelamani, R., Krohn, C. E., Krebs, J. R., Deffenbaugh, M., Anderson, J. E., and Romberg, J. K. (2008). Efficient seismic forward modeling usin simultaneous random sources and sparsity. In *78th Annual SEG Conference & Exhibition, Las Vegas*. Society of Exploration Geophysicists.

Nemirovski, A. (1999). *Optimization II: Standard Numerical Methods for Nonlinear Continuous Optimization, Lecture Notes*. Technion - Israel Institute of Technology.

Nemirovskići, A. S. and ëIěUdin, D. (1983). *Problem complexity and method efficiency in optimization*. Wiley (Chichester and New York).

Neumaier, A. (1997). On convergence and restart conditions for non-linear conjugate gradient method. *Manuscript*.

Ng, M. K., Qi, L., Yang, Y.-F., and Huang, Y.-M. (2007). On semismooth newton's methods for total variation minimization. *Journal of Mathematical Imaging and Vision*, 27(3):265–276.

Nihei, K. T. and Li, X. (2007). Frequency response modelling of seismic waves using finite difference time domain with phase sensitive detection (TD-PSD). *Geophysical Journal International*, 169:1069–1078.

Nocedal, J. and Wright, S. J. (1999). *Numerical Optimization*. New York, US : Springer.

Nocedal, J. and Wright, S. J. (2006). *Numerical Optimization*. Springer, 2nd edition.

Nolet, G. (2008). *A Breviary of Seismic Tomography*. Cambridge University Press, Cambridge, UK.

Operto, S., Brossier, R., Gholami, Y., Métivier, L., Prieux, V., Ribodetti, A., and Virieux, J. (2013). A guided tour of multiparameter full waveform inversion for multicomponent data: from theory to practice. *The Leading Edge*, September, Special section Full Waveform Inversion:1040–1054.

Operto, S., Virieux, J., Amestoy, P., L'Éxcellent, J.-Y., Giraud, L., and Ben Hadj Ali, H. (2007). 3D finite-difference frequency-domain modeling of visco-acoustic wave propagation using a massively parallel direct solver: A feasibility study. *Geophysics*, 72(5):SM195–SM211.

Operto, S., Virieux, J., Ribodetti, A., and Anderson, J. E. (2009). Finite-difference frequency-domain modeling of visco-acoustic wave propagation in two-dimensional TTI media. *Geophysics*, 74 (5):T75–T95.

Osher, S., Burger, M., Goldfarb, D., Xu, J., and Yin, W. (2005). An iterative regularization method for total variation based image restoration. *SIAM Multiscale Model Simulation*, 4(2):460–489.

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.

Plessix, R. E. (2007). A Helmholtz iterative solver for 3D seismic-imaging problems. *Geophysics*, 72(5):SM185–SM194.

Plessix, R. E. (2009). Three-dimensional frequency-domain full-waveform inversion with an iterative solver. *Geophysics*, 74(6):WCC53–WCC61.

Plessix, R.-E., Baeten, G., de Maag, J. W., and ten Kroode, F. (2012). Full waveform inversion and distance separated simultaneous sweeping: a study with a land seismic data set. *Geophysical Prospecting*, 60:733 – 747.

Plessix, R. E. and Cao, Q. (2011). A parametrization study for surface seismic full waveform inversion in an acoustic vertical transversely isotropic medium. *Geophysical Journal International*, 185:539–556.

Plessix, R. E. and Perkins, C. (2010). Full waveform inversion of a deep water ocean bottom seismometer dataset. *First Break*, 28:71–78.

Powell, M. (1977). Restart Procedures for the Conjugate Gradient Method. *Mathematical Programming*, 12:241–254.

Pratt, R. G. (1990). Frequency-domain elastic modeling by finite differences: a tool for crosshole seismic imaging. *Geophysics*, 55(5):626–632.

Pratt, R. G., Shin, C., and Hicks, G. J. (1998). Gauss-Newton and full Newton methods in frequency-space seismic waveform inversion. *Geophysical Journal International*, 133:341–362.

Pratt, R. G. and Shipp, R. M. (1999). Seismic waveform inversion in the frequency domain, part II: Fault delineation in sediments using crosshole data. *Geophysics*, 64:902–914.

Pratt, R. G. and Worthington, M. H. (1990). Inverse theory applied to multi-source cross-hole tomography. Part I: acoustic wave-equation method. *Geophysical Prospecting*, 38:287–310.

Prieux, V., Brossier, R., Gholami, Y., Operto, S., Virieux, J., Barkved, O., and Kommedal, J. (2011). On the footprint of anisotropy on isotropic full waveform inversion: the Valhall case study. *Geophysical Journal International*, 187:1495–1515.

Prieux, V., Brossier, R., Operto, S., and Virieux, J. (2013a). Multiparameter full waveform inversion of multicomponent OBC data from valhall. Part 1: imaging compressional wavespeed, density and attenuation. *Geophysical Journal International*, 194(3):1640–1664.

Prieux, V., Brossier, R., Operto, S., and Virieux, J. (2013b). Multiparameter full waveform inversion of multicomponent OBC data from valhall. Part 2: imaging compressional and shear-wave velocities. *Geophysical Journal International*, 194(3):1665–1681.

Pyun, S., Shin, C., and Bednar, J. B. (2007). Comparison of waveform inversion, part 3: amplitude approach. *Geophysical Prospecting*, 55(4):477–485.

Pyun, S., Shin, C., and Son, W. (2009). Frequency-domain waveform inversion using an L1-norm objective function. In *Expanded Abstracts*, page P005. EAGE.

Ramírez, A. C. and Lewis, W. R. (2010). Regularization and full-waveform inversion: A two-step approach. In *2010 SEG Annual Meeting*.

Ravaut, C., Operto, S., Improta, L., Virieux, J., Herrero, A., and dell'Aversana, P. (2004). Multi-scale imaging of complex structures from multi-fold wide-aperture seismic data by frequency-domain full-wavefield inversions: application to a thrust belt. *Geophysical Journal International*, 159:1032–1056.

Ribodetti, A., Operto, S., Agudelo, W., Collot, J.-Y., and Virieux, J. (2011). Joint ray+born least-squares migration and simulated annealing optimization for high-resolution target-oriented quantitative seismic imaging. *Geophysics*, 76(2):R23.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407. Mathematical Reviews number (MathSciNet): MR42668; Zentralblatt MATH identifier: 0054.05901.

Romero, L. A., Ghiglia, D. C., Ober, C. C., and Morton, S. A. (2000). Phase encoding of shot records in prestack migration. *Geophysics*, 65, (2):426–436.

Roosta-Khorasani, F. and Ascher, U. (2013). Improved bounds on sample size for implicit matrix trace estimators. *Computing Research Repository*.

Routh, P., Krebs, J., Lazaratos, S., Baumstein, A., Lee, S., Cha, Y. H., Chikichev, I., Downey, N., Hinkley, D., and Anderson, J. (2011). Encoded simultaneous source full-wavefield inversion for spectrally shaped marine streamer data. In *2011 SEG Annual Meeting*.

Roux, N. L. and Fitzgibbon, A. W. (2010). A fast natural newton method. In Fürnkranz, J. and Joachims, T., editors, *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 623–630. Omnipress.

Rudin, L., Osher, S., and Fatemi, E. (1992). Nonlinear total variation based noise removal algorithms. *Physica D*, 60:259–268.

Saad, Y. (2003). *Iterative methods for sparse linear systems*. SIAM, Philadelphia.

Sava, P. and Fomel, S. (2005). Coordinate-independent angle-gathers for wave equation migration. *75th SEG meeting, Houston, Texas, USA, Expanded Abstracts*, pages 2052–2055.

Sava, P. and Fomel, S. (2006). Time-shift imaging condition in seismic migration. *Geophysics*, 71(6):S209–S217.

Schiemenz, A. and Igel, H. (2013). Accelerated 3-d full-waveform inversion using simultaneously encoded sources in the time domain: application to valhall ocean-bottom cable data. *Geophysical Journal International*, 195(3):1970–1988.

Schittkowski, K. (2011). A robust implementation of a sequential quadratic programming algorithm with successive error restoration. *Optimization Letters*, 5:283–296. 10.1007/s11590-010-0207-9.

Schraudolph, N. N., Yu, J., and Günter, S. (2007). A stochastic quasi-newton method for online convex optimization. In *In Proceedings of 11th International Conference on Artificial Intelligence and Statistics*.

Schuster, G. T. (1996). Resolution limits for crosswell migration and traveltime tomography. *Geophysical Journal International*, 127(2):427–440.

Schuster, G. T. (2007). Basics of seismic wave theory. University of Utah.

Schuster, G. T., Wang, X., Huang, Y., Dai, W., and C., B. (2011). Theory of multisource crosstalk reduction by phase encoded statics. *Geophysical Journal International*, 184:1289–303.

Shapiro, N. M., Campillo, M., Stehly, L., and Ritzwoller, M. H. (2005). High-Resolution Surface-Wave Tomography from Ambient Seismic Noise. *Science*, 307(5715):1615.

Shen, P. and Symes, W. W. (2008). Automatic velocity analysis via shot profile migration. *Geophysics*, 73(5):VE49–VE59.

Sheng, J., Leeds, A., Buddensiek, M., and Schuster, G. T. (2006). Early arrival waveform tomography on near-surface refraction data. *Geophysics*, 71(4):U47–U57.

Shewchuk, J. R. (1994). An introduction to the conjugate gradient method without the agonizing pain. Technical Report Computer Science Technical Report CMU-CS-94-125, School of computer science, Carnegie Mellon University.

Shin, C. and Ha, W. (2008). A comparison between the behavior of objective functions for waveform inversion in the frequency and laplace domains. *Geophysics*, 73(5):VE119–VE133.

Shin, C., Jang, S., and Min, D. J. (2001a). Improved amplitude preservation for prestack depth migration by inverse scattering theory. *Geophysical Prospecting*, 49:592–606.

Shin, C. and Min, D.-J. (2006). Waveform inversion using a logarithmic wavefield. *Geophysics*, 71(3):R31–R42.

Shin, C., Min, D.-J., Marfurt, K. J., Lim, H. Y., Yang, D., Cha, Y., Ko, S., Yoon, K., Ha, T., and Hong, S. (2002). Traveltime and amplitude calculations using the damped wave solution. *Geophysics*, 67:1637–1647.

Shin, C., Pyun, S., and Bednar, J. B. (2007). Comparison of waveform inversion, part 1: conventional wavefield vs logarithmic wavefield. *Geophysical Prospecting*, 55(4):449–464.

Shin, C., Yoon, K., Marfurt, K. J., Park, K., Yang, D., Lim, H. Y., Chung, S., and Shin, S. (2001b). Efficient calculation of a partial derivative wavefield using reciprocity for seismic imaging and inversion. *Geophysics*, 66(6):1856–1863.

Sirgue, L. (2003). *Inversion de la forme d'onde dans le domaine fréquentiel de données sismiques grand offset*. PhD thesis, Université Paris 11, France - Queen's University, Canada.

Sirgue, L. (2006). The importance of low frequency and large offset in waveform inversion. In *Presented at the 68th EAGE Conference & Exhibition, Vienna, EAGE*, page A037.

Sirgue, L., Barkved, O. I., Dellinger, J., Etgen, J., Albertin, U., and Kommedal, J. H. (2010). Full waveform inversion: the next leap forward in imaging at Valhall. *First Break*, 28:65–70.

Sirgue, L., Barkved, O. I., Gestel, J. P. V., Askim, O. J., and Kommedal, J. H. . (2009). 3D waveform inversion on Valhall wide-azimuth OBC. In *Presented at the 71th Annual International Meeting, EAGE, Expanded Abstracts*, page U038.

Sirgue, L. and Pratt, R. G. (2004). Efficient waveform inversion and imaging : a strategy for selecting temporal frequencies. *Geophysics*, 69(1):231–248.

Snieder, R. (1998). The role of nonlinearity in inverse problems. *Inverse Problems*, 14:387.

Snieder, R. and Trampert, J. (2000). Linear and nonlinear inverse problems. In *Geomatic Method for the Analysis of Data in the Earth Sciences*, volume 95 of *Lecture Notes in Earth Sciences*, pages 93–164. Springer Berlin Heidelberg.

Spall, J. C. (2003). *Introduction to Stochastic Search and Oprimization: Estimation, Simulation and Control*. Wiley-Interscience Series in Discrete Mathematics and Optimization. 1st edition.

Stehly, L., Fry, B., Campillo, M., Shapiro, N., Guilbert, J., Boschi, L., and Giardini, D. (2009). Tomography of the Alpine region from observations of seimic ambient noise. *Geophysical Journal International*, 178:338–350.

Strang, G. and Nguyen, T. (1996). *Wavelets and filter banks*. Wellesley-Cambridge Press, Wellesley, MA.

Symes, W. W. (2007). Reverse time migration with optimal checkpointing. *Geophysics*, 72(5):SM213–SM221.

Symes, W. W. (2008). Migration velocity analysis and waveform inversion. *Geophysical Prospecting*, 56:765–790.

Tarantola, A. (1984a). Inversion of seismic reflection data in the acoustic approximation. *Geophysics*, 49(8):1259–1266.

Tarantola, A. (1984b). Linearized inversion of seismic reflection data. *Geophysical Prospecting*, 32:998–1015.

Tarantola, A. (1986). A strategy for non linear inversion of seismic reflection data. *Geophysics*, 51(10):1893–1903.

Tarantola, A. (2005). *Inverse Problem theory and methods for model parameter estimation*. Society for Industrial and Applied Mathematics, Philadelphia.

Tarantola, A. et al. (1984). The seismic reflection inverse problem. *Inverse problems of acoustic and elastic waves*, pages 104–181.

Tarantola, A. and Valette, B. (1982). Generalized nonlinear inverse problems solved using the least square criterion. *Reviews of Geophysical and Space Physics*, 20:219–232.

Thierry, P., Operto, S., and Lambaré, G. (1999). Fast 2D ray-Born inversion/migration in complex media. *Geophysics*, 64(1):162–181.

Tikhonov, A. and Arsenin, V. (1977). *Solution of ill-posed problems*. Winston, Washington, DC.

Tromp, J., Tape, C., and Liu, Q. (2005). Seismic tomography, adjoint methods, time reversal and banana-doughnut kernels. *Geophysical Journal International*, 160:195–216.

van Gestel, J., Kommedal, J., Barkved, O., Mundal, I., Bakke, R., and Best, K. (2008). Continuous seismic surveillance of Valhall field. *The Leading Edge*, pages 1616–1621.

van Leeuwen, T., Aravkin, A., and Herrmann, F. (2010). Seismic waveform inversion by stochastic optimization. *Tech. Rep. TR-2010-5*.

van Leeuwen, T., Aravkin, A. Y., and Herrmann, F. J. (2011). Seismic waveform inversion by stochastic optimization. *International Journal of Geophysics*, Volume 2011, ID 689041:18 pages.

van Leeuwen, T. and Herrmann, F. (2012). Fast waveform inversion without source-encoding. *Geophysical Prospecting*, 61(s1):10–19.

Vese, L. A. and Osher, S. J. (2003). Modeling textures with total variation minimization and oscillating patterns in image processing. *Journal of Scientific Computing*, 19(1-3):553–572.

Vigh, D., Moldoveanu, N., Jiao, K., Huang, W., and Kapoor, J. (2013). Ultralong-offset data acquisition can complement full-waveform inversion and lead to improved subsalt imaging. *The Leading Edge*, 32(9):1116–1122.

Vigh, D. and Starr, E. W. (2008). 3D prestack plane-wave, full waveform inversion. *Geophysics*, 73:VE135–VE144.

Virieux, J. (1984). SH wave propagation in heterogeneous media, velocity-stress finite difference method. *Geophysics*, 49:1259–1266.

Virieux, J. (1986). P-SV wave propagation in heterogeneous media, velocity-stress finite difference method. *Geophysics*, 51:889–901.

Virieux, J. and Operto, S. (2009). An overview of full waveform inversion in exploration geophysics. *Geophysics*, 74(6):WCC1–WCC26.

Virieux, J. and Operto, S. (2010). *An overview of full waveform inversion in exploration geophysics*, page ISBN: 9781560802266 (13). Society of Exploration Geophysics.

Vogel, C. (2002). *Computational methods for inverse problems*. Society of Industrial and Applied Mathematics, Philadelphia.

Vogel, C. R. and Oman, M. E. (1996). Iterative methods for total variation denoising. *Society for Industrial and Applied Mathematics Journal on Scientific Computing*, 17(1):227–238.

Wang, C., Chen, X., Smola, A., and Xing, E. (2013a). Variance reduction for stochastic gradient optimization. In *Advances in Neural Information Processing Systems*, pages 181–189.

Wang, F., Chauris, H., Donno, D., and Calandra, H. (2013b). Taking advantage of wave field decomposition in full waveform inversion. In *EAGE Technical Program Expanded Abstracts 2013*, page Tu0708.

Whitney, M. L. (2009). *Theoretical and Numerical Study of Tikhonov's Regularization and Morozov's Discrepancy Principle*. PhD thesis, Georgia State University, Department of Mathematics and Statistics.

Williamson, P. R. (1990). Tomographic inversion in reflection seismology. *Geophysical Journal International*, 100:255–274.

Wittman, T. (2012). Lecture notes on mathematical imaging processing. In *Medical Imaging*. Fields Institute for Mathematics Research, Toronto.

Woodward, M. J. (1992). Wave-equation tomography. *Geophysics*, 57:15–26.

Wu, R. S. and Toksöz, M. N. (1987). Diffraction tomography and multisource holography applied to seismic imaging. *Geophysics*, 52:11–25.

Xu, S., Chauris, H., Lambaré, G., and Noble, M. (2001). Common-angle migration: a strategy for imaging complex media. *Geophysics*, 66:1877–1894.

Zhang, Y. and Sun, J. (2009). Practical issues in reverse time migration: true amplitude gathers, noise removal and harmonic source encoding. *First break*, 27:53–59.

Zhu, C., Byrd, R. H., and Nocedal, J. (1997). L-bfgs-b: Algorithm 778: L-bfgs-b, fortran routines for large scale bound constrained optimization. *ACM Transactions on Mathematical Software*, 23(4):550–560.