Sparsity Promoting Seismic Imaging and Full-waveform Inversion

by

Xiang Li

B.Sc., Jilin University, 2007

M.Sc., Jilin University, 2009

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Geophysics)

The University Of British Columbia

(Vancouver)

May 2015

Abstract

This thesis will address the large computational costs of solving least-squares migration and full-waveform inversion problems. Least-squares seismic imaging and full-waveform inversion are seismic inversion techniques that require iterative minimizations of large least-squares misfit functions. Each iteration requires an evaluation of the Jacobian operator and its adjoint, both of which require two wave-equation solves for all sources, creating prohibitive computational costs. In order to reduce costs, we utilize randomized dimensionality reduction techniques, reducing the number of sources used during inversion. The randomized dimensionality reduction techniques create subsampling related artifacts, which we mitigate by using curvelet-domain sparsity-promoting inversion techniques. Our method conducts least-squares imaging at the approximate cost of one reverse-time migration with all sources, and computes the Gauss-Newton full-waveform inversion update at roughly the cost of one gradient update with all sources. Finally, during our research of the full-waveform inversion problem, we discovered that we can utilize our method as an alternative approach to add sparse constraints on the entire velocity model by imposing sparsity constraints on each model update separately, rather than regularizing the total velocity model as typically practiced. We also observed this alternative approach yields a faster decay of the residual and model error as a function of iterations. We provided empirical arguments why and when imposing sparsity on the updates can lead to improved full-waveform inversion results.

Preface

All of the work presented in this thesis was carried out at The University of British Columbia by me, with advice, supervision and editorial authorship by my supervisor, Dr. Felix J. Herrmann.

Chapter 1 was prepared by me.

Chapter 2 was published to Geophysical Prospecting with my supervisor Felix J. Herrmann. Felix and I worked together about the theory part, we had a very careful discussion about how to introduce sparsity promotion as well as randomized subsampling ideas to the seismic imaging problem. I was responsible for all the coding and experiments setting. The manuscript is mainly written by Felix, I prepared all the figures and parameters related to the experiments in this chapter.

Chapter 3 was published to Geophysics with my supervisor Felix J. Herrmann, Aleksandr Y. Aravkin and Tristan van Leeuwen. In the first two years of my graduate study, I was in charge of the theory part with Felix's supervising. And Tristan and Aleksandr also perfected the theory from the mathematic aspect later on when they joined our group. I was in charge of testing the theory, coding and setting up the experiments. I prepared all the figures, tables and pseudocode in this chapter. Felix helped me a lot with English by writing the most part of the manuscript.

Chapter 4 was published to SPIE with my supervisor Felix J. Herrmann, Aleksandr Y. Aravkin and Tristan van Leeuwen. This chapter is a follow-up work of the previous chapter. The theory part is based on the previous chapter, but we made it more mathematical with a lot of inputs from Tristan and Aleksandr. I was in charge of setting up all experiments with inputs from Felix. All the MATLAB code scripts that produces the results included in this chapter are written by myself. I also prepared all figures. The manuscript is written by Tristan, I helped a lot on

the experiment part of the manuscript.

Chapter 5 was submitted to a journal. I was the lead of this project and responsible for all discussions, experiments and coding. I wrote the manuscript. Felix J. Herrmann and Ernie Esser give a lot of insightful suggestions to the discussions of the theoretical part, and they all helped a lot with my writing.

Chapter 6 is my original work.

I was responsible for all the code script used in this thesis. The ℓ_1 solver is a public toolbox presented by Ewout van den Berg and Michael P. Friedlander. The ℓ_2 solver is a public toolbox presented by Chris Paige and Michael Saunders. The Curvelet toolbox is presented by Emmanuel Candes, Laurent Demanet, David Donoho and Lexing Ying. All experiments in this thesis are all prepared with MATLAB as well as its parallel computing toolbox. Relative software of this thesis for academic use can be download from https://www.slim.eos.ubc.ca/. For commercial use, please contact website administer.

Table of Contents

List of Tables

List of Figures

Glossary

CS	compressive sensing
FWI	full-waveform inversion
RTM	reverse-time migration
GN	Gauss-Newton
NLLS	nonlinear least-squares
SPG	spectral projected gradient

Acknowledgments

Firstly, I would like to give my very special thanks to our late post-doc Ernie Esser (1980-2015), who passed away from complications of pneumonia on his trip back from Europe in March 2015. He was an extremely enthusiastic and extremely talented researcher who was, above all, immensely generous with his time and ideas. He was also a perfect friend and colleague who helped me a lot on the optimization parts of this thesis with all his patience.

I would like to express my special appreciation and thanks to my supervisor Professor Dr. Felix Herrmann, you have been a tremendous mentor for me. I would like to thank you for encouraging my research and for allowing me to grow as a research geophysicist. Your advice on both research as well as on my career have been priceless. I could not have imagined having a better advisor and mentor for my Ph.D study.

I would also like to thank my committee members, professor Dr. Eldad Haber, professor Dr. Michael Bostock for serving as my committee members even at hardship. I also want to thank you for your brilliant comments and suggestions, thanks to you.

Besides my advisor and committee members, I also would give my gratitude to our technician Dr. Henryk Modzelewski and Administrator Miranda Joyce for all your help and friendship.

My grateful appreciation goes to TOTAL SA and BGP Houston for the hospitality during my internship. In particular, I want to thank Dr. Fuchun Gao (TOTAL), Dr. Nanxun Dai (BGP) for your help and insightful suggestions on my work.

It would be impossible for me to go through this six years as a PhD student without the help from a group of incredible people from SLIM/EOS department.

My sincerely thanks goes to Ian Hanlon, Brendan Smithyman, Haneet Wason and Art Petrenko for your help about my writing skill; My grateful thanks are also extended to other SLIM or previous SLIM colleagues Rajiv Kumar, Dr. Yogi Erlangga, Dr. Tristan van Leeuwen, Anais Tamalet and those who helped me but not listed here.

I would like to give a grateful credit to my Chinese fellows Tu Ning (who used to be my roommate for 2 years), Tim Lin, Zhilong Fang, Sanyi Yuan, Lina Miao, Mengmeng Yang for letting me forget I am 9,000km away from home. Without you guys, I might have something to do with my homesick.

Last but not least, I would like to give my very special thanks to Daisy Mengsu Ding who I met two years ago. A few words just can not describe how beautiful my life is with you in it. You really teach me the real meanings of love.

This work was in part financially supported by the Natural Sciences and Engineering Research Council of Canada Discovery Grant (22R81254) and the Collaborative Research and Development Grant DNOISE II (375142-08). This research was carried out as part of the SINBAD II project with support from the following organizations: BG Group, BGP, CGG, Chevron, ConocoPhillips, DownUnder GeoSolutions, Hess, Petrobras, PGS, SubSalt Solutions, WesternGeco and Woodside.

To my motherland, my parents and Liguo Han

Chapter 1

Introduction

1.1 Least-squares migration

Within the field of exploration seismology, it is important to capture a clear image of the subsurface structure. Oil and gas industries rely on this information in order to evaluate the location, size and profitability of a reservoir. In order to obtain an image of the subsurface structure, seismic waves are sent from the surface of the earth and the reflected and refracted waves are recorded. Many methods have been developed to translate the recorded data into a final subsurface image.

Least-squares migration provides a number of distinct advantages over traditional imaging methods. In the oil and gas industry, seismic imaging techniques are often conducted with the Kirchhoff migration and reverse-time migration (RTM); however, these methods only use the adjoint of the linearized forward modeling system to map the observed data to the image domain, and therefore, they do not necessarily produce correct amplitude information on changes of the subsurface velocity with respect to a known background-velocity model Guitton and Verschuur (2004); Claerbout (1985); Gray (1997). Least-squares migration, in contrast, is able to produce the correct amplitude information by using optimization techniques to minimize the difference between the observed data residual and the modeling data. Additionally, researchers have reported that least-squares migration can noticeably mitigate the migration artifacts, because linearized data from migration artifacts can not fit the observed data during the inversion process Lailly

(1983); Aoki and Schuster (2009); Nemeth et al. (1999); Kühl and Sacchi (2003). Finally, the above two advantages, taken together, results in an image of higher resolution.

Despite these advantages, the main obstacle of applying the least-squares migration method to industry size data is its prohibitively high computational cost. The least-square migration usually requires us to minimize a least-squares objective function with respect to the velocity perturbation iteratively. Each iteration requires evaluations of the migration operator and the linearized modeling operator (ie. the adjoint of migration operator). Each action requires solving two forward modeling problems for all sources Nemeth et al. (1999); Plessix and Mulder (2004). Therefore, the inversion cost grows linearly with the number of sources multiplied by the number of iterations used in the inversion.

We have used the randomized dimensionality reduction technique in order to reduce the high computational cost of least-squares migration Herrmann et al. (2009b); Neelamani et al. (2010); Herrmann and Li (2012). This approach limits the number of source experiments in the least-squares migration by replacing all the sources with a subset of randomly selected sequential sources or a subset of randomly combined simultaneous sources. In this way, we greatly improve the computing efficiency because fewer modeling problems need to be solved with subsampled sources.

Although the randomized dimensionality reduction technique lowers computational cost, there are noticeable problems with the approach. First, the randomized dimensionality reduction technique introduces subsampling related artifacts. If we use the randomly combined simultaneous sources, we introduce source cross talk; if we use the randomly selected sequential sources, we introduce nonuniform illumination Herrmann et al. (2009b); Tu et al. (2013). Second, using fewer sources will make the least-squares problem more "under-determined," meaning fewer observations than unknowns, which leads to more possible solutions.

We borrow ideas from compressive sensing (CS) to address the subsampling related problems. According to CS, a sparse signal can be recovered from far fewer observations required by the Shannon-Nyquist sampling theorem by solving an sparsity promoting problem Candès et al. (2006); Donoho (2006), and the quality of the recovered signal depends on the sparsity level instead of the Nyquist sam-

pling rate. The least-squares imaging result contains geological structures that are sparse in the curvelet domain, while the subsampling related artifacts are not Kumar (2009); Herrmann et al. (2008a). Therefore, compressive sensing, together with randomized dimensionality reduction, can easily suppress subsampling related artifacts—as well as other noise that is not sparse in curvelet domain—by promoting a sparsity constraint on the least-squares imaging result in the curvelet domain.

This thesis, therefore, provides the two contributions discussed above. First, it renders the least squares migration computationally efficient, using randomized dimensionality reduction, reducing the number of required wave equation solves that are related to the number of shots used in inversion. Second, it mitigates subsampling artifacts by utilizing a curvelet domain sparsity promotion algorithm.

1.2 Full-waveform inversion

Successful application of the least-squares migration technique depends on the accuracy of the background velocity model. If the kinetic characteristic of the velocity model is not accurate, linearized modeling will generate data at the wrong time, which can not reflect the correct position of the velocity perturbation. Therefore, least-squares migration will produce an incorrect imaging result by fitting the observed data with the wrongly modeled data Herrmann et al. (2009a); Nemeth et al. (1999); Kühl and Sacchi (2003).

In the last two decades, researchers have reported that the full-waveform inversion (FWI) technique has the potential to build up an accurate velocity model Bunks et al. (1995); Pratt et al. (1998a); Virieux and Operto (2009); Li et al. (2012); Mtivier et al. (2013). Most commonly, FWI is formed into a nonlinear least-squares optimization problem where the medium parameters are obtained by minimizing the least-squares misfit between observed and synthetic data. There are many ways to solve the FWI optimization problem, such as gradient descent Tarantola (1984a), conjugate-gradient descent Gilbert and Nocedal (1992), Gauss-newton Pratt et al. (1998a); Li et al. (2012) and quasi-Newton Liu and Nocedal. (1989). The schematic workflow is shown in Figure 1.1.

In this thesis, we investigate the Gauss-Newton method for the following rea-

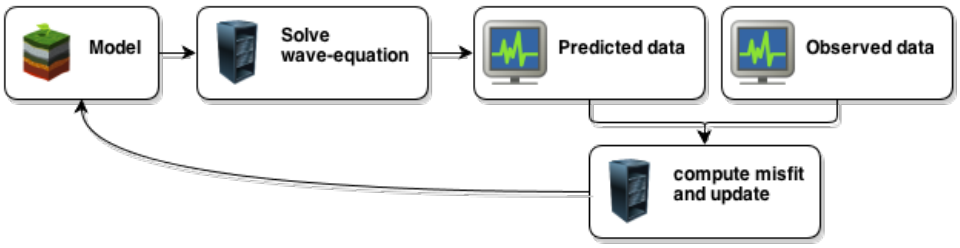


Figure 1.1: schematic FWI workflow (image courtesy Tristan van Leeuwen)

sons. Firstly, the Gauss-Newton method is a second-order optimization method that has the potential to achieve better convergence behavior compared to first-order gradient-based methods Pratt et al. (1998a); Gratton et al. (2007); Mtivier et al. (2013). Secondly, the Gauss-Newton method approximates the nonlinear least-squares problem with a sequence of linearized least-squares subproblems which require us to invert the Gauss-Newton Hessian (an approximation of the true Hessian) Gratton et al. (2007). The Gauss-Newton Hessian is symmetric and (semi-)positive definite, which can be inverted more easily than the true Hessian Hestenes and Stiefel (1952).

Using Gauss-Newton method for FWI is expensive. The Gauss-Newton Hessian of FWI objective function is formed by compounding the linearized modeling operator and its adjoint (ie. the RTM operator). To invert the Gauss-Newton Hessian, we have to evaluate the linearized modeling and the RTM operator several times, requiring a large number of wave-equation (ie. partial differential equation) solves. This could be extremely computationally expensive for large scaled datasets.

As discussed in the previous section, in order to reduce prohibitively high costs, again we utilize randomized dimentionality reduction and the sparse promoting inversion technique to mitigate subsampling related artifacts created by this approach Li et al. (2012). In our study, we find that the modified Gauss-Newton method generates better results by imposing sparsity to the updates. Upon this discovery, we analyze why it is that the modified Gauss-Newton method yields improved results and determine the situations under which it can work.

This thesis provides two main contributions. We develop a computationally efficient alogorithm which is a modification of the Gauss-Newton method, allowing

us to solve the Gauss-Newton subproblem by reducing data volume at the cost of roughly one gradient update for the fully sampled wavefield. Additionally, we analyze our modified Gauss-Newton approach and understand its behavior by means of carefully selected examples.

1.3 Thesis theme

This thesis will provide a practical and efficient approach to the seismic imaging and waveform inversion problem by using the following key strategies:

- **Randomized dimensionality-reduction** Rather than using all the sources, we use a randomly selected subset for least-square migration and the Gauss-Newton FWI.
- **Sparsity-promoting inversion** We impose a sparsity constraint on the least-square imaging result and the Gauss-Newton updates of FWI.
- **Curvelet transform** We use the curvelet to represent the least-square imaging results and the Gauss-Newton updates, allowing us to efficiently represent geological structures and mitigate incoherent subsampling related artifacts in the inversion results Herrmann (2003).

1.4 Objectives

The main objectives of the this thesis are:

- To develop sparsity-promoting seismic imaging and FWI algorithms with randomized dimensionality reduction techniques.
- To evaluate our sparsity promoting seismic inversion algorithms and understand their performances.
- To test our algorithms on datasets that have typical seismic inversion difficulties, such as lack of refraction waves, poor signal-to-noise ratio, and unmodeled shear waves.

1.5 Thesis outline

In chapter 2, we first introduce the randomized dimensionality reduction and compressive sensing techniques. Then, we explain how these techniques reduce the computational burden of large-scale imaging problems. We conclude by applying the proposed method to a seismic imaging problem with well-log constrained complexity.

Chapter 3 shows how we can lower computational costs by applying dimensionality reduction and compressive sensing techniques to the FWI problem to formulate the modified Gauss-Newton method. We test our approach with synthetic data that has a lack of refracted waves from the deep part of the velocity model, because of the low velocity structure in the middle of the model. Empirically, we find that the modified Gauss-Newton method with sparsity constraints on the updates provides better convergence behavior and resolution compared to the same method, but without promoting sparsity.

In chapter 4, we explain the modified Gauss-Newton algorithm for nonlinear optimization problems and provide a detailed formulation. We give a brief overview of stochastic optimization and CS techniques in this chapter. Finally, we illustrate the performance of the modified Gauss-Newton algorithm with a synthetic FWI problem, which is created with a realistic geological velocity model.

In chapter 5, we analyze the modified Gauss-Newton method to understand why it generates a sparse final iterate of the initial model in the curvelet domain without changing the original objective function. We demonstrate when to expect the modified Gauss-Newton method to yield a solution with a overall sparse update with respect to the starting model, and in what circumstances we should use it in place of other algorithms (such as standard Gauss-Newton). We illustrate the behavior of these methods by testing them on several different examples, such as the phase retrieval problem, synthetic FWI and the Chevron Gulf of Mexico blind FWI test.

In chapter 6, we summarize the arguments listed above, discussing recommendations for future research.

Chapter 2

Efficient least-squares imaging with sparsity promotion and compressive sensing

2.1 Introduction

Modern-day seismic imaging technology depends increasingly on computationally and data-intensive "wave-equation" migration, which relies on full acquisition and high-fidelity wavefield simulations (see e.g. Rickett, 2003; Plessix and Mulder, 2004). These challenges are compounded by a lack of available direct solvers for the time-harmonic Helmholtz equation in 3D. This is problematic because each source requires a separate partial-differential equation (PDE) solve for indirect methods and this leads to simulation costs that increase linearly with the number of source experiments. This explains current-day interest in dimensionality-reduction techniques that aim to reduce exponentially growing data volumes acquired with exceedingly many sources.

Motivated by early work of Morton and Ober (1998); Romero et al. (2000) and more recently by Ayeni (2010); Fei et al. (2010), we address the challenge of the

A version of this chapter has been published. Felix J. Herrmann and Xiang Li. Efficient least-squares imaging with sparsity promotion and compressive sensing. *Geophysical Prospecting*, 2012.

"curse of dimensionality" by decreasing the number of source experiments. As a result, we lower the computational burden of imaging significantly. To accomplish this goal, we extend the randomized dimensionality-reduction ideas presented by Herrmann et al. (2009b); Neelamani et al. (2010) to the imaging problem.

Seismic imaging entails the inversion of an extremely large, but in the absence of noise, consistent "overdetermined" systems of equations. Even though there are generally more equations than unknowns, imaging is plagued by finite aperture and shadow zones, which make this system ill conditioned (Symes, 2008b). Ill conditioning, in conjunction with extreme high costs of applying imaging operators, challenges iterative solution methods for least-squares imaging problems. To address this issue, we combine ideas from stochastic optimization (Bertsekas and Tsitsiklis, 1996; Shapiro et al., 2009; Nemirovski et al., 2009a; Haber et al., 2010a) and compressive sensing (CS—in short throughout this chapter, Candès et al., 2006; Donoho, 2006; Mallat, 2009), yielding a formulation where we invert the large linearized system by solving a sequence of much smaller subproblems that act on source-encoded "supershots" (Li and Herrmann, 2010a).

The presented approach differs from deterministic approaches, which include preconditioning, based on approximations of the wave-equation Hessian; data-dependent source syntheses, based on singular-value decompositions of the data matrix (Habashy et al., 2010); or the replacement of the Frobenius-norm (ℓ_2) by the matrix norm on the data residue (Symes, 2010). Instead, our method proposes to reduce the problem size. But contrary to Sirgue and Pratt (2004); Mulder and Plessix (2004), who select deterministic subsets of angular frequencies in their imaging, we are motivated by recent ideas from Krebs et al. (2009a); Haber et al. (2010a) who utilize source-encoding to reduce the dimensionality of full-waveform inversion.

The outline of this chapter is as follows. First, we motivate how stochastic optimization and compressive sensing are related to dimensionality reduction by randomized phase encoding in seismic imaging. Next, we introduce mini batches as collections of supershots, obtained by randomized sampling along the source and frequency axes. Inspired by stochastic optimization, we propose a solution to large-scale imaging problems via sequences of smaller dimensionality-reduced least-squares subproblems with or without sparsity constraints. Next, we identify

these constrained subproblems as relaxed sparsity-promoting problems employed by large-scale one-norm solvers. We show that this leads to an efficient algorithm, which we subsequently analyze by performing a series of controlled imaging experiments. We conclude by applying the proposed method to a seismic imaging problem with well-log constrained complexity.

2.2 Motivation

After discretization, seismic imaging involves inversion of the linearized (time-harmonic) acoustic Born-scattering matrix linking data, collected in the vector $\mathbf{b} \in \mathbb{C}^{N_f N_r N_s}$ with N_f, N_r, and N_s the number of angular frequencies, receiver, and source positions, to perturbations in the medium parameters, collected in the vector $\mathbf{x} \in \mathbb{R}^M$, with M the number of gridpoints of the model. Without loss of generality, we will keep the density of mass fixed.

Because angular frequencies and sequential sources can be treated independently, the linearized inversion has the following separable form:

$$\underset{\mathbf{x}}{\text{minimize}} \frac{1}{2K}\|\mathbf{b}-\mathbf{A}\mathbf{x}\|_2^2 = \frac{1}{2K}\sum_{i=1}^{K}\|\mathbf{b}_i-\mathbf{A}_i\mathbf{x}\|_2^2, \tag{2.1}$$

with $K = N_f N_s$ the batch size, given by the total number of monochromatic sources. The vectors $\mathbf{b}_i \in \mathbb{C}^{N_r}$ represent the corresponding vectorized monochromatic pre-processed (free of surface-multiples and direct waves) shot records. The matrix $\mathbf{A}_i$ represents the linearized scattering matrix for a specific combination of frequency and source.

Unfortunately, solving this problem is problematic because each iteration to solve Equation 2.1 typically requires $4K$ PDE solves: two to compute the action of $\mathbf{A}_i$ and two for the action of its adjoint $\mathbf{A}_i^H$ in the realistic situation where the wavefields are computed on the fly. Both actions involve solutions of the forward source and reverse-time residual wavefields (Plessix and Mulder, 2004). Thus, the inversion costs grow linearly with the number of monochromatic experiments, multiplied by the number of matrix-vector multiplies required by the solver. (Because of the size of the problem, the matrices $\mathbf{A}_i$, $i = 1 \cdots K$ can not be formed explicitly and we have to rely on iterative methods to solve equation 2.1.)

While preconditioning techniques improve convergence of Lanczos methods (Herrmann et al., 2009a), these iterative techniques require multiple evaluations of the scattering operator and its adjoint. Unfortunately, these multiple passes through the complete data are computationally intractable. To overcome this difficulty, we use a dimensionality-reduction approach where sequential sources are replaced by a reduced number of simultaneous sources made of randomized superpositions. In this way, we not only exploit linearity of the wave equation with respect to the sources but we also use the fact that randomized simultaneous sources, where sources fire at each sequential source location with random amplitude encoding, have a richer wavenumber content. This improves the image quality, albeit the resulting image can be extremely noisy. Juxtapose Figure 2.1a, obtained with a single sequential shot, with Figure 2.1b obtained from a single simultaneous shot. The key contribution of this chapter is to mitigate this noisy source cross talk while still benefiting from computational gains related to the reduction of the number of required PDE solves. Before we outline the details of our randomized imaging algorithm, let us first briefly discuss recent developments in optimization and theoretical signal analysis that provide insights and justifications for our method.

2.2.1 Stochastic optimization

In machine learning, separable optimization problems (cf. Equation 2.1) can be solved efficiently by either the stochastic-average approximation (SAA) or by the stochastic approximation (SA, Bertsekas and Tsitsiklis, 1996; Nemirovski et al., 2009a; Haber et al., 2010a). With SAA, Equation 2.1 is solved by approximating the expectation of the ensemble average by the sample average, i.e.,

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}}\, \frac{1}{2K}\|\mathbf{b}-\mathbf{A}\mathbf{x}\|_2^2 &= \mathbf{E}_{\mathbf{w}}\{\|\mathbf{b}_{\mathbf{w}}-\mathbf{A}_{\mathbf{w}}\mathbf{x}\|_2^2\} \\ &\approx \frac{1}{2K'}\sum_{i=1}^{K'}\|\mathbf{b}_i-\mathbf{A}_i\mathbf{x}\|_2^2. \end{aligned}$$

The above expression holds when $\mathbf{E}_{\mathbf{w}}\{\mathbf{w}\mathbf{w}^T\}=\mathbf{I}$ for the random vectors $\mathbf{w}$ that determine the randomized shot experiments and forms the basis for stochastic-average approximation where the expectation is approximated by selecting ran-

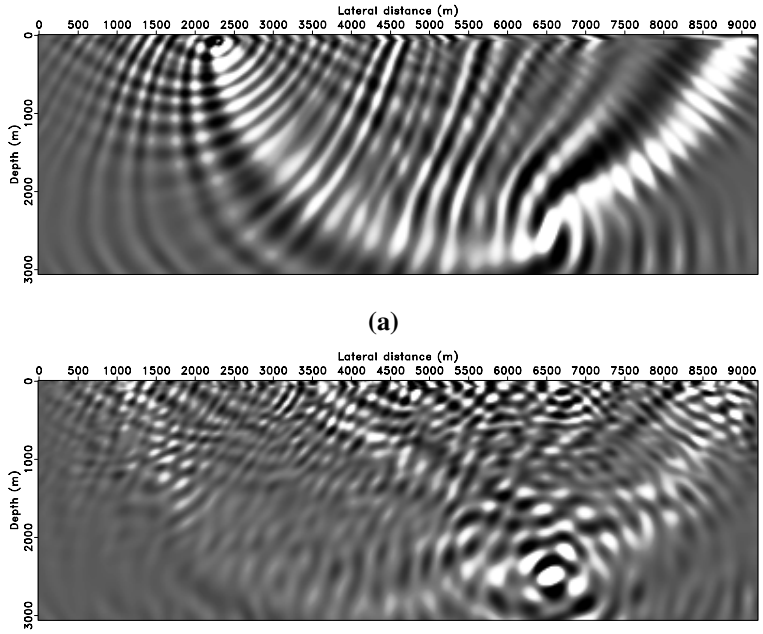


Figure 2.1: Migration of a single spike positioned in the target zone of the Marmousi model with a single sequential or simultaneous source. **(a)** Migrated spike for one sequential shot. **(b)** The same but now one single simultaneous shot. Notice the improved image of the randomized simultaneous source due to the increased wavenumber diversity.

domized subsets of frequencies and shots (possibly simultaneous shots), yielding $K' = n_f n_s \ll K$ for the reduced batch size with $n_f \ll N_f$ and $n_s \ll N_s$. For our imaging problem, this approach corresponds to carrying out least-squares migration with a randomized subset of shots. As shown by van Leeuwen et al. (2011a), the resulting error in the migrated image decay only slowly with increasing batchsize K'. This can be understood because stochastic-average approximation is essentially a Monte-Carlo sampling method, where errors decays only slowly ($\mathcal{O}(\sqrt{K'})$) with increasing batch size. Despite this disadvantage, stochastic-average approximation is popular because of its relative simplicity that offers flexibility with respect to the choice of solvers for the separable optimization in Equation 2.1. This flexibility allows us to use generic solvers such as LSQR (Paige and Saunders, 1982).

To address the relative slow convergence of stochastic-average approximation, stochastic approximation directly intervenes in first-order optimization by computing gradients on randomized subsets of data, i.e., we approximate the gradient at the k^{th} iteration of Equation (2.1) by

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \gamma_k \mathbf{A}_k^H \left(\mathbf{b}_k - \mathbf{A}_k \mathbf{x}^k \right), \tag{2.2}$$

with γ_k the step sizes. As before, $\mathbf{b}_k$ and $\mathbf{A}_k$ are the data pertaining to a randomized source experiment and the corresponding modeling operator. However, unlike before independent randomized experiments are selected after each gradient update. This procedure turns deterministic gradient descent, which involves working with all data, into stochastic-gradient descent (Bertsekas and Tsitsiklis, 1996; Nemirovski et al., 2009a; Haber et al., 2010a), which in our problem corresponds to randomly selecting a single different monochromatic source—e.g., a different $\mathbf{A}_i$ with $i \subset [1 \cdots K]$—for each gradient update pertaining to Equation 2.1. For linear problems, this approach is reminiscent of randomized "block Kaczmarz" (Strohmer and Vershynin, 2009), which was used successfully in the deterministic case by Natterer (2001) in tomography. Like stochastic-average approximation, stochastic approximation extends to nonlinear inversion problems (see e.g., Nemirovski et al., 2009a), and was recently introduced by Haber et al. (2010a) in the context of parameter-estimation problems with PDE's. Stochastic approximation also justifies recent work by Krebs et al. (2009a) and provides a theoretical explanation for observed lack of convergence— stochastic approximation only converges with $\mathcal{O}(1/k)$ with k the number of iterations (Nemirovski et al., 2009a)—and instabilities with respect to noise (van Leeuwen et al., 2011a).

2.2.2 Compressive sensing

Randomized-dimensionality reduction also underlies recent advances in sampling theory for signals that exhibit structure, which translates into transform-domain sparsity. To be more specific, compressive sensing determines conditions for which we can recover sparse vectors $\mathbf{x}$ from measurements $\mathbf{b} = \mathbf{A}\mathbf{x}$, where $\mathbf{b} \in \mathbb{R}^n$ and $\mathbf{A} \in \mathbb{R}^{n \times N}$ with $n \ll N$. When $\mathbf{x}$ is k-sparse or can be approximated accurately by its k-largest entries, stable recovery is possible for certain matrices $\mathbf{A}$ as long as

$n \gtrsim k \log(N/n)$. This results states that we can obtain an accurate estimate for $\mathbf{x}$ by solving

$$\underset{\mathbf{x}}{\text{minimize}} \|\mathbf{x}\|_1 \quad \text{subject to} \quad \mathbf{A}\mathbf{x} = \mathbf{b} \tag{2.3}$$

when we sample at a rate proportional to the sparsity k instead of the ambient dimension $N \gg k$.

As opposed to stochastic optimization, where randomization is used to reduce variance (see e.g., Hutchinson, 1990; Avron and Toledo, 2011), compressive sensing uses randomization to turn coherent subsampling-related interferences—such as aliasing and shot "cross talk"—into relatively "harmless" Gaussian noise. According to compressive sensing, the noise level depends on the degree of subsampling and transform-domain sparsity. Consequently, sampling is no longer fully determined by Nyquist, but by transform-domain sparsity (see e.g., Herrmann, 2010, for an overview of the application of CS in exploration seismology). Neelamani et al. (2010) and Herrmann et al. (2009b) both took advantage of this finding in wavefield simulations with time stepping or by inverting the time-harmonic Helmholtz system. In both cases, fully-sampled wavefields are recovered with curvelet-domain sparsity promotion from small subsets of (monochromatic) simultaneous-source experiments. This procedure leads to efficient simulations because the computational overhead of the recovery is small compared to computational gain from subsampling.

However, solving Equation 2.1 differs fundamentally from standard compressive sensing because the system in Equation 2.1 is "overdetermined"; there are more equations then unknowns. In addition, the scattering matrix is ill conditioned due to limitations in aperture that may lead to shadow zones. However, for reflectors that are in the range of the scattering operator the wave-equation Hessian $\mathbf{A}^H\mathbf{A}$ is near unitary, and curvelets are nearly invariant under the action of the Hessian (Herrmann et al., 2008b). This property—in conjunction with the optimality of curvelets on images with reflectors that may include conflicting dips—motivates us to use curvelet-domain sparse recovery to mitigate the source crosstalk caused by the dimensionality reduction.

2.3 Methodology

To solve Equation 2.1 efficiently, we combine recent ideas from stochastic optimization and compressive sensing. For this purpose, let us first mathematically define seismic mini batches consisting of "supershots". Next, we present a pragmatic optimization strategy where we cast the original imaging problem into a series of much smaller subproblems that work on different subsets of random source-encoded supershots. For linearized inversion (Herrmann and Li, 2011b), this approach corresponds to drawing a collection of supershots, followed by imaging, and using this image as a warm start for a new linearized inversion with a new independently drawn collection of supershots. We compare the performance of least-squares on these subproblems with and without sparsity constraints. Depending on these two choices, the formulation leads either to a Monte-Carlo type of algorithm, which relies on averaging to reduce the subsampling errors, or to a compressive-sensing type of algorithm, which relies on sparsity promotion to remove the source cross talk. In both cases, the error depends on the subsampling ratio K/K'.

2.3.1 The seismic mini batch: a collection of supershots

We base our algorithm on forming compressive seismic experiments, or, to use the language of machine learning, mini batches that consist of collections of small numbers of supershots. These supershots are made of randomized superpositions of sequential sources.

Mathematically, imaging experiments for mini batches with $K' \ll K$ monochromatic supershots, require the solution of the reduced system

$$\mathbb{P}_{\ell_2}(\mathbf{RM}): \qquad \underset{\mathbf{x}}{\text{minimize}}\, \frac{1}{2}\|\mathbf{RM}\,(\mathbf{b}-\mathbf{Ax})\,\|_2^2 = \frac{1}{2}\|\underline{\mathbf{b}}-\underline{\mathbf{A}}\mathbf{x}\|_2^2, \tag{2.4}$$

In this expression, we dimensionality reduced Equation 2.1 with the subsampling matrix $\mathbf{RM}$. This subsampling matrix reduces the tall expensive to compute system $\mathbf{A} \in \mathbb{C}^{(KN_r)\times M}$ to $\underline{\mathbf{A}} \stackrel{\text{def}}{=} \mathbf{RMA} \in \mathbb{C}^{(K'N_r)\times M}$ and the data to $\underline{\mathbf{b}} \stackrel{\text{def}}{=} \mathbf{RMb}$. The dimensionality-reduction matrix itself is factored into a restriction and mixing matrix. The restriction matrix $\mathbf{R}$ is defined by the Kronecker product: $\mathbf{R} \stackrel{\text{def}}{=} \mathbf{R}^{\Sigma} \otimes \mathbf{I} \otimes \mathbf{R}^{\Omega} \in \mathbb{R}^{(K'N_r)\times(KN_r)}$ with $\mathbf{R}^{\Sigma} \in \mathbb{R}^{n_s'\times n_s}$ selecting $n_s' \ll n_s$ rows uniform randomly

amongst $[1 \cdots n_s]$ and $\mathbf{R}^{\Omega} \in \mathbb{R}^{n'_f \times n_f}$ selecting $n'_f \ll n_f$ frequencies from the seismic frequency band. The matrix $\mathbf{I}$ represents the identity matrix. The mixing matrix $\mathbf{M} \in \mathbb{R}^{(KN_r)\times(KN_r)}$ is given by the Kronecker product $\mathbf{M} \stackrel{\text{def}}{=} \mathbf{M}^{\Sigma} \otimes \mathbf{I} \otimes \mathbf{I}$. As in Lin and Herrmann (2007), we follow Romberg (2009) to phase encode sequential shots via

$$\mathbf{M}^{\Sigma} \stackrel{\text{def}}{=} \text{sign}(\eta) \odot \mathbf{F}_{\Sigma}^{H} \, \text{diag}\left(e^{j\theta}\right) \mathbf{F}_{\Sigma}, \tag{2.5}$$

with $\theta = \text{Uniform}([0, 2\pi])$ a random phase rotation, and $\mathbf{F}_{\Sigma}$ the Fourier transform along the source coordinate. The vector $\eta \in N(0,1)$ is used to define a random-sign pattern by which the phase-encoded vector is premultiplied (the symbol $\odot$ represents element-wise product). This definition for the matrix $\mathbf{M}^{\Sigma}$ is fast ($\mathcal{O}(n_s \log n_s)$) and mimics the action of a matrix with Gaussian *i.i.d.* entries.

Using linearity of randomized subsampling by **RM**, in combination with linearity with respect to monochromatic sources and the separability of the imaging problem (cf. Equation 2.1), it is easy to show that the number of PDE solves required for each iteration of the solution of Equation 2.4 is slimmed down by a factor of K'/K (see also Herrmann et al., 2009b, for details). Essentially, the action of **RM** "commutes". Note that recent work by Haber et al. (2010a) uses the same principle. Unfortunately, speed ups by randomized source supersposition and subsampling go typically at the expense of leaking energy from imaged reflectors to incoherent artifacts. Hence, the key question is to find a solver that mitigates these artifacts and restores the amplitudes at an overhead small compared to the speed up.

2.3.2 Stochastic-average approximations with warm starts

To address the slow decay of the error of stochastic-average approximation and the delicacy of the stochastic approximation, we combine ideas underlining these methods by casting the original imaging problem into a series of much smaller subproblems that work on different independent subsets of random source-encoded supershots (Herrmann and Li, 2011b). For linearized inversion, this approach corresponds to drawing a collection of supershots, followed by least-squares imaging, and using these images as warm starts for a new linearized inversion with a new independently drawn collection of supershots. This process is repeated until no

further progress is made towards the solution. In Algorithm 1, we outline this procedure for a generic subproblem solver $\mathbb{P}(\mathbf{RM};\mathbf{x}_0)$ that uses warm starts $\mathbf{x}_0$.

Two subproblem solvers

Because algorithm 1 gives us flexibility regarding the subproblem solver, we propose to compare two solvers, namely $\mathbb{P}_{\ell_2}(\mathbf{RM})$—solved by a limited number of iterations of LSQR (Paige and Saunders, 1982)—and

$$\mathbb{P}_{\ell_1}(\mathbf{RM},\tau): \qquad \underset{\mathbf{x}}{\text{minimize}}\, \frac{1}{2}\|\underline{\mathbf{b}}-\underline{\mathbf{A}}\mathbf{x}\|_2^2 \quad \text{subject to} \quad \|\mathbf{x}\|_1 \leq \tau. \tag{2.6}$$

We solve the latter problem, known as the Least Absolute Shrinkage and Selection Operator (LASSO, Tibshirani, 1997) problem, by a spectral-projected gradient method (see e.g., Berg and Friedlander, 2008, for details). Large-scale sparsity-promoting solvers for Equation 2.3 often involve the solution of LASSO subproblems.

With the LSQR solver, we regularize the inverse of the wave-equation Hessian by limiting the number of iterations of LSQR (Hansen, 1997). This is necessary because otherwise we may create imaging artifacts related to the null space of the (dimensionality-reduced) Hessian. The total number of PDE solves required by LSQR is proportional to $N_{\ell_2}K'$, with N_{ℓ_2} the number of iterations required by the ℓ_2-norm solver. Conversely, we control the null space with LASSO by sparsity promotion. To take full advantage of sparsity promotion as a regularization, we include the curvelet synthesis matrix (Candès et al., 2006) in the definition of the dimensionality-reduced Born scattering operator $\underline{\mathbf{A}}$. The migrated image is then calculated by applying curvelet synthesis on the $\mathbf{x}$ that solves Equation 2.6. The total number of PDE solves required by this algorithm is proportional to $N_{\ell_1}K'$ with N_{ℓ_1} the number of iterations required by the ℓ_1-norm solver. In these computations, the computational overhead of the curvelet transform and of the solver intelligence are negligible compared to the cost of solving PDE's and are therefore ignored.

Leveraging the Pareto curve

In the noise-free case, sparsity-promoting imaging involves the solution of Equation 2.3. Efficient ℓ_1 solvers for this problem are typically based on solutions of a

series of relaxed subproblems, where components are allowed to enter into the solutions controllably. It is widely known that these approaches lead to a reduction in the number of iterations to reach the solution. The spectral projected-gradient algorithm (SPGℓ_1, Berg and Friedlander, 2008) uses this principle by solving a series of LASSO problems where the τ's are increased intelligently. In this method, the Pareto boundary—the trade-off curve delineating feasible and infeasible solutions as a function the ℓ_2-norm of the data misfit and the model's ℓ_1-norm—is exploited to compute the relaxations by root finding that uses convexity and smoothness of the Pareto curve. See Figures 2.2a and 2.2b, which illustrate this principle, and the corresponding solution path. As we can see this approach uses a limited number of matrix-vector multiplies. Because the cost of the solver is determined by this number of multiplies, this approach is particularly suitable for large-scale geophysical problems (Hennenfent et al., 2008a).

Unfortunately, the degree of randomized dimensionality reduction determines the amount of cross-talk that results from the inversion, and hence we can not reduce the problem size too much. Therefore, we improve convergence by drawing new mini batches whenever a LASSO subproblem is solved. Because the solution is maximally sparse at that point, it is natural to select the new set of supershots and continue with a warm start of the algorithm for the next subproblem. We calculate the τ's with SPGℓ_1's root finding. This principle is illustrated in Figure 2.2b where the system of equations now changes after solving each subproblem. Of course, this approach is only justified as long as K' is not too small such that Pareto curves remain similar for different realization of the **RM**'s. To verify the assumption of similarity amongst different Pareto curves, we plotted four realizations of these curves for $K' = 12$ (four simultaneous shots and three frequencies) in Figure 2.2c. These curves clearly make the case that we should be able to continue using SPGℓ_1's root finding.

2.4 Empirical performance study

To compare the proposed algorithm, we conduct a series of synthetic imaging experiments using the Marmousi model (Bourgeois et al., 1991). We use the smoothed model, plotted in Figure 2.3a, as the background velocity model for the

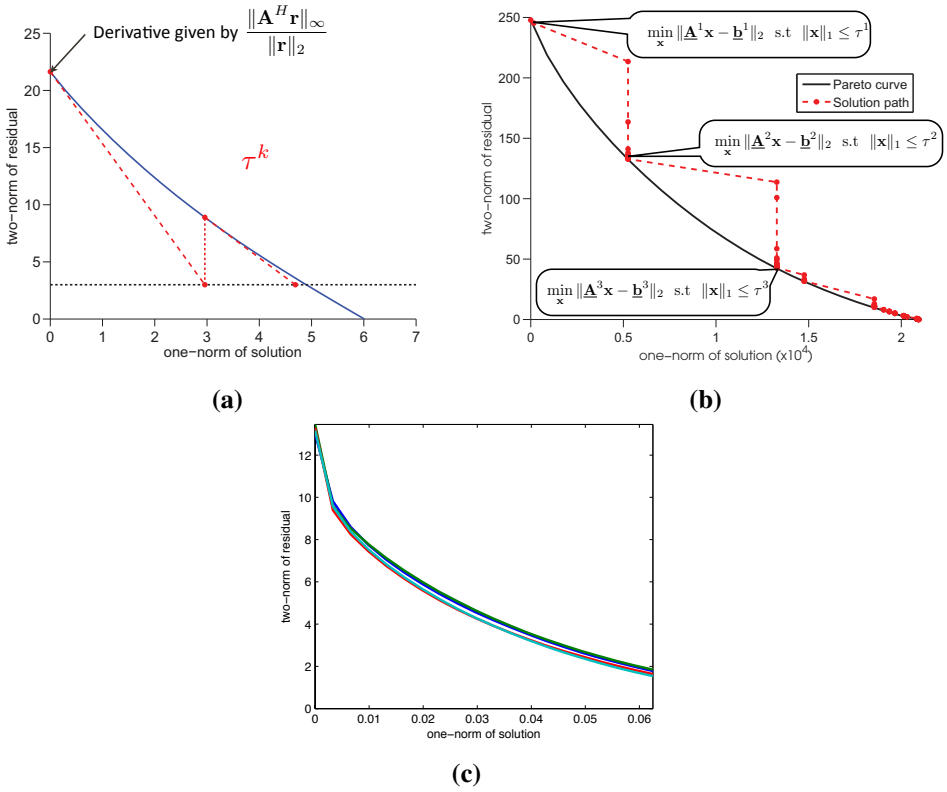


Figure 2.2: SPGℓ_1 and batching. **(a)** Newton root finding using the convexity and smoothness of the Pareto curve, which traces the two-norm of the residue $\mathbf{r}$ as a function of the one-norm of the solution τ. (adapted from Berg and Friedlander (2008)). **(b)** Series of LASSO subproblems with renewals for the collections of supershots (adapted from Berg and Friedlander (2008)). **(c)** Pareto curves for different independent realizations of the dimensionality-reduction ($K' = 12$) operator **RM**.

migration operator. This model yields the true medium perturbation included in Figure 2.3b. With this smoothed model, we generate time-harmonic data by calculating the difference between the solution of the Helmholtz equation at the receivers for the true and smoothed velocity models. This choice of using the residue, which mimics real data after processing, is realistic because it does not rely on the linearized Born approximation. Of course, when the smoothed model is close to the actual model, these two definitions become similar. We use a nine-point stencil (Jo

et al., 1996a) and absorbing boundary conditions on a 143×384 grid with a grid size of 24 m. To mimic real applications, we solve the Helmholtz systems on the fly during the inversion. The length of the time record is 2.4 s and we use 192 shot locations, with a shot spacing of 48 m, and 384 receiver positions sampled with a receiver spacing of 24 m.

To establish baselines for comparison, we compute a migrated image from only eight simultaneous shots and three random frequencies selected from the amplitude spectrum of a 12 Hz Ricker wavelet (Figure 2.4a), a migrated imaged from all 192 sequential shots and 10 randomly selected frequencies (Figure 2.4b), and a least-squares image with the same data but obtained by solving Equation 2.1 with 10 iterations of LSQR (Figure 2.4c). For these reference images, the number of PDE solves increases from $2 \times 8 \times 3 = 48$ to $2 \times 192 \times 10 = 3840$ and $4 \times 192 \times 10 \times 10 = 76800$. While there is a clear improvement in image quality, the computational costs increase sixteen-hundred fold, which illustrates the need of dimensionality reduction.

To improve the image quality of the dimensionality-reduced image (Figure 2.4a), we solve a series of dimensionality reduced subproblems with eight supershots and three frequencies ($K' = 24$) for 10 subproblems with LSQR ($\mathbb{P}_{\ell_2}(\mathbf{RM})$) and SPG$\ell_1$ ($\mathbb{P}_{\ell_1}(\mathbf{RM})$) each with and without independent renewals of $\mathbf{RM}$. The results of these experiments are summarized in Figures 2.5 and 2.6. From these experiments, we can make the following observations. First, redrawing the supershots after solving each subproblem improves the performance of both solvers. This can be understood because these renewals remove possible correlations between $\mathbf{RM}$ and the current estimate for the (curvelet-domain) velocity perturbation. Second, the images obtained by sparsity promotion (Figure 2.6) are clearly superior in quality compared to the least-squares results (Figure 2.5) even though the number of PDE solves is roughly the same. This can be explained by compressive sensing. Third, the sparsifying result with renewals, albeit noisy, compares favorably to the baseline image in certain areas; e.g. it has higher resolution and better resolved amplitudes at depth. We attribute this observation to the regularization by curvelet-sparsity promotion, which we were able to carry out by virtue of the dimensionality reduction. We argue that this improvement offsets the price we pay of the remnant random interferences. For example, these interferences average out if we use this

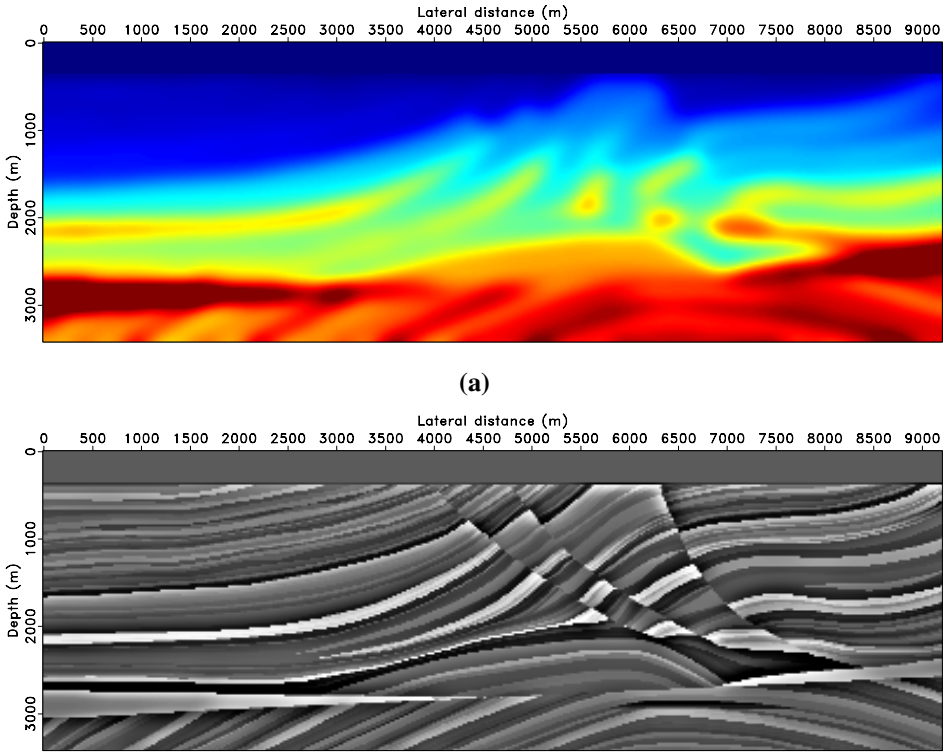
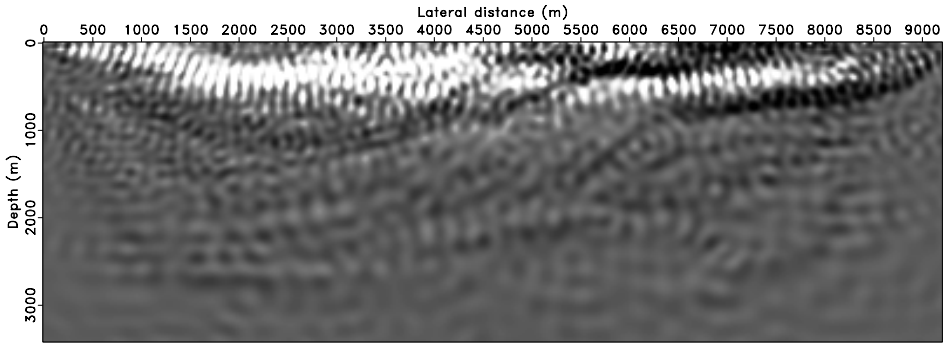


Figure 2.3: The Marmousi model. **(a)** Smoothed background-velocity model. **(b)** Velocity perturbation, defined by the difference between the true and smoothed background-velocity models.

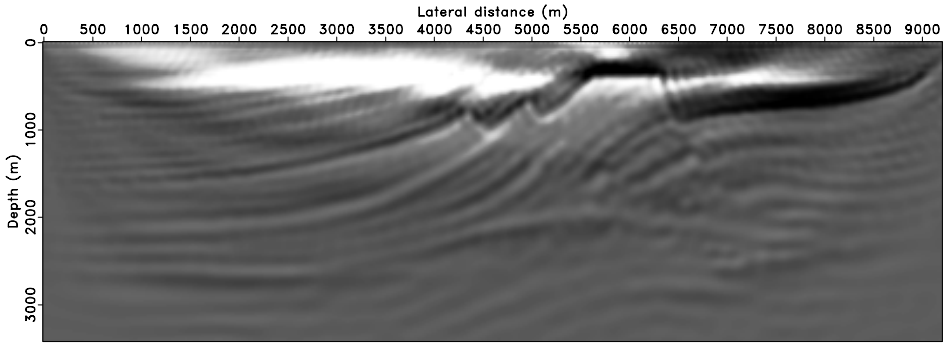
procedure as part of full-waveform inversion on which we reported elsewhere (Li and Herrmann, 2010a; Herrmann et al., 2011a). In summary, we obtained a remarkably good result with a significantly reduced computational cost. We attribute this performance to curvelet-domain compressibility, which serves as a strong prior that mitigates source crosstalk and regularizes the inversion.

2.4.1 Convergence as a function of the number of PDE solves

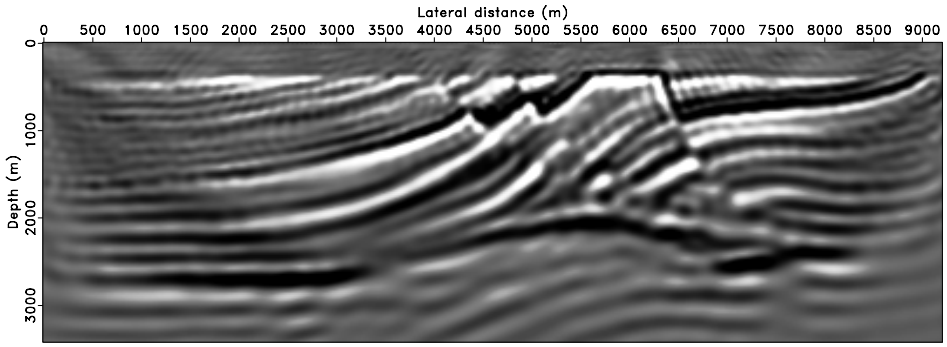
The possible gains in computation speed of our solvers hinges on the interplay between the mini batch size and the number of matrix-vector multiplies required by the solver to bring down both the data residue and to recover the artifact-free



(a)



(b)



(c)

Figure 2.4: Baseline images. **(a)** migrated image with eight simultaneous shots and three randomly selected frequencies, **(b)** migrated image calculated with all data (192 sequential sources and 10 frequencies), and **(c)** least-squares migrated image from the same data with 10 iterations of LSQR.

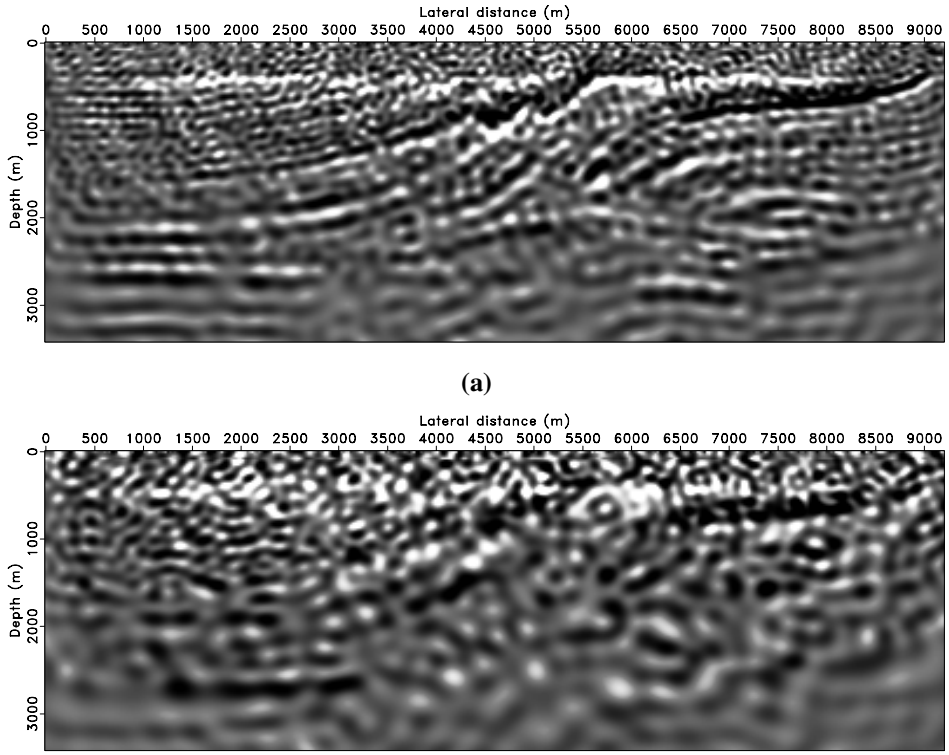


Figure 2.5: Stochastic-average approximation with LSQR. **(a)** Image obtained by Algorithm 1 with $\mathbb{P}_{\ell_2}(\mathbf{RM})$ with 10 independent redraws for **RM**. **(b)** The same but without redrawing **RM**.

image. The product of these two factors determines the number of PDE solves. To measure the performance of the proposed curvelet-based stochastic-average approximation with warm starts, we plot this number versus the model-error energy for a fixed ratio of $K/K' = 80$ and $n'_s/n'_f = 8/3$. The results of this exercise with and without redraws are plotted in Figure 2.7. This plot clearly demonstrates a more rapid decay for the model-space error (difference between true and estimated model) in case of independent redraws of the **RM**'s. Compared to the least-squares baseline problem with "all" data, we obtain approximately a fourfold speedup if we factor in the number of iterations required by the solver. On first glance, this speedup may not be significant. However, we expect a larger uplift in 3D where

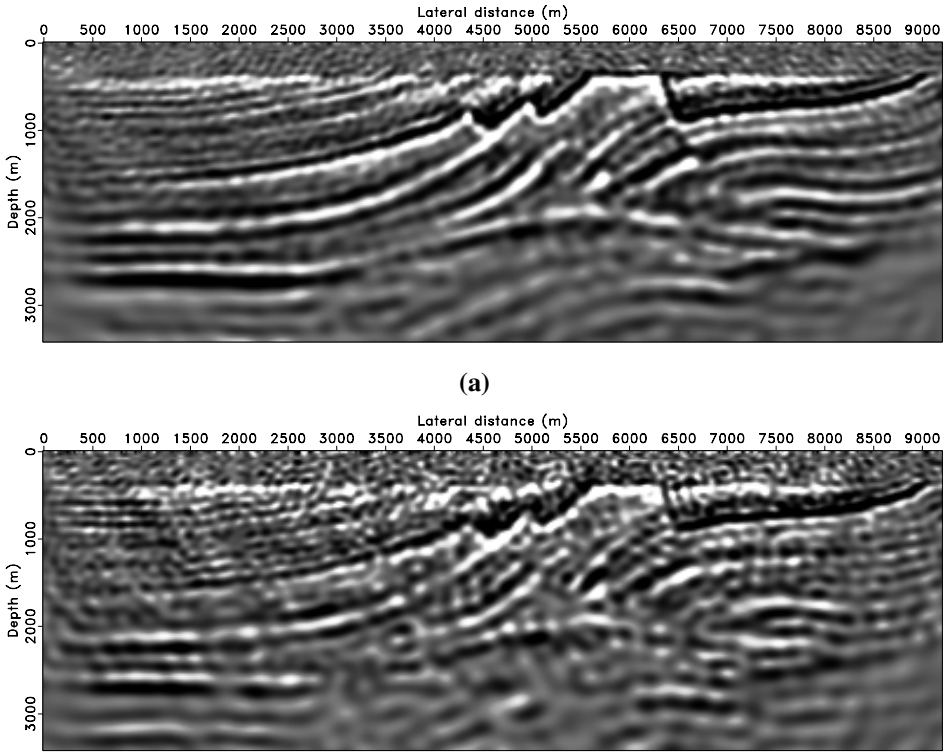


Figure 2.6: Stochastic-average approximation with SPGℓ_1. **(a)** Image obtained by Algorithm 1 with $\mathbb{P}_{\ell_1}(\mathbf{RM})$ for approximately the same number of PDE solves.**(b)** The same but without redrawing **RM**.

there are many more sources. In addition, with our dimensionality reduction we are able to approximately solve the BP problem (cf. Equation 2.3). Without the dimensionality reduction this would not have been possible because of the number of iterations required by one-norm solvers. Finally, since we are working on very small subproblems we may have the option to keep the wavefields in memory instead of computing them on the fly. We expect this to lead to a significant additional speedup.

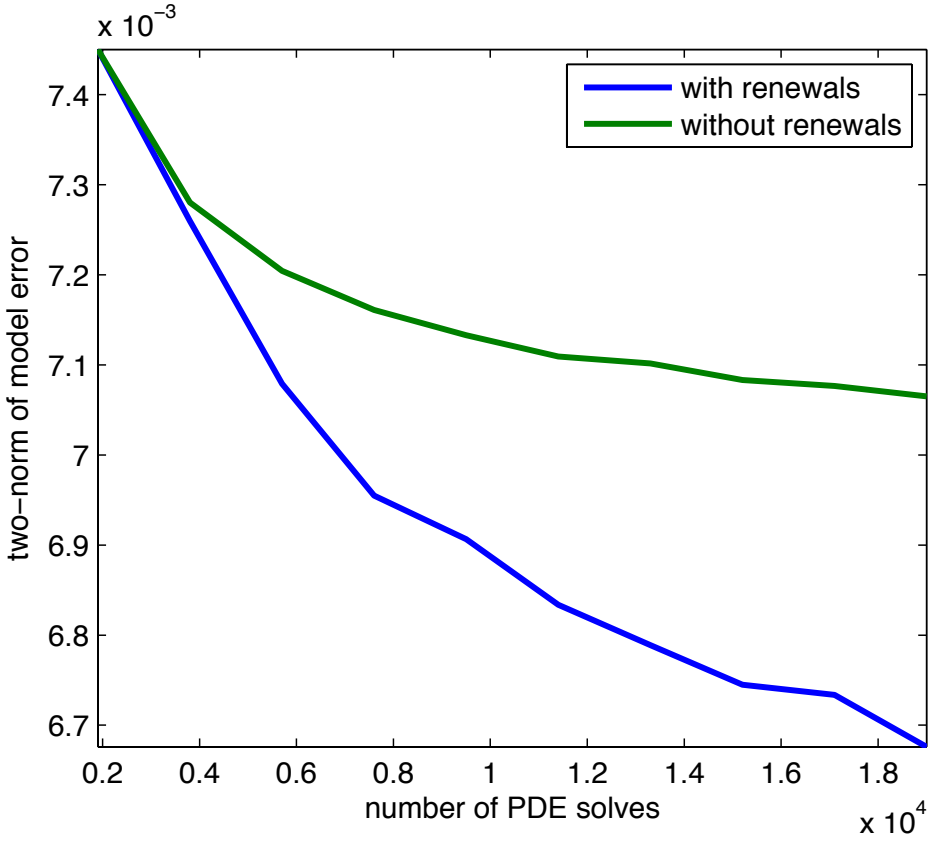


Figure 2.7: Two-norm error between the true and recovered medium perturbation as a function of the number of PDE solves (for $\mathbb{P}_{\ell_1}(\mathbf{RM})$). Convergence is clearly improved by drawing new randomized collections of supershots after each subproblem is solved.

2.4.2 Recovery quality as a function of batchsize

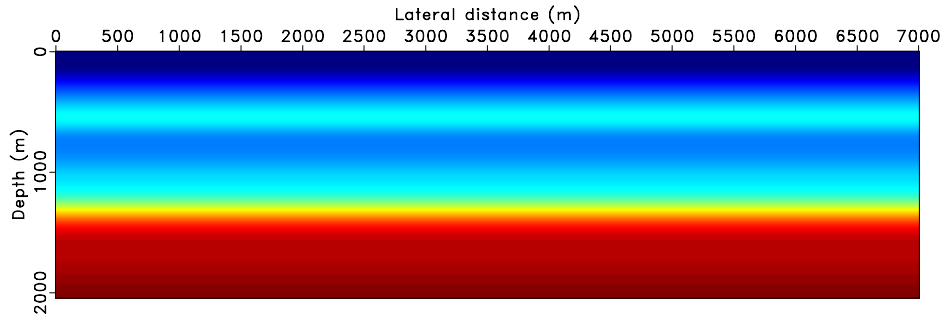
As we have seen from the previous example, computational gains can be made using the proposed stochastic-average approximation with relaxed LASSO's. To get a better understanding of the relationship between the recovery error and minibatch size, we also conduct a series of experiments where we vary the subsampling ratio's and the ratio supershots-over-frequencies while approximately fixing the number of PDE solves. The results of this exercise are summarized in Table 2.1. As expected, the numbers in Table 2.1 generally confirm increasing recovery errors for

decreasing subsampling ratios albeit there is not a very strong relationship between the recovery quality and the subsampling ratio compared to results presented in the literature (Li and Herrmann, 2010a). The fact that our results were generated with renewals offers an explanation for the weaker dependence on the subsampling ratios. Finally, we like to add that the speed improvements due to the dimensionality reduction are partly offset by the iterations required by the solver. However, without the dimensionality reduction the computational cost for the sparsity-promoting solution would have been prohibitive.

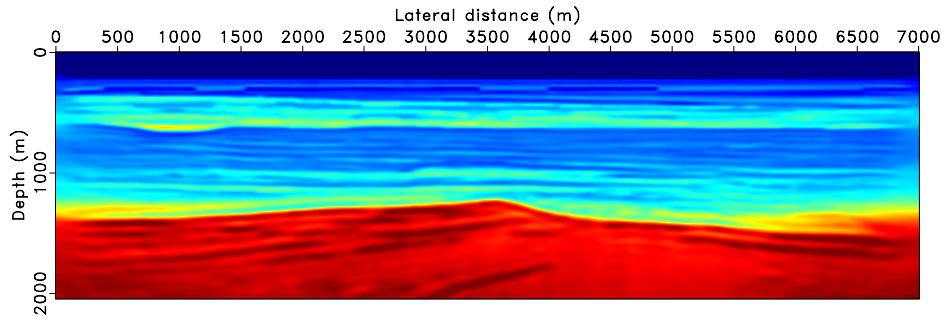
2.5 Case study: the BG Compass model

To test our imaging algorithm in a more realistic setting, we consider a synthetic velocity model with a large degree of well-constrained variability. To build the background-velocity model, we employ our modified Gauss-Newton method with sparse updates in 10 overlapping frequency bands on the interval $2.9 - 22.5$Hz and with initial model plotted in Figure 2.8a. (Note that this approach reported in Herrmann et al. (2011a) is based on a similar dimensionality reduction technique as presented in this chapter.) The output of this procedure is plotted in Figure 2.8b and is used as the background-velocity model for our imaging algorithm.

We parametrize the velocity perturbation on a 409×1401 grid with a gridsize of 5 m. Again, we use the Helmholtz solver to generate data from 350 source positions sampled at an interval of 20 m and 701 receivers sampled with an interval of 10 m. We use 10 random frequencies in our simulations selected from the interval $20 - 50$ Hz and scaled by the spectrum of a 30 Hz Ricker wavelet. The input data is given by the difference between simulations with the true and initial velocity models (Figure 2.8b). As before, we solve 10 subproblems $\mathbb{P}_{\ell_1}(\mathbf{RM})$ with and without independent redraws of $\mathbf{RM}$. The result of this exercise is summarized in Figure 2.9 and clearly show significant improvements from the redraws. Not only is the crosstalk removed more efficiently but the reflectors are also better imaged in particular at the deeper parts of the model where recovery without redraws is not able to image the events.



(a)



(b)

Figure 2.8: Full-waveform inversion result. **(a)** Initial model. **(b)** Inverted result starting from 2.9Hz with 7 simultaneous shots and 10 frequencies in each of the 10 frequency bands.

2.6 Discussion

Efforts to speed up the computation of linearized imaging roughly fall into two categories. First, there are methods that aim to "nearly diagonalize" Green's functions (Douma and de Hoop, 2007) or wave-equation Hessians (Herrmann et al., 2008b) using transform-domain techniques such as curvelets. These methods exploit the property that curvelets remain near invariant under wave propagation, which in principle, leads to fast algorithms. Unfortunately, the engineering of concrete and explicit implementations of these approximations is involved, and may carry a significant overhead (Andersson et al., 2008). Conversely, randomized dimensionality reduction is simpler because it utilizes the notion of curvelet-invariance implic-

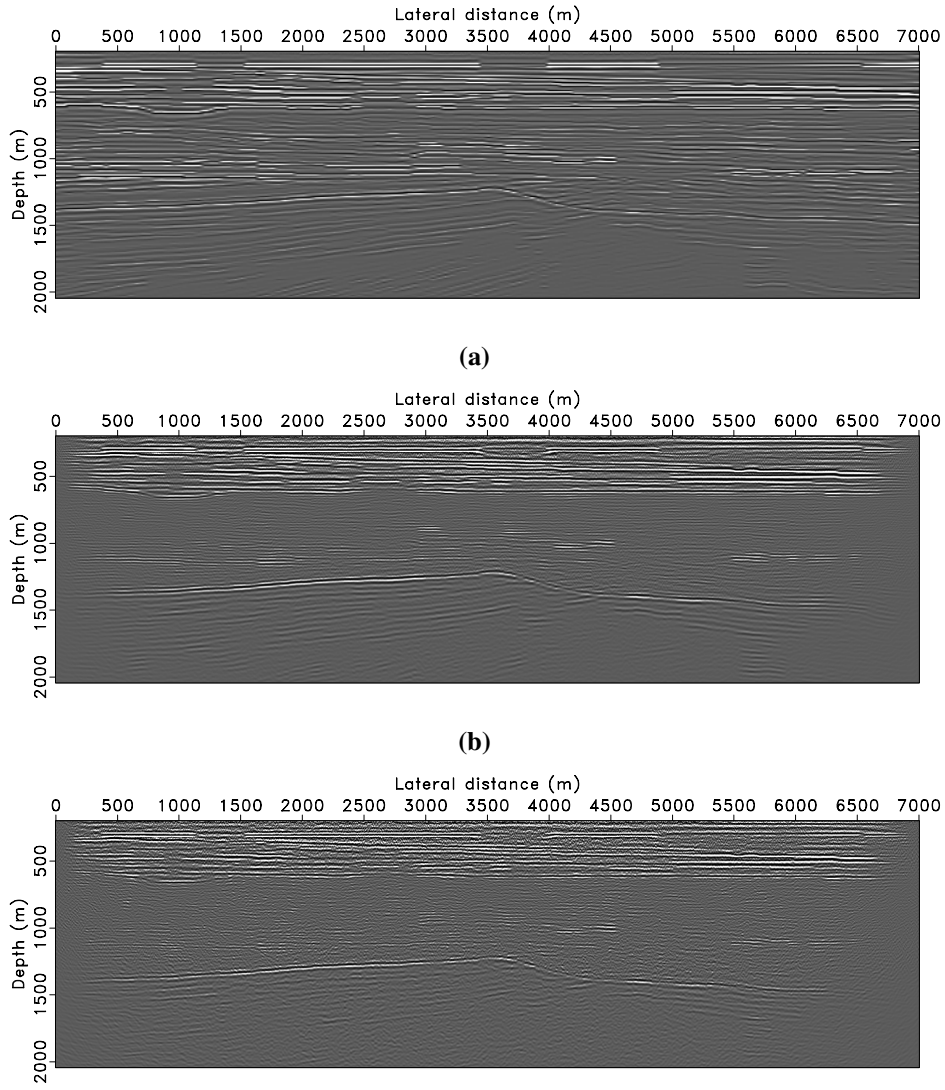


Figure 2.9: Dimensionality-reduced sparsity-promoting imaging from random subsets of 17 simultaneous shots and 10 frequencies. We used the background velocity-model plotted in Figure 2.8b **(a)** True perturbation given by the difference between the true velocity model and the FWI result plotted in Figure 2.8b. **(b)** Imaging result with redraws for the supershots. **(c)** The same but without redrawing **RM**. Notice the significant improvement in image quality when renewing collection of supershots after solving each LASSO subproblem.

itly via transform-domain sparsity. (Propagated wavefields remain compressible in curvelet frames. See Smith, 1998; Demanet and Peyré, 2011, who rigorously prove this property.) This feature explains the success of our proposed method, which benefits from the availability of wave simulators and the ability of curvelets to sparsely represent seismic images. By promoting sparsity, we are able to exploit continuity along the reflectors without requiring a data-adaptive step that requires prior information on the dip field of the reflectors (Guitton et al., 2010).

By considering dimensionality-reduced subproblems as compressive sensing-like sparse recovery problems—where the originally "overdetermined" system is turned deliberately into an underdetermined system—we remove the crosstalk artifacts and restore the amplitudes by iterating on highly dimensionality reduced subproblems. Not withstanding an ill-conditioned Hessian, spectral-projected gradients makes good progress towards the solution in relatively few matrix-vector multiplies. A possible explanation for this phenomenon and the benefit from redraws is that compressive sampling the sources does not make the dimensionality-reduced Hessian significantly more ill conditioned. This is after all the premise of compressive sensing where the condition number of sampling matrices is being controlled by design. In addition, recent work by Montanari (2010) has shown that redraws help to remove possible correlations between the solution vector and the source encoding and this increases the convergence of solutions based on iterative soft thresholding. We can argue that we are observing this effect empirically. Finally, our approach is also reminiscent of randomized "block Kaczmarz" (Strohmer and Vershynin, 2009) and to recent work by Friedlander and Schmidt (2011); van Leeuwen et al. (2011b) who also propose sampling strategies.

Because our formulation includes contributions from the wave-equation Hessian more theoretical work will be necessary to (i) compensate for the "coloring" by this operator, e.g. by solving weighted ℓ_1-norm problems and (ii) analyze the coherence and restricted-isometry properties of the dimensionality reduced Hessian, using practical techniques recently developed by Mansour et al. (2011). These latter results are particularly exciting because they allow for

- an efficient imaging technology with a controllable error. As in compressive sensing, this error depends on the subsampling ratio and on the compress-

ibility of the model in the sparsifying domain. This means that our sparse recovery algorithm returns images that can be considered as images we would have obtained by keeping only a small fraction of the largest transform-domain (curvelet) coefficients. The larger the batch size, the larger this fraction, and the better the recovery by virtue of the transform-domain compressibility. Clearly, this property differs fundamentally from Monte-Carlo techniques where the error decays slowly with the batch size.

- an integration of dimensionality reduction with acquisition. For instance, we could envisage "online" acquisition during which simultaneously acquired data is continuously inverted with a procedure reminiscent of the approach outlined in this chapter.

Finally, we would like to mention that our method relies on fixed-spread data, which makes the application of this dimensionality-reduction methodology challenging for marine data. We are working on solutions to this challenge (see e.g., van Leeuwen et al., 2011b) on which we plan to report elsewhere.

2.7 Conclusions

We introduced an efficient algorithm to solve the linearized imaging problem. Our method combines recent findings from the fields of stochastic optimization and compressive sensing and turns the originally "overdetermined" seismic imaging problem into a series of underdetermined dimensionality-reduced subproblems. By considering these subproblems as sparse-recovery problems, we were able to create high-fidelity images at a fraction of the computational cost. Our final image can be considered as a result of curvelet-domain sparsity-promoting migration because large-scale one-norm optimization relies on solving a very similar, series of subproblems.

Randomized dimensionality reduction, curvelet-domain sparsity promotion, and stochastic optimization all played essential roles in making the computations tractable, controlling the nullspace of the Hessian, removing the crosstalk, and breaking possible correlations that may develop between the current model and the randomization. The observed differences in image quality between the uncon-

strained and sparsity-constrained formulations are consistent with the predictions of compressive sensing.

In summary, our approach can be seen as an instance of a new randomized dimensionality-reduction paradigm where the costs of computations are no longer dominated by the discretization but by transform domain sparsity of the model. In this new paradigm of randomized inversion, dimensionality reductions allow us to solve (linearized) inversion problems in ways, which previously, would have been computationally infeasible. The examples presented in this chapter support this observation and show highly competitive results on synthetic model with realistic complexity.

Result: Estimate for the model $\widetilde{\mathbf{x}}$
$\mathbf{x}_0 \longleftarrow \mathbf{0}$; // initial model
$\mathbf{k} \longleftarrow \mathbf{0}$; // initial counter
while $\|\mathbf{x}_0 - \widetilde{\mathbf{x}}\|_2 \geq \varepsilon$ **do**
    $k \longleftarrow k+1$; // increase counter
    $\widetilde{\mathbf{x}} \longleftarrow \mathbf{x}_0$; // update warm start
    $\mathbf{RM} \longleftarrow \text{Draw}(\mathbf{RM})$; // draw new subsampler
    $\mathbf{x}_0 \longleftarrow \text{Solve}(\mathbb{P}(\mathbf{RM}); \widetilde{\mathbf{x}})$; // solve the subproblem
end

Algorithm 1: Stochastic-average approximation with warm restarts

Subsample ratio	0.0006	0.0013	0.0026	0.0033
n'_f/n'_s	Signal-noise ratio (dB)			
2	**1.60**	**1.63**	**1.86**	**1.95**
1	**1.62**	**1.75**	**1.87**	**1.99**
0.5	**1.63**	**1.77**	**1.98**	**2.06**
Speed up ($\times$)	1536	768	384	307

Table 2.1: Signal-to-noise ratios, SNR $= -20\log_{10}(\frac{\|\mathbf{x}-\mathbf{x}_0\|_2}{\|\mathbf{x}\|_2})$ for sparse curvelet-based recovery for different subsample and frequency-to-shot ratios. The vector $\mathbf{x}$ is the inverted perturbation and $\mathbf{x}_0$ is the true perturbation given by the difference between the true and smooth background model.

Chapter 3

Fast randomized full-waveform inversion with compressive sensing

3.1 Motivation

Full-waveform inversion (FWI) can be formulated as the separable parameter-estimation problem

$$\underset{\mathbf{m}}{\text{minimize}}\ \Phi(\mathbf{m}) := \left\{ \frac{1}{2K} \sum_{i=1}^{K} \|\mathbf{d}_i - \mathscr{F}_i[\mathbf{m}, \mathbf{q}_i]\|_2^2 = \frac{1}{2} \|\mathbf{D} - \mathscr{F}[\mathbf{m}]\mathbf{Q}\|_F^2 \right\}, \quad (3.1)$$

with $\mathbf{d}_i$ monochromatic shot records of the Earth response to monochromatic sources $\mathbf{q}_i$, $\mathscr{F}_i[\mathbf{m}, \mathbf{q}_i]$, $i = 1 \cdots K$ monochromatic nonlinear forward operators, and $K = N_f \cdot N_s$, with N_f, N_s the number of frequencies and sources. In the acoustic constant-density case, this operator is parameterized by the unknown velocity model $\mathbf{m}$ and involves the inversion of a large system of linear equations.

Solving for the velocity model is challenging for several reasons. First, the

A version of this chapter has been published. Xiang Li and Aleksandr Y. Aravkin and Tristan van Leeuwen and Felix J. Herrmann. Fast randomized full-waveform inversion with compressive sensing. *Geophysics*, 2012.

solution is non-unique due to "cycle skipping". This phenomenon gives rise to local minima in the objective function. Second, The data are incomplete because low frequencies and certain offsets are missing. Third, iterative methods for Equation 3.1 are prohibitively expensive, since they require too many PDE solves in 3D.

To address these issues, we use the following properties of FWI (cf. Equation 3.1):

- linearity with respect to the sources that gives Equation 3.1 its separable structure.
- transform-domain sparsity on the updates. This allows us to "fill in" the null space of the Hessian and to remove the source crosstalk and restore the amplitudes;
- convex composite structure of $\Phi(\mathbf{m})$—it is a composition of the convex ℓ_2-norm with the smooth $\mathscr{F}$, and thus admits the standard Gauss-Newton (GN) algorithm as well as sparsity-promoting variants.

Our main contribution is to combine these properties with existing multiscale continuation methods (Bunks et al., 1995) that are widely employed to solve Equation 3.1. The outcome is a formulation that allows us to reduce the number of PDE solves, to mitigate the effects of source crosstalk, to "fill in" the nullspace of the wave equation Hessian, and to speed up progress of the algorithm by using ideas from stochastic optimization (Haber et al., 2010a; van Leeuwen et al., 2011a).

Our paper is organized as follows. First, we discuss dimensionality reduction techniques that allow algorithms to work with only a portion of the data at each iteration. We then exploit the convex-composite structure of Equation 3.1 to design a GN subproblem that is conducive to randomized-dimensionality reduction. Next, we modify this reduced subproblem using ideas from Compressive Sensing (CS)(Candès et al., 2006; Donoho, 2006; Mallat, 2009), i.e. we remove source crosstalk by curvelet-domain sparsity promotion. Finally, we speed up the progress of the algorithm by drawing new encodings after solving each modified GN subproblem. We evaluate the performance of this algorithm on a realistic 2-D synthetic.

3.2 Methodology

An important class of algorithms used to solve the FWI problem are the GN methods that involve the pseudo-inverse of the reduced Hessian given by the combined action of the Jacobian operator $\nabla\mathscr{F}[\mathbf{m}^k;\mathbf{Q}]$ and its adjoint. Even though this method does not require explicit computation of the Hessian, each iteration for the GN subproblem requires the solution of $4K$ PDE's.

3.2.1 Dimensionality reduction by randomized source superposition

To reduce the number of required PDE solves, we combine the sources and data into a smaller volume with $K' \ll K$ simultaneous experiments by replacing Equation 3.1 with

$$\underset{\mathbf{m}}{\text{minimize}}\ \underline{\Phi}(\mathbf{m}) := \left\{ \frac{1}{2K'} \sum_{i=1}^{K'} \|\mathbf{D}\mathbf{w}_i - \mathscr{F}_i[\mathbf{m}, \mathbf{q}_i\mathbf{w}_i]\|_2^2 = \frac{1}{2}\|\underline{D} - \mathscr{F}[\mathbf{m}, \underline{Q}]\|_F^2 \right\}, \tag{3.2}$$

where $\{\underline{D}, \underline{Q}\} \overset{\text{def}}{=} \{\mathbf{DW}, \mathbf{QW}\}$ (Moghaddam and Herrmann, 2010a; Haber et al., 2010a; van Leeuwen et al., 2011a). If we choose the random weights $\mathbf{w} = [w_1, \cdots, w_{N_s}]^T$ in $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_{K'}]$ such that the expectation $\mathrm{E}\{\mathbf{w}\mathbf{w}^T\}$ equals the identity matrix, we have $\mathrm{E}\{\underline{\Phi}(\mathbf{m})\} = \Phi(\mathbf{m})$. Because Equation 3.2 can be interpreted as a sample-average, it represents an approximation of this expectation with an error that depends on K'. This error, which results in energy leaking towards source crosstalk, decreases for larger K' and this property underlies the method of the stochastic-average approximation (SAA, Bertsekas and Tsitsiklis, 1996; Nemirovski et al., 2009a).

Unfortunately the randomization in Equation 3.2 defeats the purpose of making FWI faster because the error of SAA is known to decay slowly as a function of increasing K'. To overcome this issue, Krebs et al. (2009a) proposed an approach reminiscent of the stochastic approximation where different weights are drawn for each gradient update. This method of stochastic-gradient descent is relatively well understood, and available convergence theories rely on specialized step lengths and averaging over previous model iterates (Bertsekas and Tsitsiklis, 1996; Haber et al., 2010a; van Leeuwen et al., 2011a). Unfortunately, this averaging over previous iterates significantly slows down the progress of the algorithm. Therefore,

we propose a solution method that relies on transform-domain sparsity promotion instead of averaging to reduce the error induced by the source crosstalk.

3.2.2 Exploiting the convex-composite structure

The standard GN method exploits the convex-composite structure of Equation 3.2 by linearizing the function inside the convex ℓ_2-norm. We modify the standard GN subproblem by adding transform-domain sparsity promotion, and instead recover the updates by solving the following constrained (convex) optimization problem:

$$\underset{\mathbf{x}}{\text{minimize}}\, \frac{1}{2}\|\underline{\delta D} - \nabla\mathscr{F}[\mathbf{m};\underline{Q}]\mathbf{S}^H\mathbf{x}\|_F^2 \quad \text{subject to} \quad \|\mathbf{x}\|_1 \leq \tau, \tag{3.3}$$

where $\underline{\delta D} = \underline{D} - \mathscr{F}[\mathbf{m}]\underline{Q}$, $\mathbf{S}^H$ is the inverse of the sparsifying transform determining the model update $\delta\mathbf{m} = \mathbf{S}^H\mathbf{x}$, and $\mathbf{x}$ is a vector of transform coefficients. The constraint enforces the ℓ_1-norm of $\mathbf{x}$ to be smaller than or equal to constant τ.

3.2.3 Dimensionality-reduction and compressive sensing

As long as signals exhibit structure—e.g., transform-domain compressibility where the signal's energy is concentrated in a few large coefficients—CS theory tells us that we can recover these signals from severe undersampling by solving a sparsity-promoting program.

Unfortunately, this recovery does not hold for arbitrary subsamplings. Instead, CS prescribes sampling matrices that roughly behave like matrices with Gaussian entries. In that case, subsampling artifacts are shaped into white Gaussian "noise" and the sparsity-promoting recovery separates the signal from these "noisy" interferences.

The dimensionality reduction outlined in the previous section follows these guidelines because we typically draw the random weights from a Gaussian distribution. Because our definition of the "CS sampling matrix" includes the wave-equation Jacobian there is an important difference between CS and our setting where we need to invert a tall system of equations to obtain the GN updates. So, for CS-type of arguments to hold, we need $\nabla\mathscr{F}^H[\mathbf{Q}]\nabla\mathscr{F}[\mathbf{Q}]$ to be near unitary with incoherent off-diagonals. Since this property is already true for $\mathbf{W}\mathbf{W}^H$, it would

then hold for $\nabla\mathscr{F}^H[\mathbf{QW}]\nabla\mathscr{F}[\mathbf{QW}]$.

For "high" frequencies, this argument holds because the wave-equation Hessian is "near unitary" as long as we are in the range of the Jacobian (Herrmann et al., 2008b). This means that the dimensionality reduction keeps the problem in the regime where the insights from CS are applicable (Herrmann and Li, 2011b,a). Consequently, it is useful to modify dimensionality-reduced GN subproblems to incorporate sparsity-promotion.

3.3 Modified Gauss-Newton

To arrive at a practical and fast GN formulation, we need to address the following issues: (i) convergence guarantees, (ii) selection of the proper sparsifying transform, (iii) selection of the one-norm solver and sparsity levels, and (iv) data-volume reduction to decrease the number of PDE solves. We discuss each of these issues by considering the pseudo code of Algorithm 2 in detail.

Proof of convergence [lines 2–9, Algorithm 2]: The convergence algorithm for the standard GN method is well known (Burke, 1990), and can be shown to apply to our modified algorithm as long as the values of the series of increasing sparsity levels τ_k remain bounded and the weights $\mathbf{W}$ remain the same for each iteration—i.e., $\mathbf{W}^k = \mathbf{W}$ for all k (Herrmann et al., 2011a).

Sparsifying transform [lines 6–7, Algorithm 2]: Selection of the appropriate sparsifying transform for the updates is important for two reasons. First, transform-domain sparsity leads to a concentration of the update's energy into a few large transform-domain coefficients. Second, the wave-equation reduced Hessian—also known as the demigration-migration operator—has a null space and requires regularization to stabilize its inversion. Here, transform-domain sparsity-promotion serves as a prior that fills in the nullspace (see e.g., Daubechies et al. (2005) or chapter 11, Mallat (2009)).

To guarantee a fast decay for the magnitude-sorted transform coefficients on the updates, we require the transform to detect "wavefronts", possibly with conflicting dips, and to be nearly "invariant" under wave propagation. This allows

us to sparsely represent both the medium perturbations and the updates even in situations where the current model iterate is far from the true model. Finally, to avoid non-physical artifacts at the boundaries of the model, we use a mirror extended discrete curvelet transform (Candès et al., 2006; Demanet and Ying, 2007), which decomposes the updates with respect to a collection of multiscale and multidirectional "localized plane waves". We denote the 2D mirror extended curvelet transform by the symbol $\mathbf{C}_2$.

One-norm solver and relaxation [lines 5-6, Algorithm 2] An essential component of our algorithm requires solution of sparsity-inducing GN subproblems (cf. Equation 3.3). We solve these subproblems with a spectral-gradient method (SPG, see e.g. Berg and Friedlander, 2008). We choose this first-order method because of its algorithmic simplicity and the fact that the Hessian is near unitary.

Because solutions of the GN subproblems depend on the sparsity level τ—i.e., small τ's lead to sparse solutions with a large residue—we need to carefully select the τ^k for each GN subproblem. For this purpose, we follow Berg and Friedlander (2008); Hennenfent et al. (2008a) and introduce the function $v(\tau)$, which is the ℓ_2-norm of the residue in Equation 3.3 as a function of the sparsity level τ. For each GN subproblem, this curve is decreasing, convex, and smooth (Berg and Friedlander, 2008), which allows us to obtain reasonable values of τ^k for each subproblem. The energy of the residual depends on the the current iterate $\mathbf{m}^k$ and on the solution of the GN subproblem, which is unknown (Doel and Ascher, 2011). However, because $\delta\mathbf{m}^k$ is a descent direction this residual will go down with each GN solve. It is therefore reasonable to assume that the value function decreases by some fraction $0 < \alpha < 1$ such that we have $v_k(\tau_k) = \alpha v_k(0)$ for the k^{th} GN subproblem. We can calculate the τ^k using the linearization $\tau_k \approx (\alpha - 1)v_k(0)/v_k'(0)$ with v' given by (Berg and Friedlander, 2008, Theorem 2.1): $v_k'(0) = -\|\mathbf{C}_2 \nabla \mathscr{F}^H[\mathbf{m}^k; \underline{Q}]\underline{\delta D}^k\|_\infty$, which yields a closed-form solution $\tau^k = \frac{(1-\alpha)\|\underline{\delta D}^k\|_2}{\|\mathbf{C}_2 \nabla \mathscr{F}^H[\mathbf{m}^k; \underline{Q}]\underline{\delta D}^k\|_\infty}$.

Algorithm speed-up using stochastic optimization [lines 3 and 7, Algorithm 2] While the proposed dimensionality reduction technique allows us to solve the GN subproblems with a reasonable computational effort, the overall costs of the inver-

sion remains prohibitively expensive. Therefore, we rely on insights from Stochastic Optimization (Haber et al., 2010a; van Leeuwen et al., 2011a) by drawing independent $\mathbf{W}$'s for each GN subproblem. This extends our work on the linear case of least-squares migration where these renewals led to a significant improvement in the convergence (Herrmann and Li, 2011b,a). In this case it is not clear how to select the steplength γ^k when sampling $\mathbf{W}^k$ at every iteration. Motivated by stochastic-gradient descent (Bertsekas and Tsitsiklis, 1996), we use a fixed sequence of steplengths ($\gamma_k = 1$), which works well in practice.

3.4 The BG compass model

To test our inversion algorithm in a realistic setting, we consider a synthetic velocity model with a large degree of variability constrained by well data. We use this model to generate data with a 12 Hz Ricker wavelet. We use a smooth starting model without lateral information (Fig. 3.1b) and we start the inversion at 2.9 Hz.

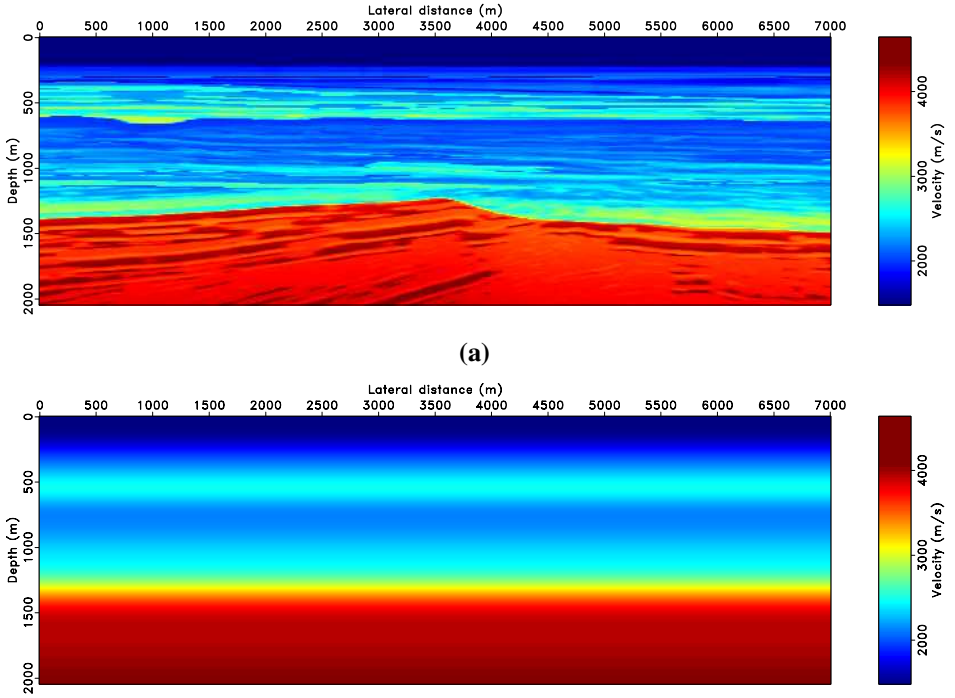
All simulations are carried out with 350 shot and 701 receiver positions sampled at 20m and 10m intervals, respectively, yielding a maximum offset of 7km. To avoid local minima and to improve convergence, the inversions are carried out sequentially in 10 overlapping frequency bands on the interval $2.9 - 22.5$Hz (Bunks et al., 1995), each using 7 different randomly selected simultaneous shots and 10 selected frequencies. For each subproblem, we use roughly 20 iterations of SPGℓ_1. Because we want to reduce the residue as much as possible, we set $\alpha = 0$. This requires roughly 4% of the cost of solving FWI with all data, using ten iterations of LSQR per GN iteration (Paige and Saunders, 1982). Since the costs to carry out this inversion for all data are prohibitive, we rely on a limited-memory quasi-Newton method (l-BFGS, Nocedal and Wright, 2006b) instead, which uses approximately twice the number of PDE solves compared to the new algorithm. The results for l-BFGS and our algorithm with and without independent redraws for the $\mathbf{W}$'s are included in Figure 3.2. We can make the following observations from these results. First, the results from the modified GN have higher resolution and recover the different layers more accurately. This is remarkable, and demonstrates the ability of curvelet sparsity promotion on the updates to "fill in" the null space of the reduced Hessian. Second, the renewals remove visual artifacts in the estimated velocities

(juxtapose Figure 3.2b and 3.2c).

Even though the above example shows that excellent inversion results are attainable working with randomized superpositions, our dimensionality reduction relies on fixed-spread acquisition where each source sees the same receivers. Unfortunately, our reliance on this type of reduction technique limits the applicability of our approach to marine data where the sources and receivers both move and where near and far offsets are missing. This dependence can be avoided if we replace the Gaussian i.i.d. columns of the $\mathbf{W}$'s by randomly selected columns of the Dirac basis. This choice, which corresponds to selecting random subsets of sequential shots, opens the possibility to work with marine data. To test the performance of this type of dimensionality reduction for a marine acquisition with near offsets (up to 100 m and far offsets (from $3000-7000$ m) missing, we rerun the above GN examples with and without renewals for this type of dimensionality reduction. The results are included in Figure 3.3 and clearly illustrate the importance of changing the random subsets of shots for the different GN updates. From the perspective of compressive sensing, we can also explain the slight loss in resolution and quality of the recovered discontinuities. According to this theory, the Gaussian matrix yields better recovery and comparing Figures 3.2b and 3.3a confirms this prediction.

3.5 Discussion

Efforts to control source artifacts related to fast inversions that act on subsets of the data rely mostly on averaging In contrast, we employ sparsity-promotion on the model updates to remove the interferences related to the dimensionality reduction. In that sense, our method is somewhat reminiscent of "gradient preconditioning" (see e.g. Fichtner, 2011) via smoothing but it differs because we leverage curvelet-domain sparsity promotion, which preserves wavefront-like features, removes source crosstalk, and does not need model-specific information such as local dips (Guitton et al., 2010). Our approach has the additional advantage that it, like hybrid (van Leeuwen et al., 2011b; van Leeuwen and Herrmann, 2011) and robust methods (Aravkin et al., 2011), also works with marine data, without relying on correlation/phase-based misfit functionals (Routh et al., 2011).



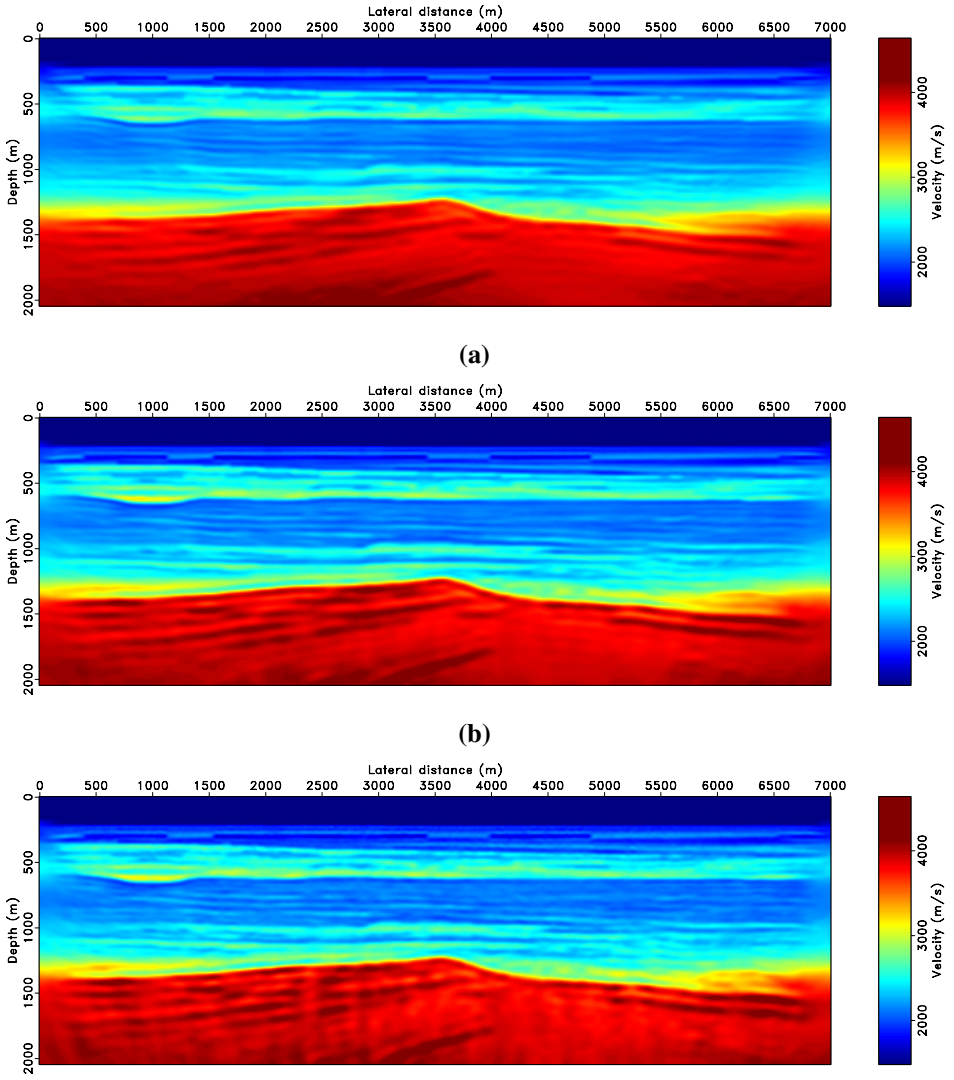
(a)

(b)

Figure 3.1: BG Compass model. **(a)** original model ($\mathbf{m}$), **(b)** initial model ($\mathbf{m}_0$) used to start FWI.

3.6 Conclusions

We introduced an efficient algorithm to solve the full-waveform inversion (FWI) problem by incorporating insights from convex optimization, stochastic optimization, and compressive sensing. By exploiting the convex-composite, multiscale, and separable structure of FWI, we modified the Gauss-Newton (GN) method to produce a new algorithm that permits us to consider GN subproblems as compressive-sensing experiments with dramatically reduced numbers of sources. This reduction, which can be achieved both for time- and frequency-domain formulations of FWI, leads to corresponding speed improvements for the evaluation of the data misfit functional, the Jacobian, and its adjoint. For fixed randomized subsets of data, we were able to establish convergence of our method, which promotes curvelet-domain sparsity on the model updates. We also demonstrated a significant speed-up

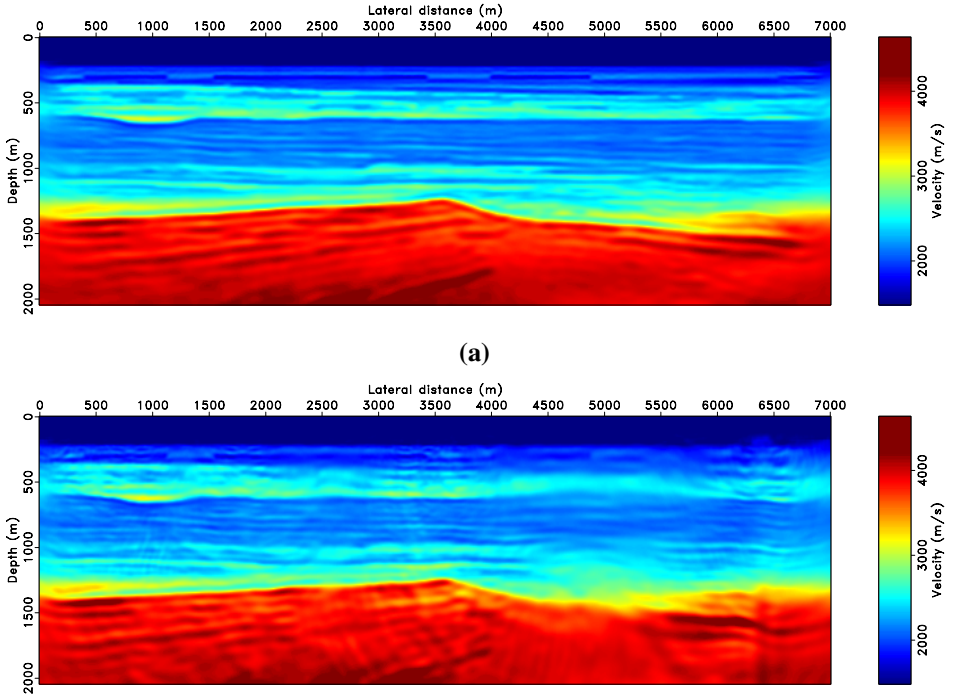


(a)

(b)

(c)

Figure 3.2: Full-waveform inversion results starting from 2.9Hz over 10 frequency bands. **(a)** Inverted result for all data using l-BFGS. **(b)** Inverted result with the modified GN method using 7 simultaneous shots and 10 frequencies. **(c)** the same as **(b)** but without renewals.



(a)

(b)

Figure 3.3: Full-waveform inversion results starting from 2.9Hz over 10 frequency bands. **(a)** Inverted result with the modified GN method using 7 randomly selected sequential shots and 10 frequencies. **(b)** the same as **(a)** but without renewals.

attained by selecting independent randomized subsets of the data for each GN update. While we argue that these renewals remove possible correlations between the subsampling and model iterates, a formal convergence proof of the optimization algorithm still needs to be established.

Application of our algorithm to a complex synthetic data set leads us to the following conclusions. First, dimensionality reduced GN with curvelet-domain sparsity promotion yields higher quality inversion results than do quasi-Newton methods. Second, sparse recovery in combination with randomized dimensionality reduction allows us to speed FWI significantly by iterating on small subsets of the data only. Third, we were able to obtain inversion results from reduced experi-

ments based on either randomized simultaneous sources or on randomized subsets of sequential sources. The latter has the advantage of being suitable for marine acquisition at the cost of a moderately inferior inversion result relative to fixed-spread acquisition. Finally, we find that renewals make a significant difference, in particular for dimensionality reduction based random subsets of sequential shots.

Result: Output estimate for the model $\mathbf{m}$

$k \longleftarrow 0$; $\mathbf{m}^k \longleftarrow \mathbf{m}_0$; // initial model

while not converged **do**

  $\{\underline{\mathbf{D}}^k, \underline{\mathbf{Q}}^k\} \longleftarrow \{\mathbf{D}\mathbf{W}^k, \mathbf{Q}\mathbf{W}^k\}$ with $\mathbf{W}^k \in N(0,1)$; // indep. draw.

  $\underline{\delta\mathbf{D}}^k \longleftarrow \underline{\mathbf{D}}^k - \mathscr{F}[\mathbf{m}^k; \underline{\mathbf{Q}}^k]$; // residual

  $\tau^k \longleftarrow (1-\alpha)\,\|\underline{\delta\mathbf{D}}^k\|_F \,/\, \|\mathbf{C}_2 \nabla \mathscr{F}^*[\mathbf{m}^k; \underline{\mathbf{Q}}^k]\underline{\delta\mathbf{D}}^k\|_\infty$; // one-norm LASSO

  $$\delta\mathbf{x} \longleftarrow \begin{cases} \arg\min_{\delta\mathbf{x}} \frac{1}{2}\|\underline{\delta\mathbf{D}}^k - \nabla\mathscr{F}[\mathbf{m}^k; \underline{\mathbf{Q}}^k]\mathbf{C}_2^H \delta\mathbf{x}\|_F^2 \\ \text{subject to} \quad \|\delta\mathbf{x}\|_1 \leq \tau^k \end{cases}$$

  $\mathbf{m}^{k+1} \longleftarrow \mathbf{m}^k + \gamma^k \mathbf{C}_2^H \delta\mathbf{x}$; // update with line search

  $k \longleftarrow k+1$;

end

Algorithm 2: Dimensionality-reduced Gauss Newton with sparsity

Chapter 4

A modified, sparsity promoting, Gauss-Newton algorithm for seismic waveform inversion

4.1 Introduction

Seismic data can be used to image structures inside the earth on various scales, similar to how a CT scan reveals images of the human body. Earthquake data are used to study the structure of the earth's crust and the core-mantle-boundary. Active seismic experiments, conducted mainly by oil and gas companies, can be used to infer structural information up to about 10 km deep with a typical resolution of $50 - 100$ meters. In such experiments, sources and receivers are placed on the surface or towed through the water. The response of the sequentially detonated explosive sources is measured by as many as 10^6 channels covering areas of 100s of km^2. These experiments produce enormous amounts of data which then have to processed. Most of the data consists of reflected energy with a frequency content of roughly $[5 - 100]$ Hz. Current acquisition practice is moving towards recording

A version of this chapter has been published. Felix J. Herrmann, Xiang Li, Aleksandr Y. Aravkin, and Tristan van Leeuwen. A modified, sparsity promoting, Gauss-Newton algorithm for seismic waveform inversion. *Proc. SPIE*, 2011.

lower frequencies and using larger apertures to capture refracted (transmitted) energy. When the underlying geological structure is simple, the reflection data may be interpreted directly. However, with the ever increasing need for fossil fuels, the industry is moving into geologically more complex areas. The data cannot be interpreted directly and have to be imaged using specialized algorithms. "Migration" is an example of such basic imaging that is widely used in the geophysical community. The basic idea is to correct for the wavepaths along which the reflected data traveled. Most industry practice is still based on a geometric optics approximation of wave propagation. Such algorithms need a smooth "velocity model" that describes the propagation speed of the waves in the subsurface. In general, little is known about the velocity variations on this scale (as opposed to the global scale, where good models exist) and this has to be determined from the data as well.

In contrast to the geometric optics approach, Full-waveform inversion (FWI) relies on modeling the data by solving the wave equation (with finite-differences, for example), and adapting the model parameters (i.e., the coefficients of the PDE) to minimize the least-squares data misfit. This method, first proposed in the early '80sTarantola (1984b), tries to infer a gridded model from the data directly, without making the distinction between the (smooth) velocity model and the image. It quickly became apparent that this approach needs a very good initial guess of the velocity structure to circumvent local minima in the misfit functional that are related to "loop skipping" Tarantola (1984b). The basic idea is that we have to provide information on the low wavenumbers that are missing from the data. With better data (lower frequencies, larger aperture) we need a less detailed initial model. Now that such data is becoming available, waveform inversion may become a viable alternative to more traditional imaging procedures. In order to make waveform inversion feasible for industrial-scale applications, inversion formulations and algorithms must take advantage of dimensionality reduction techniques for working with exceedingly large data volumes. In this paper, we design a modified Gauss-Newton method for FWI that uses dimensionality reduction techniques and ideas from stochastic optimization. The modification we propose promotes transform-domain sparsity on the model updates. Consequently, we are able to incorporate curvelet frames Candes and Demanet (2004); Candès et al. (2006); Hennenfent and Herrmann (2006a) into our framework that offer a compressible representation for

wavefields, which improves FWI.

4.1.1 Full-waveform inversion

Full-waveform inversion is a data fitting procedure that relies on the collection of seismic data volumes and sophisticated computing to create high-resolution models. The corresponding nonlinear least-squares (NLLS) optimization problem is as follows:

$$\underset{\mathbf{m}}{\text{minimize}}\, \phi(\mathbf{m}) := \frac{1}{2}\sum_{i=1}^{K} \|\mathbf{d}_i - \mathscr{F}[\mathbf{m};\mathbf{q}_i]\|_2^2 \,, \tag{4.1}$$

where K is the batch size (number of sources), $\mathbf{d}_i$ represents the data corresponding to the i^{th} (known) source $\mathbf{q_i}$, both organized as vectors, and $\mathscr{F}[\mathbf{m};\mathbf{q}_i]$ is the forward operator for the i^{th} source. The vector of unknown medium parameters is denoted by $\mathbf{m}$. The forward operator $\mathscr{F}$ acts linearly on the sources $\mathbf{q}_i$; that is

$$\mathscr{F}[\mathbf{m}; a\mathbf{q}_i + b\mathbf{q}_j] = a\mathscr{F}[\mathbf{m};\mathbf{q}_i] + b\mathscr{F}[\mathbf{m};\mathbf{q}_j] \,. \tag{4.2}$$

Formulation (4.1) assumes a fixed receiver array.
If we organize the sources and the data as matrices: $D = (\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_K)$ and $Q = (\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_K)$, we may write the objective in (4.1) as

$$\phi(\mathbf{m}) = \frac{1}{2}\|D - \mathscr{F}[\mathbf{m};Q]\|_F^2 \,, \tag{4.3}$$

where $\|\cdot\|_F$ is the Frobenius norm.
We will pause to make several important observations about (4.3). First, the forward operator $\mathscr{F}$ involves the solution of a PDE with multiple right-hand-sides, so the work load is directly proportional to K. Since the sources and receivers may number in the millions, dimensionality reduction techniques become essential for making any headway on the problem. Second, the objective in (4.3) is nonlinear and non-convex. It is, however, convex-composite, meaning that we can write

$$\phi(\mathbf{m}) = \rho(\mathscr{G}(\mathbf{m})), \tag{4.4}$$

with ρ convex, and $\mathscr{G}$ smooth (differentiable). Here, $\rho(X) = \frac{1}{2}\|X\|_F$, and $\mathscr{G}(\mathbf{m}) = D - \mathscr{F}[\mathbf{m};Q]$. This structure allows for natural design and analysis of algorithms

to solve (4.3). The natural approach to FWI motivated by this structure is the Gauss-Newton method, which involves iterative linearization of $\mathscr{G}(\mathbf{m})$ and solution of least-squares problems of the form

$$\underset{\delta\mathbf{m}}{\text{minimize}} \|\delta D - \nabla\mathscr{F}[\mathbf{m}, Q]\delta\mathbf{m}\|_F^2 , \tag{4.5}$$

where $\delta D = D - \mathscr{F}[\mathbf{m}, Q]$ and $\nabla\mathscr{F}$ is the "Jacobian tensor" of $\mathscr{F}$, which acts linearly on $\delta\mathbf{m}$ and produces a matrix as output. This matrix has shot records, i.e, the single-source experiments, organized in its columns. Throughout this paper we refer to the above optimization problem as the Gauss-Newton (GN) subproblem.

4.1.2 Main contribution and relation to existing work

In earlier developments in seismic acquisition and imaging, several authors have proposed reducing the computational cost of FWI by randomly combining sources Krebs et al. (2009b); Moghaddam and Herrmann (2010b); Boonyasiriwat and Schuster (2010); Li and Herrmann (2010b); Haber et al. (2010b). We follow the same approach, but focus the exposition on the Gauss-Newton subproblem, setting the stage for further modifications. We replace (4.5) by

$$\underset{\delta\mathbf{m}}{\text{minimize}} \|\delta DW - \nabla\mathscr{F}[\mathbf{m}, QW]\delta\mathbf{m}\|_F^2 , \tag{4.6}$$

where W is a matrix with i.i.d. random entries with $\widetilde{K} \ll K$ columns. The main computational cost lies in solving a wave-equation for each column of Q, and this strategy aims to significantly reduce this number, replacing Q by QW. We may link this directly to ideas from stochastic optimization by recognizing this modified subproblem as being the GN subproblem of a modified misfit, given by:

$$\widetilde{\phi}(\mathbf{m}; W) = \frac{1}{2}\|DW - \mathscr{F}[\mathbf{m}; QW]\|_F^2. \tag{4.7}$$

If we choose W with unit covariance (i.e., $\mathsf{E}\{WW^H\} = I$) we find that

$$\mathsf{E}\{\widetilde{\phi}(\mathbf{m}; W)\} = \phi(\mathbf{m}). \tag{4.8}$$

Specialized algorithms to deal with such problems go back to the '50s and a detailed overview is given in a later section. The main idea of these algorithms is to make some progress using random realizations of the gradient, relying on using sufficiently many realizations to eventually converge.
We may also view the reduced GN subproblems from the vantage point of compressed sensing, which studies theory and algorithms for the recovery of sparse vectors from severely undersampled systems. Herrmann et. al. Herrmann and Li (2011b) successfully used this approach to make sparsity-promoting seismic imaging more efficient.

Under certain assumptions on the matrix we can recover a sparse vector from such a system by solving a sparsity promoting problem. This is promising, since we need not rely on a Monte-Carlo type sampling strategy common to stochastic methods to recover the solution. It does require, however, that we find a representation in which the solution is sparse (or at least compressible) Donoho (2006). Fortunately, the curvelet frame offers a very efficient (sparse) representation for waveields Candes and Demanet (2004); Candès et al. (2006); Hennenfent and Herrmann (2006a). Curvelets can be thought of as a higher dimensional generalization of wavelets, which capture local information at different directions and scales. Motivated by the optimally sparse representation of wavefields in the curvelet frame Candes and Demanet (2004), we regularize the reduced GN subproblem with an ℓ_1 constraint:

$$\underset{\delta\mathbf{x}}{\text{minimize}} \|\delta DW - \nabla\mathscr{F}[\mathbf{m}, QW]C^H\delta\mathbf{x}\|_F^2 \quad \text{s.t.} \quad \|\delta\mathbf{x}\|_1 \leq \tau, \tag{4.9}$$

where C represents the curvelet transform. The key point here is that the solution $\delta\mathbf{m}$ to the Gauss-Newton subproblem may be interpreted as a wavefield, as we demonstrate in section 4. The formulation (4.9) is a way to regularize the GN-subproblem to take advantage of wavefield sparsity. Because of the convex composite structure of (4.7), the solution to (4.9) is still a descent direction for $\widetilde{\phi}$.
The final algorithm, then, combines ideas from both stochastic optimization and compressed sensing and is shown to be highly effective for our particular application.

4.1.3 Outline of the chapter

The main contribution of the paper is the development of a novel modified Gauss-Newton method for FWI that combines ideas from stochastic optimization and compressed sensing (CS). We therefore first give a brief overview of stochastic optimization and CS techniques in sections 4.2 and 4.3. In section 4.4, we formulate a modified GN method for a particular realization of $\widetilde{\phi}(\mathbf{m}, W)$ and present a convergence proof for it. The practical implementation of both the the modified GN method and the modeling operator is discussed in section 4.5. In the practical version of the method we resample the matrix W (encoding the simultaneous shots) at each realization, which significantly improves the quality of the recovery, but precludes a rigorous convergence theory. Numerical results obtained using the new method are presented in section 4.6, and conclusions follow in section 4.8.

4.2 Stochastic Optimization

Stochastic optimization deals with optimization problems of the form

$$\underset{\mathbf{m}}{\text{minimize}} \big\{ \phi(\mathbf{m}) = \mathsf{E}\{\widetilde{\phi}(\mathbf{m}; W)\} \big\}. \tag{4.10}$$

This approach does not require access to the full misfit and gradient, ϕ and $\nabla\phi$. Instead, we have access to 'noisy' realizations $\widetilde{\phi}$ and $\nabla\widetilde{\phi}$, which are correct on average. In this section we briefly outline two main approaches to essentially get rid of the "noise" in the approach, called Stochastic Average (SA) and Sample Average Approximation (SAA).

4.2.1 Sample Average Approximation (SAA)

A natural approach to pick a "large enough" batchsize $\widetilde{K}$ (i.e., such that $WW^H \approx I$), effectively replacing the expectation by a sample average. This is often referred to in the literature as the Sample Average Approximation (SAA)Nemirovski et al. (2009b). Once drawn, the batch W is fixed; the idea is simply replace the full objective $\phi(\mathbf{m})$ by the subsampled variant $\widetilde{\phi}(\mathbf{m}; W)$. When some additional assumptions are satisfied, the optimal value of $\widetilde{\phi}$ converges to the optimal value of ϕ with probability 1 (see Shapiro (2003); Shapiro and Nemirovsky (2005)) . From

a practical point of view, the SAA approach is appealing because it allows flexibility in the choice of algorithm for the solution of the subsampled problem. In particular, we may directly use the GN method to minimize the reduced misfit.

4.2.2 Stochastic Approximation

A second alternative is to apply specialized stochastic optimization methods to problem (4.10) directly. This is often referred to as the Stochastic Approximation (SA). The main idea of such algorithms is to pick a new random realization W^k for each iteration k. Notably, some methods include averaging over past iterations to suppress the noise introduced by the randomized source encoding. This approach yields an iterative algorithm of the form

$$\mathbf{m}^{k+1} = \mathbf{m}^k - \gamma_k \nabla \mathbf{s}^k \,,$$

where the search direction is typically given by a realization of the gradient: $\mathbf{s}^k = \nabla\widetilde{\phi}(\mathbf{m}^k; W^k)$ The batch size is typically very small ($K = \mathscr{O}(1)$), and $\{\gamma_k\}$ represent step sizes taken by the algorithm, which are picked ahead of time.
(Betrsekas and Tsitsiklis, 2000, propopsition 3) provides a convergence theory for a class of SA algorithms for directions $\mathbf{s}^k + \omega^k$, as long as several technical conditions hold:

1. ϕ is differentiable with $\nabla\phi$ Lipshitz continuous.
2. $\mathsf{E}[\omega] = 0$.
3. The expected value of the search directions are descent directions for ϕ, i.e. $\nabla\phi(\mathbf{m}^k)^H \mathsf{E}[\mathbf{s}^k] < 0$.
4. There exist positive constants c_1, c_2 and c_3 so that
   (a) $c_1\|\nabla\phi\|^2 \leq -\nabla\phi(\mathbf{m}^k)^T\nabla(\mathbf{s}^k + \omega^k)$,
   (b) $\|\mathbf{s}^k + \omega^k\| \leq c_2(1 + \|\nabla\phi\|)$, and
   (c) $\mathsf{E}[\|\omega\|^2] \leq c_3(1 + \|\nabla\phi\|^2)$.
5. $\sum_{\nu=0}^{\infty}\gamma_\nu = \infty$, $\sum_{\nu=0}^{\infty}\gamma_\nu^2 < \infty$. A common example is $\gamma_\nu \propto \frac{1}{\nu}$.

Even though the modified GN algorithm in section 4 has a convergence theory for a particular random realization W (i.e. for solving the SAA problem $\widetilde{\phi}(\mathbf{m}, W)$), the SA theory presented here does not apply to the practical algorithm where the W's are redrawn (presented in Section 6). In particular we cannot guarantee that condition 3 above is satisfied by the model update $\delta\mathbf{m}$ derived from the modified Gauss-Newton subproblem. However, redrawing W's substantially improves our recovery, as shown in section 6.

4.3 Sparsity Regularization and Compressive Sensing

Compressed or Compressive Sensing (CS) provides theory centered around recoverability of sparse signals using linear measurements Candes (2006); Donoho (2006). The basic problem is to solve an underdetermined linear system $RM\mathbf{f} = \mathbf{b}$, where RM is a flat matrix consisting of the measurement matrix M and the restriction matrix R, and $\mathbf{f}$ is known to be sparse in some basis. This latter fact can be written as $\mathbf{f} = S^H\mathbf{x}$, where S is the basis (or frame) , and $\mathbf{x}$ is sparse or compressible. Denoting $A = RMS^H$, we now want to find the sparsest solution of the system $A\mathbf{x} = \mathbf{b}$, or

$$\underset{\mathbf{x}}{\text{minimize}} \|\mathbf{x}\|_0 \quad \text{s.t.} \quad A\mathbf{x} = \mathbf{b}, \tag{4.11}$$

where $\|\cdot\|_0$ denotes the ℓ_0 "norm" given by the number of nonzero elements of a vector. Unfortunately, solutions of this type of non-convex optimization problems are nearly impossible to compute for large problems because they require a combinatorial search over all possible subsets of columns of A to find the solution with the fewest nonzero elements. One of the major findings of CS is that under some conditions on A and $\mathbf{x}$, the solution can be recovered by solving the convex optimization problem

$$\underset{\mathbf{x}}{\text{minimize}} \|\mathbf{x}\|_1 \quad \text{s.t.} \quad A\mathbf{x} = \mathbf{b}\,. \tag{4.12}$$

Whether solving this problem, known as Basis Pursuit (BP), recovers the correct sparse signal depends on the sparsity level of $\mathbf{x}$, the number of measurements, and the *Restricted Isometry Property* (RIP) constant of the matrix A. Roughly speaking, the RIP constant measures how far the matrix A is from a unitary matrix when

acting on sparse vectors. Again, checking this condition for arbitrary matrices requires a combinatorial search over subsets of columns of A. To overcome this difficulty, the *mutual coherence*, which is the maximum (normalized) inner product between any two columns of A (i.e., the maximum off-diagonal entry of $A^H A$), is an often-used heuristic. Low mutual coherence is necessary for recovery guarantees of a sparse signal by solving a sparsity promoting program.

When noise is present in the data we may instead solve the Basis Pursuit Denoise (BPDN) problem

$$\underset{\mathbf{x}}{\text{minimize}} \|\mathbf{x}\|_1 \quad \text{s.t.} \quad \|A\mathbf{x} - \mathbf{b}\|_2 \leq \sigma \,, \tag{4.13}$$

where σ is the expected noise level in the data van den Berg and Friedlander (2008a). This problem is hard to solve, but turns out to be equivalent to two related formulations

$$\underset{\mathbf{x}}{\text{minimize}} \|A\mathbf{x} - \mathbf{b}\|_2^2 + \lambda \|\mathbf{x}\|_1 \tag{4.14}$$

and

$$\underset{\mathbf{x}}{\text{minimize}} \|A\mathbf{x} - \mathbf{b}\|_2^2 \quad \text{s.t.} \quad \|\mathbf{x}\|_1 \leq \tau \tag{4.15}$$

known as the QP and LASSO problems, respectively. The equivalence is true in the sense that for each σ, there are unique values for λ and τ so that the solutions of (4.13, 4.14, 4.15) all coincide. However, the values of these parameters are not known ahead of time. Therefore, most algorithms that solve (4.13) do some sort of continuation either in λ (see Kim et al. (2007)) or in τ (see van den Berg and Friedlander (2008a)). In both cases, the iterates start sparse and additional components are allowed to enter the solution to bring down the residue. Even though continuation in parameters λ and τ for QP and LASSO subproblems can be used to solve (4.13), the LASSO-based approach offers two advantages: the Spectral Projected Gradient method can be used to quickly solve (4.15) for very large linear systems, and the continuation in τ can be naturally derived using the graph of the value function for (4.15). The spectral projected gradient (SPG) method for (4.15) is detailed in ((van den Berg and Friedlander, 2008a, Algorithm1). SPG is

an iterative method, with iterates taking the form

$$\mathbf{x}^{k+1} = P_\tau[\mathbf{x}^k + \gamma_k \mathbf{s}^k] \,,$$

where the search direction $\mathbf{s}^k$ is the negative gradient of the objective ($\mathbf{s}^k = -2A^T(A\mathbf{x}^k - \mathbf{b})$), P_τ is the projection operator onto the one-norm ball of radius τ (the set $\{\mathbf{x} : \|\mathbf{x}\|_1 \leq \tau\}$), and γ_k is a line search parameter chosen according to the Barzilai-Borwein scheme (see e.g. (Birgin et al., 2010, Algorithm 2.1)). Since the SPG method has already been proven to be very successful in solving large-scale CS problems in seismic exploration Hennenfent et al. (2008a), we use it as a subroutine in the current convex-composite formulation to solve the modified Gauss-Newton LASSO subproblems.

4.4 Modified Gauss-Newton Method for SAA Approach

Recall the dimensionality reduced misfit function defined in (4.7):

$$\widetilde{\phi}(\mathbf{m}; W) = \frac{1}{2}||DW - \mathscr{F}[\mathbf{m}; QW]||_F^2.$$

This problem has the same convex composite structure (see (4.4)) as the full misfit, and we exploit this structure to design an algorithm for solving (4.7). We begin with a basic Gauss-Newton method, which is an iterative algorithm of the form

$$\mathbf{m}^{k+1} = \mathbf{m}^k + \gamma_k \delta\mathbf{m}^k \,,$$

where $\delta\mathbf{m}^k$ solves

$$\underset{\delta\mathbf{m}}{\text{minimize}} \, \|\delta D^k W - \nabla\mathscr{F}[\mathbf{m}^k; QW]\delta\mathbf{m}\|_F^2 \,,$$

the quantity $\delta D^k W$ is the dimensionality-reduced linearized data residual $DW - \mathscr{F}[\mathbf{m}^k; QW]$, and γ_k is a line search parameter.
We adapt the Gauss-Newton method by using the following key observations:

- **The scattering operator is diagonal in phase space, and thus has low**

mutual coherence: The normal operator $\nabla\mathscr{F}[\mathbf{m}^k;QW]^H\nabla\mathscr{F}[\mathbf{m}^k;QW]$ has a very special structure in exploration seismology. Namely, in the high-frequency limit, this operator is diagonal in phase space (more precisely, it is a pseudo-differential operator)Beylkin (1984); ten Kroode et al. (1998); de Hoop and Brandsberg-Dahl (2000); Stolk and Symes (2003) for point sources. More specifically, we can write

$$\nabla\mathscr{F}[\mathbf{m}^k;Q]^H\nabla\mathscr{F}[\mathbf{m}^k;Q] = BL^{\frac{1}{2}},$$

where B is a positive-definite scaling matrix and L is a discrete LaplacianSymes (2008a); Herrmann et al. (2008a, 2009a). From this factorization, we expect a very low mutual coherence between the columns of the scattering operator. We do not expect either the curvelet frame or the random mixing of the sources to increase mutual coherence, since $\mathsf{E}[WW^H] = I$ and $CC^H = I$.

- **The GN search direction is sparse in curvelets:** The gradient of the misfit, $\nabla\phi = \nabla\mathscr{F}^H\delta D$, can be computed by correlating two wavefields (see (4.27)), and this correlation is again a wavefield. As already noted, curvelets give an optimally sparse representation of wavefields, so we expect the gradient to be sparse in this frame. Next, the solution to the standard Gauss-Newton method is given by

$$\delta\mathbf{m} = -(\nabla\mathscr{F}^H[\mathbf{m}^k;Q]\nabla\mathscr{F}[\mathbf{m}^k;Q])^{-1}\nabla\phi. \tag{4.16}$$

Using the special structure of the Hessian, outlined above, we argue that $\delta\mathbf{m}$ can be interpreted as a scaled wavefield, and hence can also be sparsely represented with curvelets. Note that this argument does not depend on a correct or nearly correct velocity model $\mathbf{m}^k$, but only on the form of $\delta\mathbf{m}$ in (4.16). Thus updates are expected to be sparse in curvelets even when $\mathbf{m}^k$ is far away from the true solution, e.g. at the beginning of FWI.

These observations motivate us to replace the standard GN subproblem by a sparsity promoting LASSO variant

$$\begin{aligned} \delta\mathbf{x}^k &= \arg\min_{\delta\mathbf{x}} \|\delta D^k W - \nabla\mathscr{F}[\mathbf{m}^k; QW]C^H\delta\mathbf{x}\|_F^2 \quad \text{s.t.} \quad \|\delta\mathbf{x}\|_1 \le \tau_k \\ \delta\mathbf{m}^k &= C^H\delta\mathbf{x}^k\,, \end{aligned} \tag{4.17}$$

where τ_k are parameters to be selected. Note that taking $\tau_k = 0$ forces $\delta\mathbf{m}^k = 0$, while taking τ_k to be very large gives us the ordinary Gauss-Newton solution update for $\widetilde{\phi}$. As discussed above, this subproblem can be solved using the SPG method. Denote by v_k the value function for the k-th subproblem (4.17):

$$\begin{aligned} v_k(\tau_k) &= \min_{\delta\mathbf{x}} \|\delta D^k W - \nabla\mathscr{F}[\mathbf{m}^k; QW]C^H\delta\mathbf{x}\|_F^2 \quad \text{s.t.} \quad \|\delta\mathbf{x}\|_1 \le \tau_k \\ &= \|\delta D^k W - \nabla\mathscr{F}[\mathbf{m}^k; QW]\delta\mathbf{m}^k\|_F^2\,. \end{aligned} \tag{4.18}$$

The value function is carefully studied in van den Berg and Friedlander (2008a) , where its graph is dubbed the "Pareto trade-off curve". The graph of the value function traces the optimal trade-off between the two-norm of the residual and the one-norm of the solution. Because the value function is continuously differentiable, convex and strictly decreasing, the LASSO formulation (4.17) has a corresponding BPDN problem (4.13) for a unique σ. Hence, our approach can be thought of as finding the sparse search direction for to the full problem from subsampled measurements. The noise level in this formulation, however, refers to the error in the linearization and the question is how to choose the magnitude of this mismatch σ, or correspondingly, how to choose the right τ, for each subproblem. This question is addressed in Section 4.5.

The above interpretation—where LASSO problems are argued to recover significant transform-domain coefficients à la CS—has no rigorous justification, particularly due to lack of CS results for frames and lack of RIP constants for $\nabla\mathscr{F}$ in the seismic application. Nonetheless, the point here is that the LASSO problem (4.17) is particularly well-tailored to ideas related to sparsity promotion and CS. To arrive at a convergence theory for the modified GN algorithm, the solution of

the modified subproblem must be a descent direction of $\widetilde{\phi}$. This condition is ensured by the convex-composite structure of $\widetilde{\phi}$. For any convex-composite function $\rho(\mathscr{G}(\mathbf{m}))$, the directional derivative $\rho'(\mathbf{m}, \delta\mathbf{x})$ exists and satisfies

$$\rho'(\mathbf{m}; \delta\mathbf{x}) \leq \rho(\mathscr{G}(\mathbf{m}) + \nabla\mathscr{G}^H \delta\mathbf{x}) - \rho(\mathscr{G}(\mathbf{m})) \tag{4.19}$$

from (Burke, 1990, Lemma 1.3.1). Because the subproblem (4.17) minimizes $\rho(\mathscr{G}(\mathbf{m}) + \nabla\mathscr{G}^H \delta\mathbf{x})$ over the one norm ball of radius τ_k, we know that $\delta\mathbf{m}^k$ satisfies

$$v_k(\tau_k) = \|\delta D^k W - \nabla\mathscr{F}[\mathbf{m}^k; QW]\delta\mathbf{m}^k\|_F \leq \|\delta D^k W\|_F \,,$$

with equality only if we have stationarity. Combining this with (4.19), we have

$$\widetilde{\phi}'(\mathbf{m}^k; \delta\mathbf{m}^k) \leq v_k(\tau_k) - \|\delta D^k W\|_F < 0 \,,$$

unless $\mathbf{m}^k$ is a stationary point, in which case $\widetilde{\phi}'(\mathbf{m}^k; \delta\mathbf{m}^k) = 0$. Therefore, the k-th the LASSO subproblem yields a descent direction for the full nonlinear SAA problem for any $\tau_k > 0$, unless we have already reached a local minimum. The full development is shown in Algorithm 3. If we make the additional assumptions

1. The sequence $\{\tau_k\}$ is bounded, and
2. $\nabla\mathscr{F}[\mathbf{m}; QW]$ is uniformly continuous on the convex closure of $\mathbf{m}$ satisfying $\widetilde{\phi}(\mathbf{m}, W) < \widetilde{\phi}(\mathbf{m}_0, W)$,

where 2 above is a standard technical assumption, then hypotheses of (Burke, 1990, Theorem 2.1.2) and (Burke, 1990, Corollary 2.1.2) are satisfied, yielding a global convergence theory for Algorithm 3.

4.5 Practical implementation

4.5.1 Modified Gauss-Newton approach

We propose slight modifications to Algorithm 3. Motivated by the SA approach, we found that resampling the matrix W at each linearization (i.e. using W^k instead of W) improves recovery significantly. Intuitively, it makes sense that using several

initialize $\mathbf{m}$, $k \leftarrow 0$, $\Delta_1 \leftarrow 1$, ε, c.
while $\Delta_k > \varepsilon$ **do**
  $k \leftarrow k+1$
  Compute residual $\delta D^k W = D^k W - \mathscr{F}[\mathbf{m}^k; QW]$
  $\delta\mathbf{x}^k \leftarrow \arg\min_{\delta\mathbf{x}} \left\{ \begin{array}{l} \|\delta D^k W - \nabla\mathscr{F}[\mathbf{m}^k; QW] C^T \delta\mathbf{x}\|_F^2 \\ \text{s.t.} \quad \|\delta\mathbf{x}\|_1 \le \tau_k \end{array} \right\}$
  $\delta\mathbf{m}^k = C^T \delta\mathbf{x}^k$
  $\Delta_k = \|\delta D^k W - \nabla\mathscr{F}[\mathbf{m}^k; Q]\delta\mathbf{m}^k\|_F^2 - \|\delta D^k W\|_F^2$
  Pick λ_k to ensure $\widetilde{\phi}(\mathbf{m}^k + \lambda_k \delta\mathbf{m}^k; W) < \widetilde{\phi}(\mathbf{m}^k; W) + c\lambda_k\Delta_k$ (sufficient decrease condition)
  $\mathbf{m}^{k+1} \leftarrow \mathbf{m}^k + \lambda_k \delta\mathbf{m}^k$
end while

Algorithm 3: GN-Method for FWI with Sparse Updates

different realizations may improve the result, as we remove the bias introduced by a particular random sampling.

As shown above, the direction $\delta\mathbf{m}^k$ is a descent direction for $\widetilde{\phi}$ at $\mathbf{m}^k$ for any positive τ_k, with the requirement that the sequence $\{\tau_k\}$'s are are bounded imposed by the convergence theory for the SAA objective $\widetilde{\phi}$. Nonetheless, practical implementation requires a systematic way to select τ_k. A reasonable approach is to require $v_k(\tau_k) = \alpha v_k(0)$ for some given $\alpha < 1$. Using a linear approximation of v_k we find:

$$\tau_k \approx (\alpha - 1) v_k(0) / v_k'(0) \,. \tag{4.20}$$

A closed-form expression for v' is computed in (van den Berg and Friedlander, 2008a, Theorem 2.1); in our context (4.20) is given by

$$v_k'(0) = \frac{(\alpha-1)\|\delta D^k W^k\|_2}{\|C\nabla\mathscr{F}[\mathbf{m}^k; W^k]^H \delta D^k W^k\|_\infty} \,. \tag{4.21}$$

When sampling W^k at every iteration, it is not clear what linesearch criterion to use. In FWI, we are typically interested in doing a fixed number of iterations (as much as computing resources allow); motivated by SA algorithms which use prescribes fixed sequences of steplengths, we picked the steplengths to be constant, and found this to work well in practice.

The practical implementation with full details is given in Algorithm 4.

initialize $\mathbf{m}$, $k \leftarrow 0$, $\Delta_1 \leftarrow 1$, $\varepsilon \leftarrow 10^{-6}$.
while $\Delta_k > \varepsilon$ and $k < k_{\max}$ **do**
  $k \leftarrow k+1$
  Sample to get W^k
  Compute residual $\delta D^k W^k = DW^k - \mathscr{F}[\mathbf{m}^k; QW^k]$
  $\tau_k = (\alpha-1)\|\delta D^k W^k\|_2 / \|C\nabla\mathscr{F}[\mathbf{m^k}; QW^k]^T(\delta D^k W^k)\|_\infty$
  $\delta\mathbf{x}^k \leftarrow \arg\min_{\delta\mathbf{x}} \|\delta D^k W^k - \nabla\mathscr{F}[\mathbf{m}; QW^k]C^T\delta\mathbf{x}\|_F^2 \quad \text{s.t.} \quad \|\delta\mathbf{x}\|_1 \le \tau_k$
  $\delta\mathbf{m}^k = C^T\delta\mathbf{x}^k$
  $\Delta_k = \|\delta D^k W^k - \nabla\mathscr{F}[\mathbf{m}; Q]\delta\mathbf{m}^k\|_F^2 - ||\delta D^k W^k||_F^2$
  $\mathbf{m}^{k+1} \leftarrow \mathbf{m}^k + \gamma_k\delta\mathbf{m}^k$
end while

Algorithm 4: GN-Method for FWI in Practice

4.5.2 Modeling operator

The modeling operator, $\mathscr{F}[\mathbf{m}, Q]$ is implemented via a frequency-domain finite-difference method. The wavefield for a single frequency ω is obtained by solving a discrete Helmholtz system:

$$H[\omega;\mathbf{m}]U(\omega) = Q(\omega), \tag{4.22}$$

where H is a 9-point, mixed-grid discretizationJo et al. (1996b) of the Helmholtz operator $\omega^2\mathbf{m} + \nabla^2$. The data for a single frequency are obtained by sampling the wavefield at the receiver locations: $D(\omega) = PU(\omega)$, where P is the sampling operator. The modeling operator, finally, produces data for several frequencies and stacks the results. The action of the scattering operator on a vector $\delta\mathbf{m}$ for each frequency ω can be computed as follows:

$$\text{Solve} \qquad H[\omega,\mathbf{m}]U(\omega) = Q \tag{4.23}$$

$$\text{Solve} \qquad H[\omega,\mathbf{m}]^H\delta U(\omega) = \omega^2\mathsf{diag}(\delta\mathbf{m})U(\omega)\,. \tag{4.24}$$

The action of the adjoint for each frequency ω is calculated as follows:

$$\text{Solve} \qquad H[\omega, \mathbf{m}]U(\omega) = Q \tag{4.25}$$

$$\text{Solve} \qquad H[\omega, \mathbf{m}]^H V(\omega) = P^H \delta D(\omega) \tag{4.26}$$

$$\text{Compute} \qquad \delta\mathbf{m} = \sum_{\omega} \omega^2 \mathsf{diag}(UV^H)\,. \tag{4.27}$$

These formulas can be derived via the adjoint-state methodLailly (1983); Tarantola (1984b). We refer to Plessix (2006) for a detailed overview of such techniques in geophysics.

4.5.3 Inversion strategy

A well-known strategy in full waveform inversion is to invert the data starting from low frequencies and gradually moving to higher frequenciesBunks et al. (1995); Pratt et al. (1996). This helps to mitigate some of the issues with local minima. In this case we simply apply the proposed GN algorithm for a fixed number of iterations on a certain frequency band and use the end result as initial guess for the next frequency band.

4.6 Results

We test the proposed method on a part of the BG-Compass synthetic benchmark model. The velocity is depicted in figure 4.2. The data are generated for 350 sources and 700 receivers, all regularly spaced along the top of the model. The initial model for the inversion is depicted in figure 4.2. We used 7 simultaneous sources (columns in W) for this experiment. Hence, a single evaluation of the misfit is 50 times cheaper than an evaluation of the full misfit. The subproblems are solved using 20 SPG iterations. The cost of calculating the update in this case is then comparable to one evaluation of the full misfit.

The inversion is carried out in 10 partially overlapping frequency bands with 10 frequencies each, starting at 2.9 Hz and going up to 25 Hz. We perform 10 GN iterations for each frequency band and use the end result as starting model for the next band. The result with and without renewals are shown in figure 4.3. As a

benchmark, we also show the result obtained with L-BFGS on the full data.
The convergence history in terms of the model mistmatch is also shown. The renewals clearly benefit the inversion, giving a less noisy final result as well as a smaller ℓ_2 model mismatch. The renewals appear to be especially beneficial in the later stage of the inversion. The modified GN method outperforms L-BFGS in this example.

4.7 Discussion

In this paper, we designed a modified Gauss-Newton algorithm for seismic waveform inversion using ideas from stochastic optimization and compressive sensing. Stochastic optimization techniques and dimensionality reduction are used to yield a method that makes fast progress on the whole problem but works only on small randomized subsets of the data at a time. The randomsubsampling weights are periodically redrawn, to remove any bias introduced by a particular weighting matrix and further speed up the progress of the method.
The randomly subsampled Gauss-Newton subproblems may be seen as Compressive Sensing experiments, where a sparse vector is reconstructed from randomly undersampled measurements. Together with the compressibility of seismic images Herrmann and Li (2011a) and wavefieldsSmith (1998); Demanet and Peyré (2011) in Curvelets, this motivated the proposed modification of the Gauss-Newton subproblem to include curvelet transform-domain sparsity of the updates.

The main innovation of the new method is a rigorous way to exploit the compressibility of seismic wavefields and images in the curvelet domain in the context of a large-scale application with a nonlinear forward model. Specifically, the Gauss-Newton subproblems are subsampled using randomization techniques and then regularized by a constraint on the one-norm of the curvelet representation of the update, turning them into LASSO problems. The purpose of this regularization is to "fill in" the null space of the wave-equation Hessian, using curvelet-domain sparsity promotion. While the LASSO problems are are harder to solve, they remain feasible when dimensionality reduction techniques are used.

The sparse regularization of the updates may also be a way to get around the problem of "loop skipping" (local minima). While good starting models and mul-

tiscale continuation methods have successfully mitigated some of the ill effects of these local minima, curvelets and sparsity promotion may be an additional safeguard for getting trapped in a local minimum. This is because strict constraints on the ℓ_1-norm of the updates forces components to enter the solution slowly, and as a result curvelets in model space map to "curvelet images" in the data space that share characteristics of the scale and direction of the corresponding curvelet in the model space. As a consequence, the misfit functional is calculated over relatively small subsets of "curvelet images" that have some support and direction and this reduces the effects of "loop skipping".

4.8 Conclusions

We present a modified Gauss-Newton algorithm for seismic waveform inversion. Using random source superposition, we reduce the computational cost involved in solving Gauss-Newton subproblems. Our approach can be seen as an instance of the Sample Average Approximation method, which introduces random noise as source crosstalk in the updates. The noise level is controlled by the batch size (the number of randomized sources); with larger batch size corresponding to lower noise level. To regularize the subproblems and to suppress the noisy source-cross talk, we add an ℓ_1 constraint on the curvelet coefficients of the updates. The rationale for adding this constraint lies in curvelet-domain compressibility of seismic wavefields, which is due to the special representation of updates as correlations of source and residual wavefields. This argument in combination with curvelet-domain compressibility of seismic images motivated us to develop and implement a modified Gauss-Newton method with LASSO subproblems.

Using the convex-composite structure of the problem, we provide a global convergence theory for this algorithm for a single fixed random realization (batch) of simultaneous shots. We supplemented this theoretical proof with a heuristic argument justifying the redrawing of random source weights after solving each Gauss-Newton subproblem. Even though the convergence theory does not extend to this case, we argue that these renewals remove bias introduced by a particular realization of the random weights, and show that incorporating renewals leads to better results.

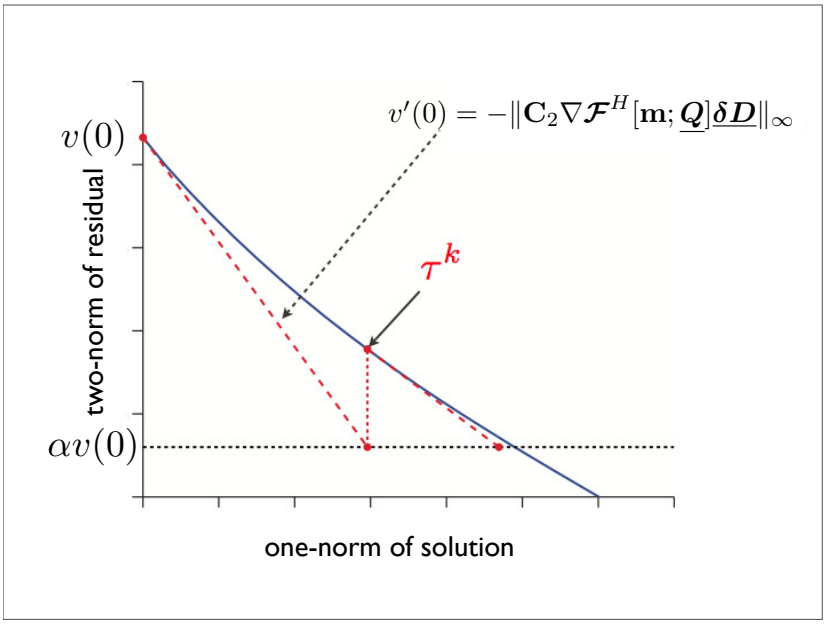


Figure 4.1: Schematic depiction of pareto curve used to select τ for the GN subproblems.

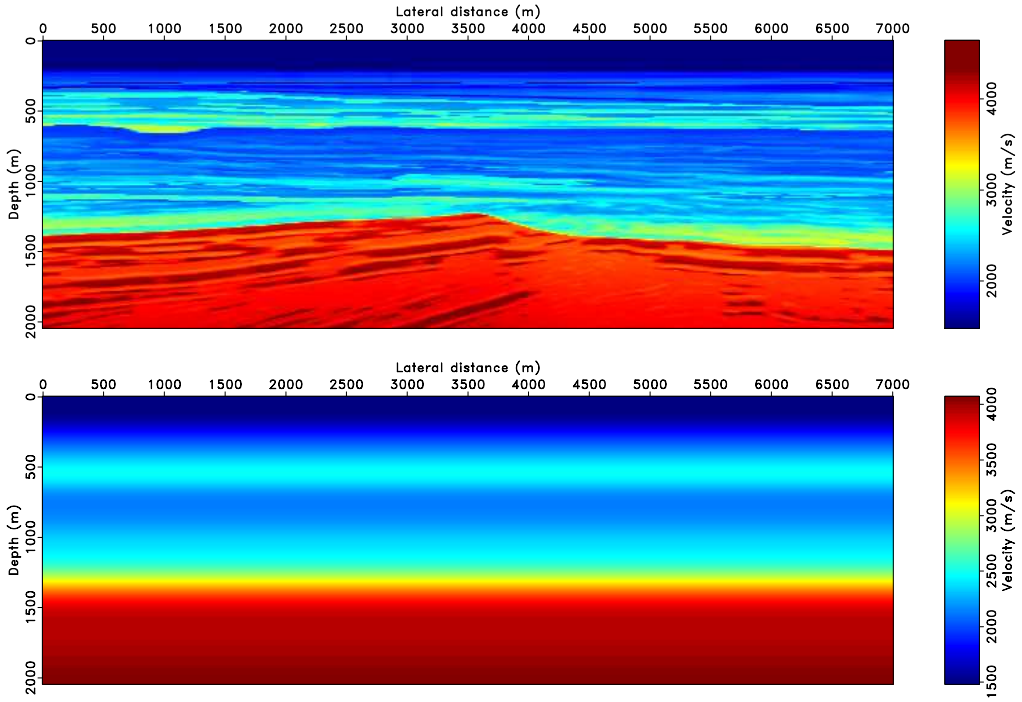


Figure 4.2: Compass Benchmark model with velocities ranging from 1480 - 4500 m/s (top) and initial model used for the inversion (bottom). Note the total lack of lateral variation in the initial model.

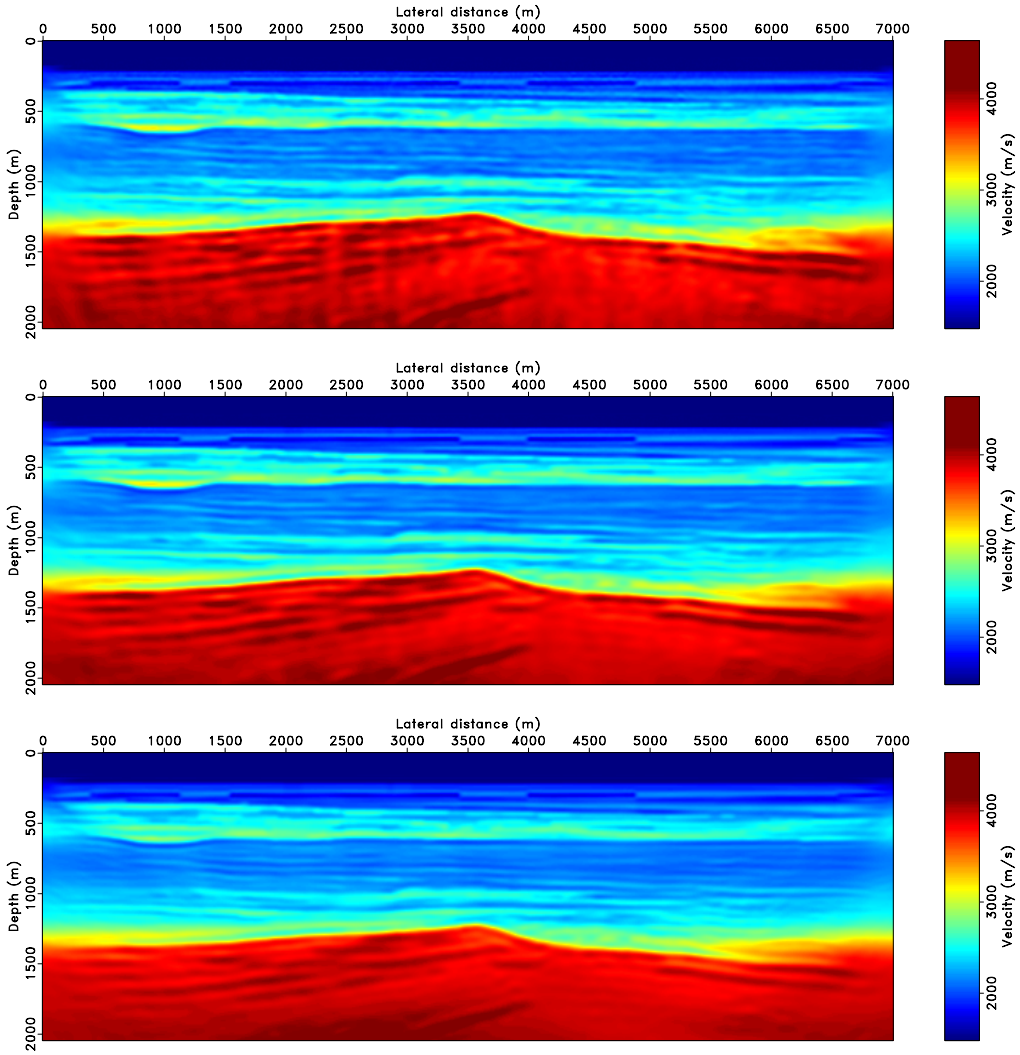


Figure 4.3: Inversion result for the modified Gauss-Newton algorithm without (top) and with (middle) renewals. The result obtained with a standard Quasi-Newton approach is depicted in bottom. The latter approach does not include dimensionaility reduction techniques.

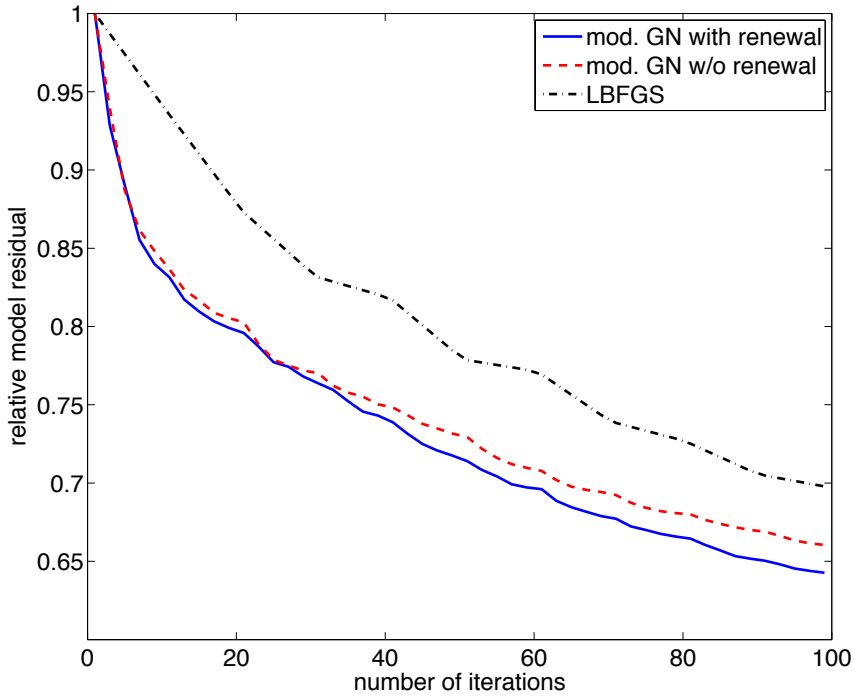


Figure 4.4: Convergenve in terms of the reconstruction error of the modified GN approach (with and without renewals) and the Quasi-Newton approach.

Chapter 5

Modified Gauss-Newton full-waveform inversion explained—why sparsity-promoting updates do matter

5.1 Introduction

Full-waveform inversion (FWI, Pratt et al., 1998b; Virieux and Operto, 2009) aims to reap information on underground physical medium parameters, such as spatial velocity and density distributions, from observed seismic measurements collected at the surface or within well bores. Mathematically, FWI corresponds to partial-differential equation (PDE) constrained optimization problem where the PDE constraints are generally eliminated and where the medium parameters are obtained by minimizing the least-squares misfit between observed and modeled data (Tarantola, 1984a; Pratt et al., 1998a). In the last twenty years, numerous first-order gradient-based methods have been developed to solve this inversion problem, in-

A version of this chapter has been submitted to Geophysics.

cluding gradient descent (Pratt et al., 1998a; Warner et al., 2013), nonlinear conjugate gradients (Gilbert and Nocedal, 1992; Mora, 1987; Tarantola, 1986; Crase et al., 1990) and so on. However, as reported by Pratt et al. (1998a) and Shin et al. (2001)], first-order methods may suffer from slow convergence which is, in part, related to difficulties in calculating reliable step lengths, and also, to the fact that first-derivative information only is used. This slow convergence may lead to inferior results in situations where one can only afford a small number of iterations that utilize all observed data.

As shown by Mtivier et al. (2013), Pratt et al. (1998a) and Gratton et al. (2007), second-order methods have the potential to achieve better convergence than first-order methods when the starting model is reasonably close to the true model. In this situation, inverting the Hessian matrix—i.e., the matrix that contains second-derivative information—compensates for source-related blurring, limited aperture and other amplitude-related effects. Unfortunately, true Hessian matrices are impossible to explicitly form for large scale problems, and they are challenging to invert iteratively since this matrix is not guaranteed to be positive definite. For this reason, it is common to approximate the Hessian by the semi-positive definite Gauss-Newton Hessian, which can readily be inverted using iterative methods (Hestenes and Stiefel, 1952). While incorporating partial second-order information can lead to significant improvements in the convergence (Mtivier et al., 2013; Pratt et al., 1998a), the computational costs of inverting the Gauss-Newton Hessian iteratively, for each model update, become quickly prohibitive expensive because evaluation of the action of the Gauss-Newton Hessian—formed, for instance, by compounding Jacobian (linearized Born modeling) matrix and its adjoint (migration)—require multiple expensive PDE solves.

To overcome the generally prohibitive expensive costs of computing Gauss-Newton updates, each of which involve the (approximate) iterative least-squares solution of the wave-equation Jacobian, we propose a curvelet-domain sparsity-promoting method (Li et al., 2012) that only works with randomized subsets of source experiments. By limiting the number of PDE solves, we obtain an order-of-magnitude improvement in computational efficiency where least-squares inversions of the Jacobian could be computed at the cost of roughly one reverse-time migration with all data [chapter 2](Tu et al., 2013). In this approach, we base our

argument on the observations that Gauss-Newton updates are sparse in the curvelet domain and random subsampling related artifacts are not sparse in this domain, thereby creating favorable conditions for sparsity promotion. During these inversions, incoherent (and therefore non-sparse) energy is mapped via curvelet-domain sparsity-promotion to coherent events in the Gauss-Newton updates, in a similar way as we demonstrate for least-squares migration (Herrmann and Li, 2012; Tu et al., 2013). While the resulting modified Gauss-Newton method (MGN, Herrmann et al., 2011b) yields encouraging results in improving the computational performance, and more importantly, the quality of FWI (Li et al., 2012), our approach becomes problematic and differs fundamentally from existing regularization methods for inverse problems because it imposes ℓ_1-constraints on the Gauss-Newton updates rather than on the model iterates themselves. The latter is more common and undergirds recent work by (Gauthier et al., 1986; Hansen, 1998; Askan et al., 2007). The main aim of our work is to provide arguments explaining why and when our approach forms an attractive alternative to imposing sparsity constraints on the model directly.

In spite of the fact that regularizing model iterations, rather than model updates, seems to make more intuitive sense, it appears that these type of regularization methods rely critically on prior knowledge of the sparsity level (the ℓ_1-norm) of the (unknown) model. This critical dependence may hamper application of these methods to large scale problems or at least calls for a continuation technique that somehow relaxes the constraint, a topic of active research in current-day inverse problems (van den Berg and Friedlander, 2008b; Lin and Herrmann, 2013; Hennenfent et al., 2008b). As we will demonstrate, imposing constraints on the linearized, and therefore convex, subproblems by using a combination of theoretical convex-composite and problem-specific arguments shows that under certain circumstances sparse models can be obtained from model updates that solve sparsity-promoting Gauss-Newton subproblems. By means of carefully selected examples, we demonstrate for which type of problems and how this can be accomplished. We find that it suffices to choose a series of conservative sparsity levels for the ℓ_1-norm constraints on the updates as long as the supports—i.e., the locations of the non-zeros—do not differ too much for the different Gauss-Newton updates so that their sum, and therefore the model iterate itself, remains sparse as well. As our

examples will demonstrate, the SPGℓ_1-framework (van den Berg and Friedlander, 2008b), which holds for convex problems with convex constraints, can be used to determine sparsity levels that lead to sparse updates that meet the above criterion for certain problems. These include problems where the model and the updates—i.e., the model, the difference between the starting model and the true model and any update—permit sparse representations in some transformed domain.

This chapter is organized as follows. First, we introduce the Gauss-Newton method and how it can be extended to include an ℓ_1-norm constraint. Next, we compare this approach to the modified Gauss-Newton method where ℓ_1-norm constraints are imposed on the model updates. Afterwards, we compare results from the modified Gauss-Newton method with ℓ_1- or ℓ_2-norm constraints and without constraints in order to understand the importance of sparsity promotion. We, then, carry these experiments out for two-parameter problems so we can plot the objective, solution path, and constraints in a 2D plane in order to illustrate how these constraints factor into the optimization. Finally, to further validate the proposed method, we consider the notoriously difficult problem of phase retrieval (Bauschke et al., 2002), and two seismic examples, one of which is blind.

5.2 Optimization algorithms

Full-waveform inversion (FWI), like many other linear and nonlinear geophysical problems, involves the solution of an optimization problem. Depending on problem specifics and the prior knowledge, these optimization problems can take different forms, e.g. they can be constrained or unconstrained; first- or second-order. Without being all inclusive, we briefly introduce the types of optimization problems relevant to solving nonlinear sparsity-promoting inversion problems.

5.2.1 The unconstrained least-squares objective

FWI can be considered as an unconstrained least-squares (LS) minimization problem (Nocedal and Wright, 2006a)

$$\textbf{LS}: \qquad \min_{\mathbf{m}} \Phi(\mathbf{m}) := \left\{ \frac{1}{2} \|\mathbf{d} - \mathscr{F}[\mathbf{m}]\|_2^2 \right\},$$

where the vector $\mathbf{d}$ represents the observed data that we want to fit with the nonlinear forward modeling operator $\mathscr{F}$, parameterized by the discrete and vectorized model $\mathbf{m}$. We assume the source function to be known and fixed throughout this chapter.

FWI is challenging for the following reasons. First, evaluations of the forward modeling operator $\mathscr{F}[\mathbf{m}]$ are expensive because they involve the solution of large-scale PDE's (Helmholtz systems). Second, the forward map is nonlinear in the medium properties and solutions of the wave equation are oscillatory, which leads to multiple local minima related to cycle skipping when starting models are not close enough to the true model.

To reduce reliance of accurate starting models, several types of regularization have been proposed to include prior information (Hansen, 1998; Vogel and Oman, 1996; Abubakar and van den Berg, 2002). In this chapter, we limit ourselves to sparsity promoting priors that exploit structure on model updates with respect to the starting model. Before specializing our finding to FWI, we first introduce constrained and unconstrained formulations as well as our modified Gauss-Newton formulation for arbitrary forward modeling operators $\mathscr{F}$, which are assumed to be differentiable functions with respect to the model $\mathbf{m}$.

5.2.2 The ℓ_1-norm constrained least-squares objective

Motivated by sparsity exhibited by certain model updates—think for example of velocity perturbations in a FWI setting that are known to be compressible in the curvelet domain (Candès et al., 2006)—we would like to find solutions of **LS** that yield sparse updates with respect to the initial model $\mathbf{m}_0$. We can accomplish this by adding a sparsity constraint on the model update minimizing the least-squares objective (**LS**). We have

$$\mathbf{LS}\ell_1: \qquad \min_{\mathbf{x}} \Phi(\mathbf{x}) := \left\{ \frac{1}{2} \|\mathbf{d} - \mathscr{F}[\mathbf{S}^{\mathrm{H}}\mathbf{x}]\|_2^2 \right\} \quad \text{subject to} \quad \|\mathbf{x} - \mathbf{x}_0\|_{\ell_1} \leq \tau,$$

where $\mathbf{S}^{\mathrm{H}}$ is the inverse sparsifying transform. In this expression, the vector $\mathbf{x_0}$ denotes the transform domain coefficients of $\mathbf{m_0}$—i.e., $\mathbf{x}_0 = \mathbf{S}\mathbf{m}_0$ of the known starting model while $\mathbf{x}$ represent the synthesis coefficients that minimize the least-

squares objective subject to a ℓ_1-norm on the model update guaranteeing $\|\mathbf{x} - \mathbf{x}_0\|_{\ell_1} \leq \tau$. The choice for the sparsity level τ depends on the ℓ_1-norm of the model update (the difference between the true and starting models). For now, we will assume that this sparsity level is known. Unfortunately, in practice we cannot make this assumption and this requirement forms part of the motivation regularizing descent directions instead via ℓ_1-norm constraints.

5.2.3 Gauss-Newton for the unconstrained least-squares objective

There are numerous ways to solve (un)constrained optimization problems of the type **LS** and **LS**ℓ_1. Compared to first-order gradient based methods, second-order Gauss-Newton methods generally yield improved descent directions, converge faster, and are amenable to imposing structure on descent directions via ℓ_1-norm minimization. We arrive at the Gauss-Newton formulation by linearizing $\mathscr{F}$ within the $\|\cdot\|_2$norm brackets of **LS**. We compute the Gauss-Newton descent direction at the k^{th} by solving the following linear least-squares problem (Nocedal and Wright, 2006a):

$$\delta\mathbf{m}_k = \underset{\delta\mathbf{m}}{\arg\min} \|\delta\mathbf{d}_k - \nabla\mathscr{F}[\mathbf{m}_k]\delta\mathbf{m}\|_2^2. \tag{5.1}$$

In this expression, $\nabla\mathscr{F}[\mathbf{m}_k]$ represents the Jacobian evaluated at the model iterate of the k^{th} iteration $\mathbf{m}_k$. The vector $\delta\mathbf{d}_k = \mathbf{d} - \mathscr{F}[\mathbf{m}_k]$ contains the corresponding data residual. As outlined in Algorithm 5 where α is the step length, we repeat these iterations until the ℓ_2-norm of this residual is below some user-selected threshold ξ. While Gauss-Newton iterations are known to require fewer iterations compared to first-order gradient descent methods, incorporating second-order information comes at the price of having to solve least-squares problems (cf. Equation 5.1) for each model iterate. This can be problematic for large-scale problems, such as FWI, or for problems where the Jacobian has a null space—i.e., the Jacobian in Equation 5.1 is ill conditioned.

Output: Solution $\widetilde{\mathbf{m}}$ of the Gauss-Newton problem for starting model $\mathbf{m}_0$, tolerance ξ, and step length α.

1. $k = 0$
2. **while** $\|\mathbf{d} - \mathscr{F}[\mathbf{m}_k]\|_2^2 \geq \xi$ **do**
3. $\quad \delta\mathbf{m}_k = \arg\min_{\delta\mathbf{m}} \|\delta\mathbf{d}_k - \nabla\mathscr{F}[\mathbf{m}_k]\delta\mathbf{m}\|_2$ `// descent direction`
4. $\quad \mathbf{m}_{k+1} = \mathbf{m}_k + \alpha\delta\mathbf{m}_k$ `// model update`
5. $\quad k = k+1$
6. **end**
7. $\widetilde{\mathbf{m}} \longleftarrow \mathbf{m}_k$

Algorithm 5: Gauss-Newton method for unconstrained objective (**LS**).

5.2.4 Gauss-Newton method for the ℓ_1-norm constrained least-squares objective

Gauss-Newton methods can readily be extended to solve ℓ_1-norm constrained objectives (**LS**ℓ_1). In that case, the descent direction at the k^{th} iteration becomes

$$\delta\mathbf{m}_k = \mathbf{S}^{\mathrm{H}} \arg\min_{\delta\mathbf{x}} \|\delta\mathbf{d}_k - \nabla\mathscr{F}[\mathbf{m}_k]\mathbf{S}^{\mathrm{H}}\delta\mathbf{x}\|_2^2 \quad \text{subject to} \quad \|\delta\mathbf{x} + \mathbf{x}_k - \mathbf{x}_0\|_{\ell_1} \leq \tau. \tag{5.2}$$

In this constrained formulation, the Gauss-Newton subproblems optimize over the transform-domain coefficients and the descent direction at the k^{th} iteration is obtained by inverse transforming these coefficient via $\mathbf{S}^H$. Replacing the unconstrained least-squares problem on line 3 of Algorithm 5 by the constrained least-squares problem of Equation 5.2.

5.2.5 Modified Gauss-Newton method for unconstrained least-squares objective

Imposing ℓ_1 constraints on the descent directions $\delta\mathbf{m}_k$ themselves can lead to algorithms that are not only computationally efficient but that are also less sensitive to the sparsity level when following a scheme that carefully relaxes sparsity constraints on the descent directions. We introduced such an approach in the context of FWI, coined the modified Gauss-Newton method as outlined in Algorithm 6 below. Solutions of the ℓ_1-norm constrained Gauss-Newton subproblems of the type

$$\delta \mathbf{m}_k = \mathbf{S}^{\mathrm{H}} \arg\min_{\delta \mathbf{x}} \|\delta \mathbf{d}_k - \nabla \mathscr{F}[\mathbf{m}_k] \mathbf{S}^{\mathrm{H}} \delta \mathbf{x}\|_2^2 \quad \text{subject to} \quad \|\delta \mathbf{x}\|_{\ell_1} \leq \tau_k \qquad (5.3)$$

lie at the heart of this approach (Li and Herrmann, 2010a). Contrary to Equation 5.2 — where we impose a single ℓ_1-norm constraint on the transform-domain coefficients of the difference between the sum of the current model iterate and Gauss-Newton update at iteration k and the transform-domain coefficients of the starting model — we impose different sparsity constraints τ_k on the descent directions for each linearized Gauss-Newton subproblem. Since the Gauss-Newton subproblems are convex, we choose the τ_k for each subproblem (Equation 5.3) using a root-finding algorithm on the Pareto tradeoff curve (van den Berg and Friedlander, 2008b; Hennenfent et al., 2008b; Lin and Herrmann, 2013). For our purpose, it is sufficient to solve for each Gauss-Newton subproblem Equation 5.3 with the sparsity level set to $\tau_k = \frac{\|\delta \mathbf{d}_k\|_2}{\|\mathbf{S} \nabla \mathscr{F}^{\mathrm{T}}[\mathbf{m}_k] \delta \mathbf{d}\|_\infty}$ where $\|\cdot\|_\infty$ is the ℓ_∞, which corresponds taking the maximal value. This value for the sparsity level corresponds to the first τ selected by SPGℓ_1 (van den Berg and Friedlander, 2008b; Hennenfent et al., 2008b; Lin and Herrmann, 2013) and is a very conservative value for the sparsity level on the transform coefficients of the descent directions. Remark that after each iteration we update the model, which leads to a new Gauss-Newton subproblem. As we will show below, this empirical strategy can be applied successfully to nonlinear (non-convex) problems and that for linear problems this strategy is equivalent to the root-finding method undergirding SPGℓ_1. We will also demonstrate that the above choice of sparsity levels for the Gauss-Newton subproblems does not require detailed information on the ℓ_1-norm of the transform coefficients of the difference between the starting and true models. Instead, the algorithm needs as conservative estimate of the ℓ_1-norm for the updates. As long as the sparsity levels are bounded, the ℓ_1-norm constrained descent direction remain descent directions and the algorithm provably converges (Burke, 1992). In Algorithm 6, we summarize the details of the modified Gauss-Newton method.

Output: Solution $\widetilde{\mathbf{m}}$ of the modified Gauss-Newton problem for starting model $\mathbf{m}_0$, tolerance ξ, and step length α.

1. $k = 0$
2. **while** $\|\mathbf{d} - \mathscr{F}[\mathbf{m}]\|_2^2 \geq \xi$ **do**
3. $\quad \tau_k = \|\delta\mathbf{d}_k\|_2 / \|\mathbf{S}\nabla\mathscr{F}^{\mathrm{H}}[\mathbf{m}_k]\delta\mathbf{d}_k\|_\infty$
4. $\quad \delta\mathbf{m} = \arg\min_{\delta\mathbf{x}} \|\delta\mathbf{d} - \nabla\mathscr{F}[\mathbf{m}_k]\mathbf{S}^{\mathrm{H}}\delta\mathbf{x}\|_2^2 \quad$ subject to $\quad \|\delta\mathbf{x}\|_{\ell_1} \leq \tau_k$ `// Gauss-Newton update`
5. $\quad \mathbf{m}_{k+1} = \mathbf{m}_k + \alpha\mathbf{S}^{\mathbf{H}}\delta\mathbf{x}_k$ `// update with linesearch`
6. $\quad k = k + 1$
7. **end**

Algorithm 6: Modified Gauss-Newton method with sparse update for unconstrained LS objective function.

5.3 Comparisons on stylized two-parameter examples

Our main goal is to provide a justification for our modified Gauss-Newton method for a particular class of problems where the difference between the starting model and true is sparse in some transformed domains. For this purpose, we will conduct a series of stylized examples designed to demonstrate the superior performance of our method compared formulations that either do not exploit sparsity—i.e., **LS** and Equation 5.1, or do exploit sparsity by either constraining the model as in **LS**ℓ_1 and Equation 5.2 or descent directions as in Equation 5.3. We conduct our study to substantiate our claim that for conservative chosen sparsity levels, the modified Gauss-Newton method yields for particular problems sparse solutions without requiring prior knowledge on the sparsity level. Our stylized examples are divided into convex problems where local minima is global minima coincide and non-convex problems that may have local minima. For convex problems, we show that Algorithm 6 converges to the solution while Gauss-Newton applied to ℓ_1-norm constrained objective function (Algorithm 5 with line 3 defined by Equation 5.2) will only converge to the true solution if it is inside the constraint set. For non-convex problems, the modified Gauss-Newton method is more likely to converge to a global or local minimum as long as the difference between the starting and true model permits a sparse representation that is a sparse perturbation of the initial model.

5.3.1 Solution paths of a determined convex problem with a unique solution

To get a better understanding of the behavior of the above optimization methods, we compare solution paths for Algorithm 5 with and without ℓ_1-norm constraints for correct and wrong values of the sparsity levels on a simple determined two-dimensional problem. We also do this for the modified Gauss-Newton method outlined in Algorithm 6. By setting $\mathscr{F}[\mathbf{m}] \stackrel{\text{def}}{=} \mathbf{Am}$ with $\mathbf{A} = \begin{bmatrix} 2 & 4 \\ 6 & -3 \end{bmatrix}$ we arrive at a linear convex problem that has a unique (denoted by the green dot in Figure 5.1) global solution for the two model parameters $\mathbf{m} = \begin{bmatrix} m_1 \\ m_2 \end{bmatrix}$ (Local minima correspond to global minima for convex problems, and Jacobian of this problem is $\mathbf{A}$). For the data given by $\mathbf{d} = \begin{bmatrix} -6 \\ -3 \end{bmatrix}$ the solution equals $\mathbf{m}_{\text{true}} = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$. As we can see from Figure 5.1, this solution corresponds to the global minimum of the least-squares objective as a function of the two model parameters. As expected, solutions paths for the unconstrained (Algorithm 5) and correct ℓ_1-norm constrained (Algorithm 5 with line 3 replaced by Equation 5.2) Gauss-Newton methods both arrive at the correct global minimum, and therefore, yield the correct solution. Irrespective of the starting model, we can expect this behavior as long as the restriction to the "ℓ_1-norm ball" — i.e., the diamond-shaped constraint set denoted by the green dashed line in Figure 5.1b — includes the global minimum. However, if we choose a sparsity level so small that it no longer includes the global minimum, the constrained formulation proceeds, as illustrated in Figure 5.1c, to the wrong solution. Instead of finding the correct global minimum, the algorithm converges to a solution that minimizes the least-squares objective while meeting the ℓ_1-norm constraint. This example clearly shows the potential danger of including ℓ_1-norm or other constraints. The global minimum needs to be within the constraint set in order to find the correct solutions.

Results from the modified Gauss-Newton method, on the other hand, arrive at the correct global minimum irrespective of the type of norm (diamond-shaped ℓ_1-norm ball as in Figure 5.1d or the circular-shaped ℓ_2-norm ball as in Figure 5.1e). This behavior, where the descent directions are constrained, is consistent with the-

oretical findings of Burke (1992), who states that Algorithm 6 converges despite the fact that we impose constraints on the search directions. For illustrative purposes, we imposed the ℓ_2-norm as well by simply replacing line 3 in Algorithm 6 by $\tau_k = \|\delta \mathbf{d}_k\|_2 / \|\mathbf{A}^{\mathrm{T}} \delta \mathbf{d}_k\|_2$ and the ℓ_1-norm in line 4 by the ℓ_2-norm. While this example clearly shows that global minima can still be found when imposing constraints on the updates, it does not demonstrate the added value of these constraints for more challenging problems that do not have a unique solution.

5.3.2 Solutions paths of an underdetermined convex problem with multiple solutions

To further study the behavior of the above listed optimization problem, we conduct the same experiments for the underdetermined case where $\mathbf{A} = \begin{bmatrix} 2 & 4 \end{bmatrix}$ and $\mathbf{d} = -4$. Without imposing prior knowledge, this problem has infinitely many solutions. Again, we compare the performance of the different optimization problems by plotting the least-squares objective in color code, the solutions minimizing the least-squares objective that lie on the line $-4 + 2m_1 + 4m_2 = 0$, and the solutions paths in Figure 5.2 from starting models located at $\begin{bmatrix} 3 \\ 4 \end{bmatrix}$.

Since our problem currently does not have a unique solution, we will impose sparsity on the solution. This means that we are looking for solutions that align with the principle axes in which case one of the two model parameters is zero. As expected, the solution path that minimizes the least-squares objective only misses the sparse minimum at $\begin{bmatrix} 3 \\ -0.5 \end{bmatrix}$ denoted by the green dot. Similarly, the ℓ_1-norm constrained formulations also fail to find the correct minimum in cases where the sparsity level τ is too high (Figure 5.2b), in which case the solution corresponds to the unconstrained least-squares solution, or too small (Figure 5.2c) in which case the solution is sparse but wrong in amplitude. As long as we know the exact sparsity level in advance, we can inflate the ℓ_1 ball so one of its corners touches the least-squares objective (see Figure 5.2d), which yields a sparse solution where one of the two model parameters is zero.

As illustrated in Figure 5.2e, solution paths for our modified Gauss-Newton method with ℓ_1-norm constraints on the descent directions also find the sparse so-

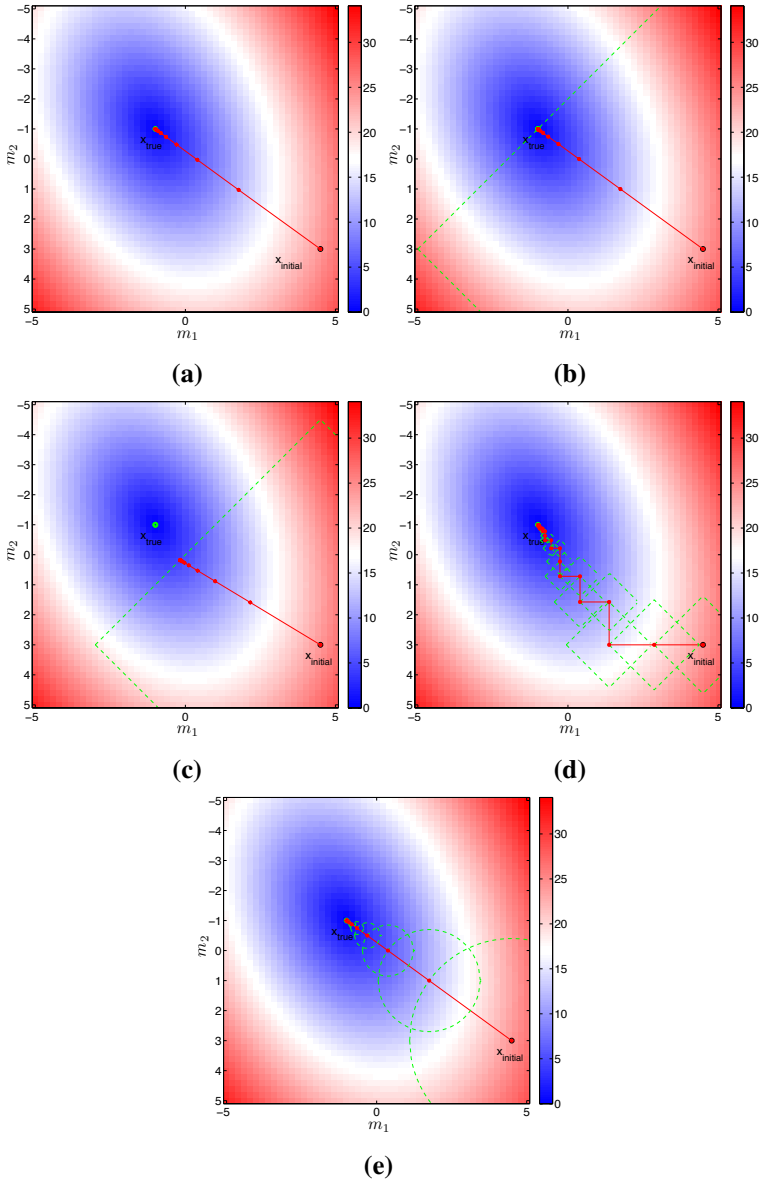


Figure 5.1: Solution path of different methods for convex problem that has a unique solution: (a) Algorithm 5; (b) Algorithm 5 with line 3 defined by Equation 5.2 (correct ℓ_1 constraint); (c) same as Figure 5.1b but with wrong ℓ_1 constraint; (d) Algorithm 6 but with ℓ_1 constraint on the updates; (d) Algorithm 6 but with ℓ_2 constraint.

lution without having prior knowledge on the true sparsity level. This example illustrates that we can find sparse solutions by imposing ℓ_1-norm constraints on the model updates according to Algorithm 6. Figure 5.2f shows the ℓ_1-norm plays a crucial role since constraining the ℓ_2-norm on the updates does not lead to a sparse solution. Contrary to the ℓ_2-norm constrained descent directions, descent directions constrained by the ℓ_1-norm are all sparse and have the same locations for the significant entries while moving to the sparse solution as shown in Figure 5.2e.

The above observations for convex problems with ℓ_1-norm constraints provide an intuitive explanation why imposing ℓ_1-norm constraints on updates may make sense in certain circumstances. This comes not as a surprise because the modified Gauss-Newton for these problems is derived from the same principle as the SPGℓ_1 (van den Berg and Friedlander, 2008b; Hennenfent et al., 2008b) solver, which solves sparsity-promoting problems by solving a number of relaxed ℓ_1-norm constrained least-squares problems. Since the descent directions apparently share the same sparse support, we argue that for certain problems our modified Gauss-Newton approach exhibits the same behavior for certain non-convex problems.

5.3.3 Solutions paths for undetermined non-convex problems with multiple solutions

Many inverse problems in geophysics are nonlinear, and therefore, non-convex. Unfortunately, FWI is no exception. For our stylized two-parameter problem this means that the minima for the least-squares objective no longer lie on a straight line, but instead, on an arbitrary curve. While it is still relatively straightforward to find a minimum, by using classical derivative-based optimization methods (Ruszczyński, 2006), finding the global minimum is far more difficult (Horst et al., 2000), due to the existence of local minima. To make matters worse, imposing sparsity as prior information via ℓ_1-norm constraints, an approach we used successfully to solve underdetermined convex problems, is also jeopardized since it may no longer be likely that the ℓ_1-norm ball touches the least-squares objective at its corners, and therefore, it would no longer yield a sparse solution for the update with respect to the starting model—i.e., the starting vector for the model parameters.

To illustrate this phenomenon, we consider a nonlinear quadratic problem by

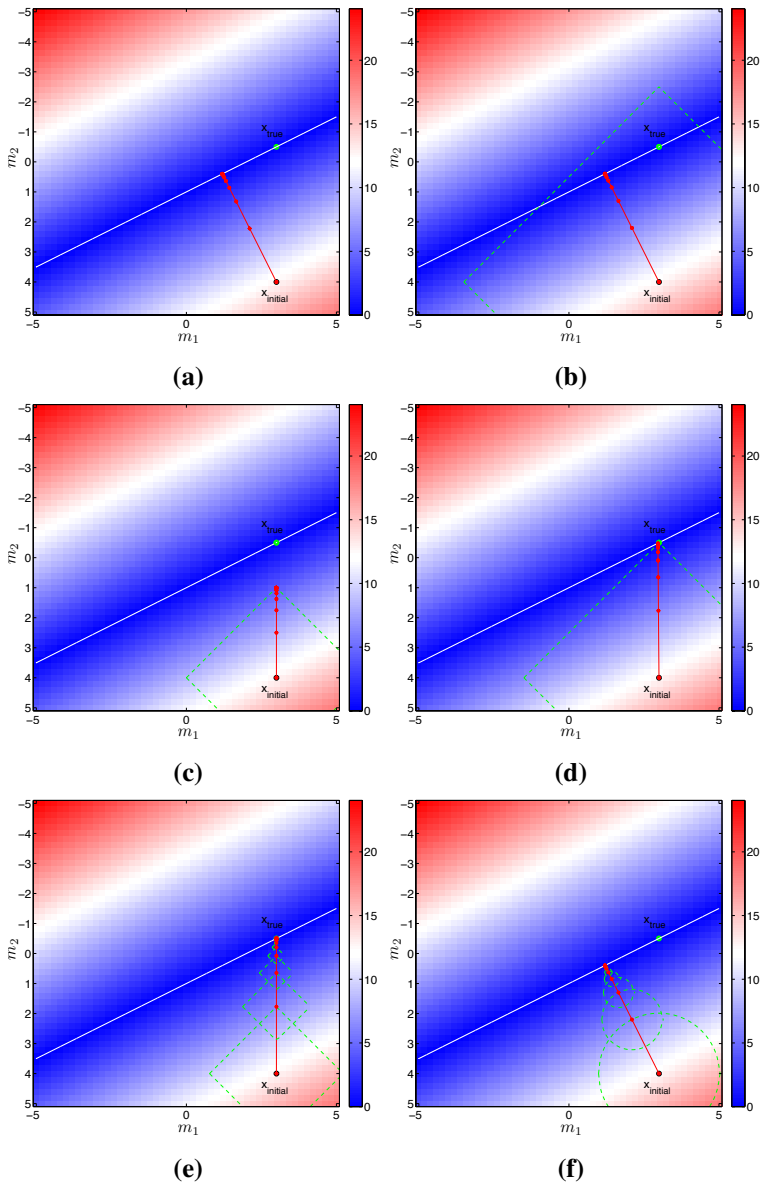


Figure 5.2: Solution path of different methods for a linear problem that has multiple solutions: (a) Algorithm 5; (b) Algorithm 5 with line 3 defined by Equation 5.2 with wrong constraint ($\tau > \tau_{true}$); (c) same as Figure 5.2b with wrong constraint ($\tau < \tau_{true}$) ; (d) same as Figure 5.2b but with the right constraint ($\tau = \tau_{true}$); (e) Algorithm 6; (f) Algorithm 6 but with ℓ_2 constraint on the updates.

setting $\mathscr{F}[\mathbf{m}] \stackrel{\text{def}}{=} (\mathbf{Am})^T\mathbf{Am}$ with $\nabla\mathscr{F}[\mathbf{m}] = (\mathbf{Am})^T$. In this formulation, the data is no longer linear but quadratic in the model parameters. As we can see from Figure 5.3, this problem yields non-unique minima for the least-squares objective that lie on an ellipse (denoted by the white line) given by the following expression $4m_1^2 + m_2^2 = 4$ for $\mathbf{A} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$ and $\mathbf{d} = -4$. As before, we conduct our experiments comparing the different optimization formulations and plot the solutions paths for the unconstrained (Figure 5.3a); constrained with a ℓ_1-norm constraint τ that is too large (Figure 5.3b, $\tau > \tau_{\text{true}}$); correct (Figure 5.3c, $\tau = \tau_{\text{true}}$) or gradually relaxed (Figure 5.3d) from a small τ to a large τ. In all cases (Figures 5.3a—5.3c), no sparse solutions (denoted by the green dot) are found because of the curvature of the ellipsoid delineating the minimum of least-squares objective for this quadratic optimization problem. This example illustrates possible limitations of ℓ_1-norm constraints when solving non-convex problems even in cases where the sparsity level is known. This observation was also recently made in the literature (see Van Den Doel et al., 2012).

However, this is not the end of the story as we can see when we closely inspect the solution path yielded by the modified Gauss-Newton method with ℓ_1-norm constraints. In that case (Figure 5.3e), Algorithm 6 gives rise to descent directions that continue to make progress towards the sparse minimum until the solutions of the Gauss-Newton subproblems bend upwards to hit the least-squares objective. The ℓ_1-norm constraints on the descent directions are responsible for this behavior because we do not observe the same behavior when we impose ℓ_2-norm constraints instead (juxtapose Figure 5.3e and 5.3f). We explain the relative success of the modified Gauss-Newton compared to imposing (the correct) ℓ_1-norm constraint on the model (Figure 5.3c by virtue of the fact that the ℓ_1-norm balls for our modified Gauss-Newton method are smaller—because the ℓ_1-norm of the model updates goes to zero as the algorithm converges to the minimum—and consequently this method may be less sensitive to the curvature of the least-squares constraint (the ellipsoid in this case) as the solution approaches the minimum of the objective. Since the problem permits a sparse solution (denoted by the green dot), this phenomenon provides an explanation why the modified Gauss-Newton method finds descent directions that are sparse while sharing approximately the same non zeros—i.e.,

support. As a consequence of sharing the same support, we would expect sparse, or at least relatively close to sparse, solutions that only become "dense" as we approach the minimum of the quadratic objective. This is, arguably, good enough because in most instances we cannot afford enough iterations to bring us close to a minimum due to the size and numerical complexity of geophysical problems.

Obviously, these results are encouraging for the following two reasons. First, we proposed an algorithm that yields sparse solutions as long as the sparsified descent directions have approximately the same support—i.e., have approximately the same locations for the non-zeros. The question now is what type of problems exhibit this type of behavior. Second, the proposed modified Gauss-Newton method seems to be less sensitive to the curvature (read conditioning of the Hessian, Van Den Doel et al. (2012)) and does not require prior information on the sparsity level. Before we apply the modified Gauss-Newton method to realistic (blind) FWI problems, let us first examine its performance on a larger scale nonlinear problem.

5.3.4 Application to the "phase-retrieval" problem

With some intuition built from the stylized two-parameter convex and non-convex problems of the previous section, we now study the performance of the modified Gauss-Newton method on a more challenging non-convex underdetermined optimization problem, referred to as the "phase retrieval" problem:

$$\textbf{Phase}: \qquad \min_{\mathbf{m}} \Phi(\mathbf{m}) := \left\{ \frac{1}{2} \|\mathbf{d} - \text{diag}(\mathbf{Am})(\mathbf{Am})\|_2^2 \right\},$$

where we choose $\mathbf{A}$ to be a slightly underdetermined 400×512 random Gaussian matrix. Given this choice for $\mathbf{A}$, $\mathscr{F}[\mathbf{m}] \overset{\text{def}}{=} \text{diag}(\mathbf{Am})(\mathbf{Am})$ and $\nabla\mathscr{F}[\mathbf{m}] = \text{diag}(\mathbf{Am})\mathbf{A}$ our task is to recover the model parameters from data collected in the vector $\mathbf{d} = \mathscr{F}[\mathbf{m}_{\text{true}}]$. For simplicity, we will assume the data to be noise free and our goal is to recover the vector $\widetilde{\mathbf{m}}$ from the square of the entries yielded by applying the slightly underdetermined system $\mathbf{A}$ to the true model vector.

While seemingly harmless, this type of non-convex optimization problem is because of the nonlinearity of the forward operator $\mathscr{F}[\mathbf{m}]$ notoriously difficult to

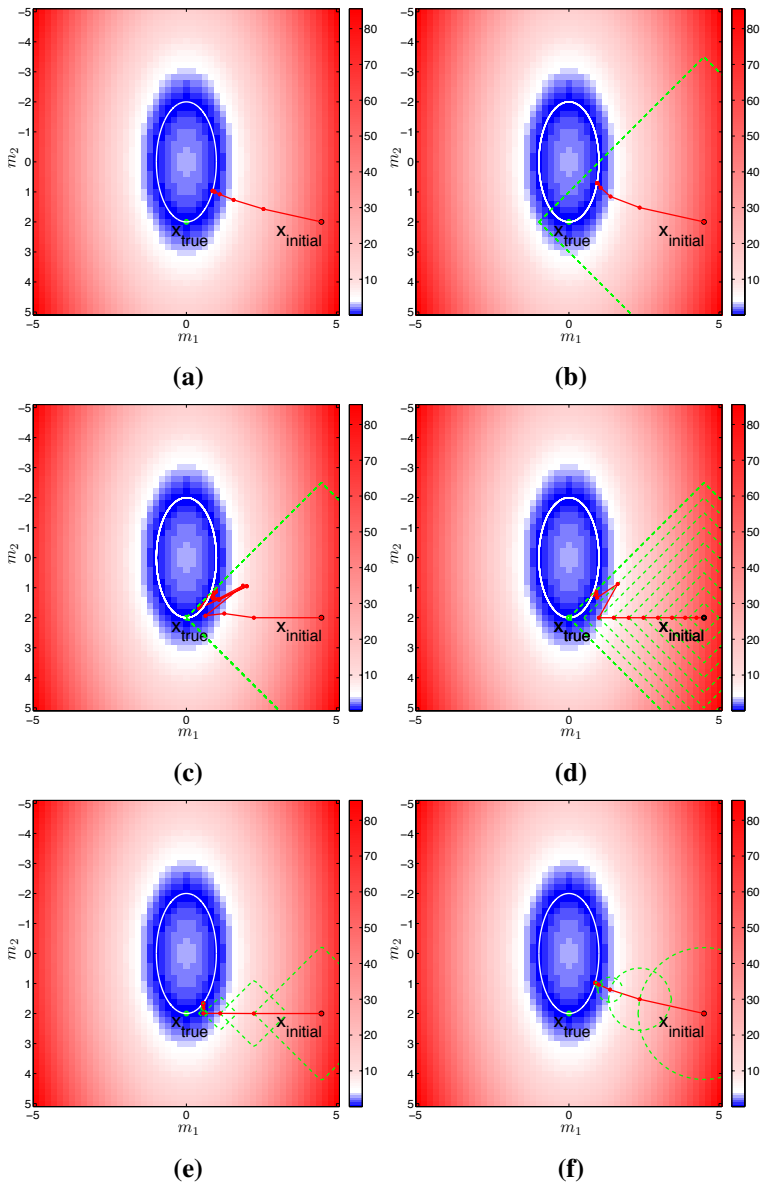


Figure 5.3: Solution path of different methods for a nonlinear problem that multiple solutions: (a) Algorithm 5; (b) Algorithm 5 with line 3 defined by Equation 5.2 with wrong constraint ($\tau > \tau_{true}$); (c) same as Figure 5.3b but with the right constraint; (d) same as Figure 5.3b but with the gradually relaxed constraint; (e) Algorithm 6; (f) Algorithm 6 with ℓ_2 constraint on the updates.

solve without prior knowledge on $\mathbf{m}$. However, if we choose the model vector and starting models in such as way that their difference is sparse (see Figure 5.4), e.g. by choosing a smooth (background) model and sparse-spike model permutations, the above optimization problem may become easier to solve with ℓ_1-norm constraints.

Figures 5.4b and 5.4c contain results of applying the unconstrained **LS** and constrained formulations **LS**ℓ_1 to this "phase-retrieval" problem with a correct starting model and a correct sparsity level for the spiky perturbations. From these plots, it is clear that both formulations are not able to recover the sparse spike train despite accurate knowledge on the starting model and sparsity level of the spiky perturbations. The result for **LS** is noisy because of the random "crosstalk" generated by the Gaussian measurement matrix $\mathbf{A}$. Adding a ℓ_1-norm constraint to the least-squares objective removes this interference noise but yields the wrong sparse solution. The modified Gauss-Newton method, on the other hand, is capable of accurately resolving both the locations and magnitudes of the spikes.

The reason for the superior performance of the modified Gauss-Newton method is twofold. First, the modified Gauss-Newton method exploits the sparse structure of the perturbations. While this may seem unrealistic but as we will demonstrate below this sparsity assumption is valid for FWI. Second, and more importantly, the ℓ_1-norm constrained descent directions of the modified Gauss-Newton method share a substantial fraction of their support from iteration to iteration, yielding a solution that preserves the smooth component and accurately recovers the sparse difference. We illustrate this behavior in Figure 5.5, where we plot the support (locations of the non zeros) for the ℓ_1-norm constrained Gauss-Newton search directions. Compared to the locations of the true spikes, plotted in red at the bottom of Figure 5.5, the sparsity patterns of the Gauss-Newton descent direction remain sparse with non-zero patterns that are coincident with the true support in red. This observed behavior partly explains the successful recovery of the modified Gauss-Newton method in a situation where the other two methods failed. As we can see from Figure 5.4e promoting sparsity via the ℓ_1-norm again plays a crucial role because modified Gauss-Newton with ℓ_2-norm constraints fails. Also relaxing the ℓ_1-norm constraint added to the ℓ_2-norm objective does not result in the correct sparse solution (juxtapose Figures 5.4c and 5.4f).

Our observations are also confirmed by plots for the relative ℓ_2 norms for the

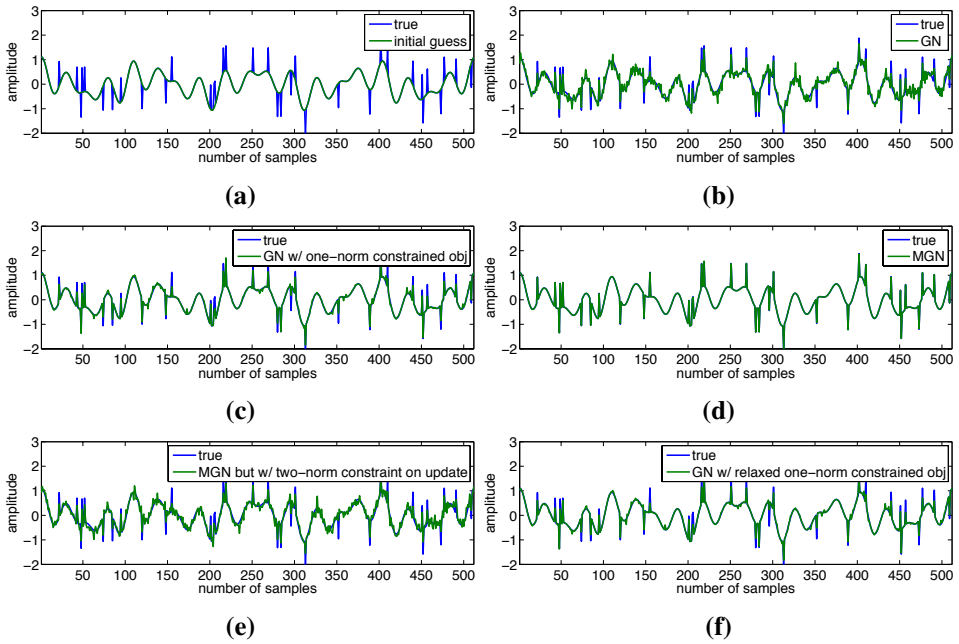


Figure 5.4: Results for phase retrieval example from difference methods: (a) true solution and initial guess; (b) solution of Algorithm 5; (c) solution of Algorithm 5 with line 3 defined by Equation 5.2; (d) solution of Algorithm 6; (e) solution of Algorithm 6 but with ℓ_2 constraint on the updates; (f) same as Figure 5.4c but with gradually relaxed ℓ_1 constrained objective function;

residuals and model errors as a function of the number of Gauss-Newton iterations included in Figure 5.6. These plots clearly show that fitting the data with accurate knowledge on the sparsity level by itself is not sufficient. While prior knowledge on the sparsity level helps, progress to the true solution stalls for **LS** and **LS**ℓ_1 because both algorithms get trapped in local minima. Conversely, the modified Gauss-Newton method of Algorithm 6 continues to make progress towards the solution bringing both the relative data residual and model errors down as Algorithm 6 progresses.

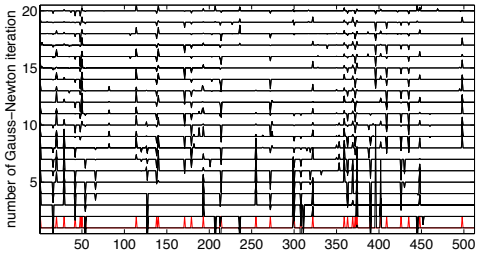
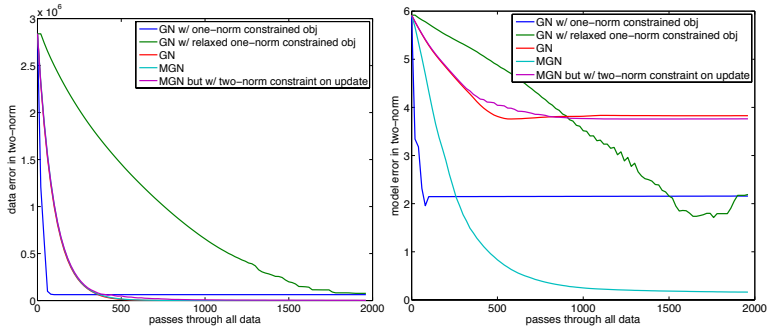


Figure 5.5: All modified Gauss-Newton updates for the phase retrieval problem.



(a) **(b)**

Figure 5.6: Data misfit and relative model error for phase retrieval example: (a) data misfit; (b) relative model error.

5.4 Application to FWI

In the previous section, we were able to demonstrate that under certain conditions, the modified Gauss-Newton method can lead to accurate results. We will now argue that this method can also perform well on FWI for which the method was originally developed (Li et al., 2012). Before we apply this method to two realistic synthetic examples, let us first briefly present the original randomized formulation for the modified Gauss-Newton method, followed by a brief motivation why we expect this approach to perform well given our findings so far.

5.4.1 Randomized modified Gauss-Newton

FWI is extremely challenging for several reasons including the problem size and sensitivity to cycle skipping. In earlier work (Li et al., 2012), we demonstrated that the excessive demands on computational resources of multi-experiment FWI can be overcome by working with small randomized subsets of the data (read small numbers of randomized composite shots or randomly selected shots) during the modified Gauss-Newton iterations. As we have shown in the past, working with randomized subsets of shots and angular frequencies turns the now dimensionality-reduced Gauss-Newton subproblems into underdetermined problems that give rise to sub-sampling related artifacts as we already observed in Figure 5.4b.

In the FWI setting, we remove these sub-sampling related artifacts by promoting curvelet-domain sparsity on the descent directions by solving for the k^{th} Gauss-Newton iteration the following ℓ_1-norm constrained optimization problem (Li et al., 2012):

$$\delta\mathbf{m}_k = \mathbf{S}^H \arg\min_{\delta\mathbf{x}} \|\delta\underline{D}_k - \nabla\mathscr{F}[\mathbf{m}_k, \underline{Q}_k]\mathbf{S}^H\delta\mathbf{x}\|_F^2 \quad \text{subject to} \quad \|\delta\mathbf{x}\|_{\ell_1} \leq \tau_k. \quad (5.4)$$

In this expression, the optimization is carried out over the curvelet coefficients, which are brought back to the physical domain via the inverse curvelet transform, given by the adjoint, denoted by the symbol ^{H}, of the forward curvelet transform $\mathbf{S}$. At each iteration, the descent directions themselves are calculated over randomly selected frequencies in overlapping windows for data residuals (for each shot record in the columns) $\{\delta\underline{D}_k = \underline{D} - \mathscr{F}[\mathbf{m}_k, \underline{Q}_k]\}$ and sources $\underline{Q}_k$ that are also randomly selected. We denote these randomly sub-sampled quantities by the underbar. To accelerate the convergence, we select independent subsets for each modified Gauss-Newton problem, which are only solved approximately. For completeness, we included this randomized modified Gauss-Newton method in Algorithm 7. For a detailed description of this algorithm, we refer to the literature (Li et al., 2012).

We now evaluate this algorithm on two different synthetic models, namely the BG North Sea COMPASS model and the blind Chevron Gulf of Mexico (GOM) model. Both models are generated with real geological information but reflect completely different geological settings. The BG model was designed to evaluate

FWI potential ability to resolve fine reservoir-scale variations in the rock properties, which would make it ideal for our modified Gauss-Newton method because this model can be well approximated by only one percent of the largest curvelet coefficients. Since we have access to the true model, this example will allow us to quantitatively compare the different algorithms. The blind Gulf of Mexico example, on the other hand, is much more challenging. It is very deep, well beyond the penetration depth of turning waves from which FWI normally reaps its information; it is noisy and contains challenging high-velocity salt bodies that are difficult to delineate.

Output: Solution $\widetilde{\mathbf{m}}$ of the randomized modified Gauss-Newton problem for starting model $\mathbf{m}_0$, tolerance ξ, and step length α.

1. $\widetilde{\mathbf{m}} \longleftarrow \mathbf{m_0}$, and ξ `// initial guess and expected residual`
2. **while** $\|\delta\underline{D}_k\|_2 \geq \xi$ **do**
4. $\tau_k = \|\delta\underline{D}_k\|_F / \|\mathbf{S}\nabla\mathscr{F}^{\mathrm{H}}[\mathbf{m}_k, \underline{Q}_k]\delta\underline{D}_k\|_\infty$
5. Solve Equation 5.4
6. $\mathbf{m}_{k+1} = \mathbf{m}_k + \alpha\mathbf{S}^{\mathbf{H}}\delta\mathbf{x}_k$ `// update with linesearch`
7. **end**

Algorithm 7: Modified Gauss-Newton with curvelet-domain sparsity promotion and randomization.

5.4.2 BG COMPASS model

The BG COMPASS model (Figure 5.7a) contains a large amount of variability constrained by well data. We use this velocity model to generate synthetic data by running a time-domain finite-difference code with a 15 Hz Ricker wavelet. In total, we simulated 350 shots with 20m shot intervals; all shots share the same 700 receiver positions with 10m receiver intervals, yielding a maximum offset of 7km.

All inversions based on our frequency-domain methods start from 5 Hz with a heavily smoothed velocity model without lateral variations (Figure 5.7b). To avoid local minima, the inversions are carried out in 8 increasing sequential but overlapping frequency bands on the interval $5-15$ Hz (Bunks et al., 1995), each using 20 different randomly selected simultaneous shots and 3 random selected frequencies. We use 10 modified Gauss-Newton iterations (Equation 5.3) for each

frequency band. For each modified Gauss-Newton subproblem, we use roughly 10 iterations of SPGℓ_1 (van den Berg and Friedlander, 2008b). Figure 5.7 contains our results for the inverted velocity by the different algorithms, namely unconstrained **LS** is plotted in Figure 5.7c; constrained **LS**ℓ_1 with correct sparsity constraint level as in Figure 5.7e and wrong sparsity constraint level as in Figure 5.7d; our modified Gauss-Newton method is plotted in Figure 5.7f.

From these examples, the following observations can be made. First, without sparsity-promoting constraints, the inversions fail to recover the velocity model using a relatively small number of randomly selected shots and frequencies. Consequently, we are able to significantly speedup the inversion, which is consistent with our observations reported in the literature (Li et al., 2012). Second, imposing sparsity as additional constraint for the least-square objective does not give a satisfying inversion result, even when we use the correct ℓ_1-norm constraint. Again, imposing ℓ_1-norm constraints on the updates yields the best results as plotted in Figure 5.7f. As before, replacing the ℓ_1-norm constraint by a ℓ_2-norm constraint leads to noisy and inferior results (Figure 5.7g). These observations are also reflected in the behavior of the relative data error (plotted in Figure 5.8a) where the modified Gauss-Newton method is the most successful in bringing the relative residual down. The behavior of the relative model errors (Figure 5.8b) paints an even more drastic picture where all but the result from the modified Gauss-Newton have relative model errors that are not only inferior but also diverge after a certain number of iterations. While the iteration-to-iteration percentage of overlapping curvelet coefficients decreases somewhat, more than 50 % of the support of the curvelet coefficient overlap, explaining that the final result remains sparse in the curvelet domain, as shown in Figure 5.9.

5.4.3 Blind Gulf of Mexico example

Aside from accelerating FWI, where each Gauss-Newton subproblem can be considered as a compressive-sensing type of recovery problem, the constrained updates can also be considered as curvelet-domain "denoised" model updates. The "noise" in this case refers to subsampling artifacts related to the acceleration and to unmodeled components in the data. The latter include elastic (converted) energy but

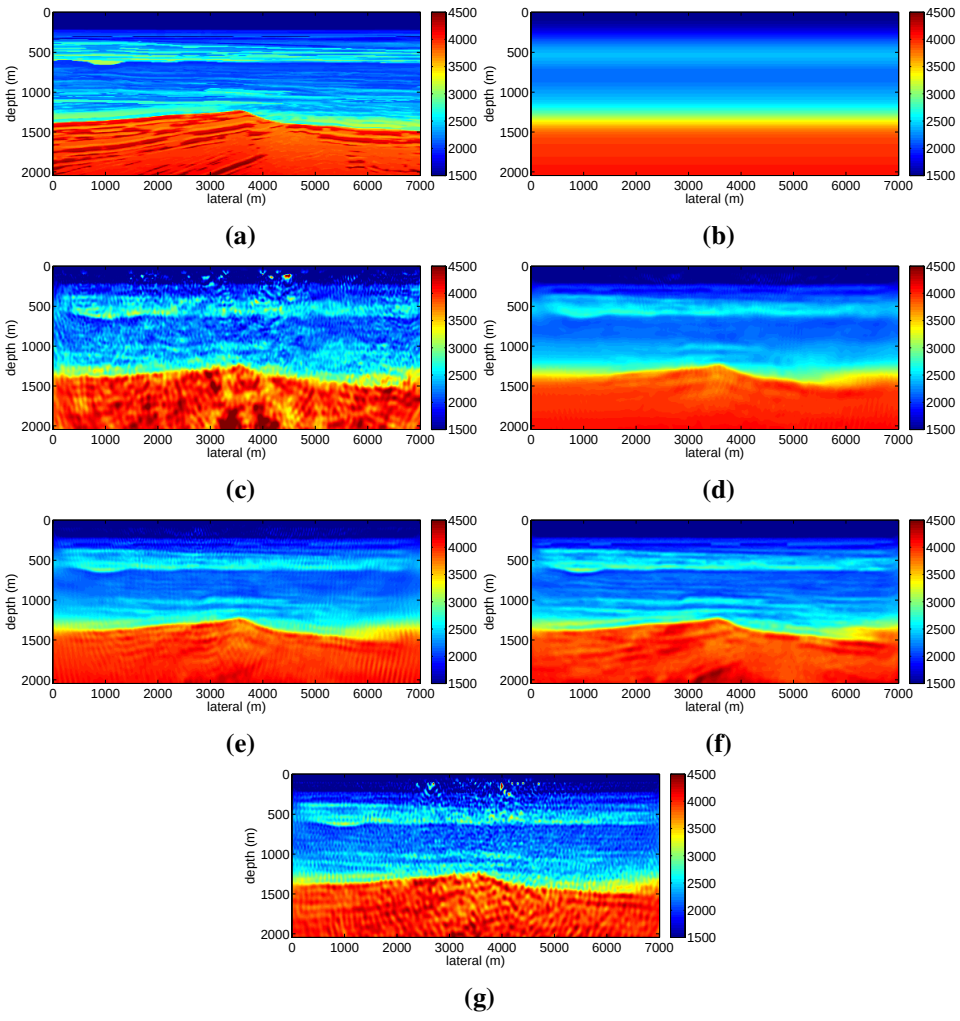
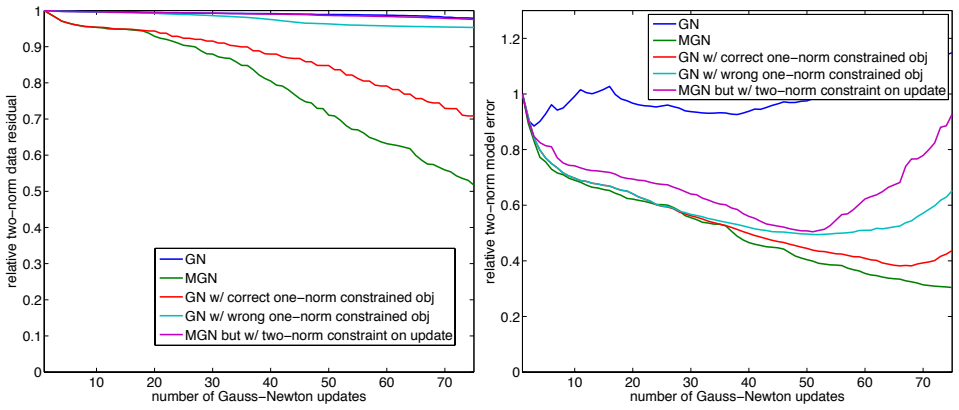


Figure 5.7: FWI result from BG COMPASS model data set: (a) true model used to generate observed data; (b) starting model for FWI; (c) Gauss-Newton result with unconstrained objective function; (d) Gauss-Newton result with incorrect ℓ_1 constrained objective function; ($\tau < \tau_{true}$); (e) same as Figure 5.7d but with correct ℓ_1 constrained objective function; ($\tau = \tau_{true}$); (f) modified Gauss-Newton result with ℓ_1 constraint on the updates; (g) same as Figure 5.7f but with ℓ_2 constraint on the updates.



(a) **(b)**

Figure 5.8: Data misfit and relative model error for BG model example: (a) data fitting residual; (b) relative model error.

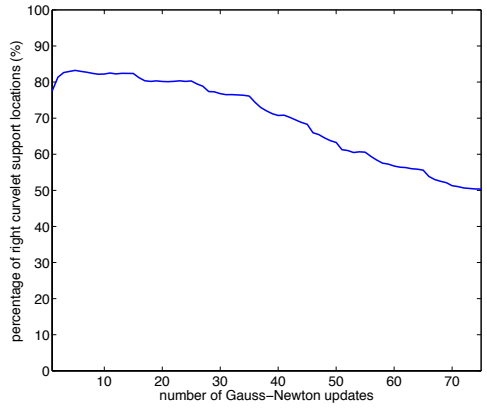


Figure 5.9: Percentage of curvelet coefficients that are at the right support positions.

also, to some extend reflection events that normally would have been muted during the FWI workflow.

The results for different inversions, using the starting model plotted in Figure 5.10a, are included in Figure 5.10b for unconstrained Gauss-Newton; in Figure 5.10c for modified Gauss-Newton with ℓ_2; and in Figure 5.10d for modified Gauss-Newton with ℓ_1. As described in (Herrmann et al., 2013), the starting model was obtained by carrying out ray-based travel-time tomography yielding

a root-mean-square traveltime misfit of only 11ms. To handle low-frequency noise in the data, consisting of 3201 shots with a 25m shot interval, we performed curvelt-domain denoising on selected monochromatic frequency slices (Kumar, 2009; Hennenfent and Herrmann, 2006b) on the interval 2-5Hz in the source-offset domain. The receiver spacing was 25m and the maximal offset of this streamer data set 20km.

In addition to the bad signal-to-noise ratio at the low frequencies, this data set is extremely challenging because of the limited offset, presence of complex high-velocity salt bodies, and large depth. As we can see from the inversion results, this combination exposes certain shortcomings of FWI to recover the deeper portions of the model, the top of the salt and complexity within the salt. Despite these shortcomings, the inversion results in Figure 5.10 are important for the following reasons. First, the results are obtained automatically without human intervention by running Algorithm 7 for 25 iterations and using 600 randomly selected shots for each MGN iteration. This means that these results are reproducible. Second, the unmodeled "noise" leads to major artifacts if we do not impose constraints on the Gauss-Newton updates as we can observe from Figure 5.10a. One the other hand, imposing constraints on the norm of the curvelet coefficients of the model updates improves the inversion results (Figure 5.10d). As before, promoting curvelet-domain sparsity via the ℓ_1 performs the best. This can be understood because this sparsity constraint acts as denoiser during which only the largest, and therefore most significant, curvelet coefficients are allowed into the model updates. This prevents overfitting of components that do not lie in the range of the forward modeling operator. While it is clear that standard FWI is unable to handle this type of data, comparison between the data misfit for the starting and final models shows that certain phases in the data that were originally cycle skipper have a better fit as we can see in Figure 5.11.

5.5 Conclusion

Full-waveform inversion is challenging due to the fact that it is computationally expensive, and also, requires accurate starting models and modeling engines. Our main contribution has been to demonstrate how to reduce computational cost while

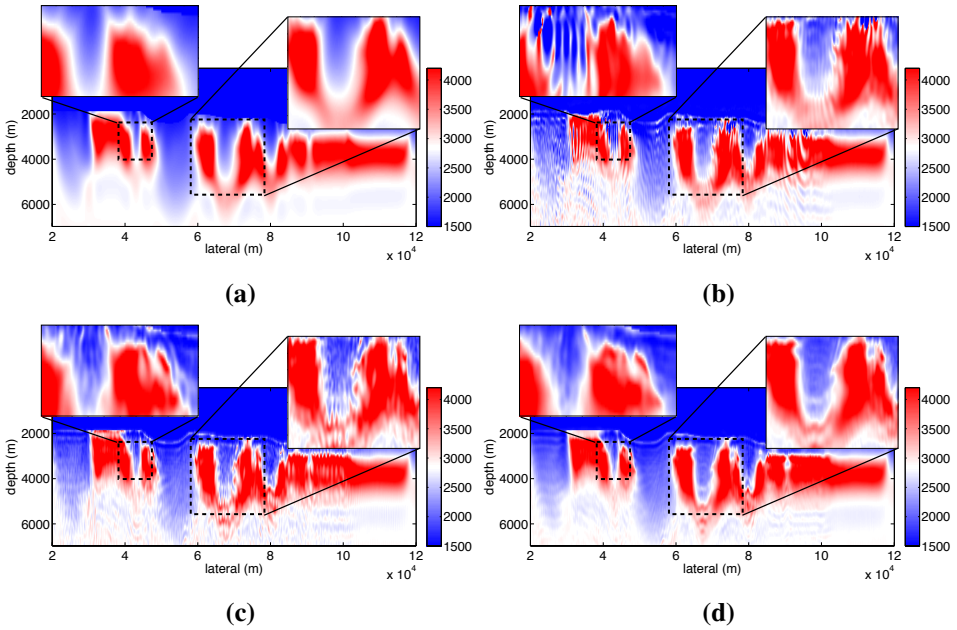
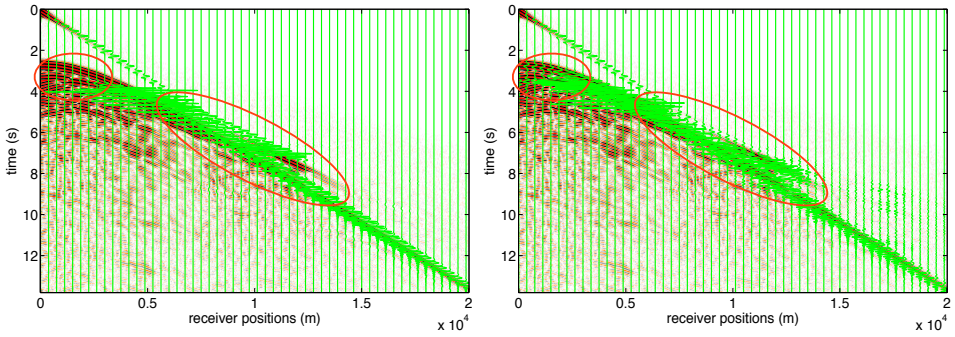


Figure 5.10: FWI result from Chevron Gulf of Mexico data set: (a) ray based tomography starting model for FWI; (b) Gauss-Newton result with unconstrained objective function; (c) inverted result with modified Gauss-Newton with ℓ_2 constraint. (d) inverted result with modified Gauss-Newton with ℓ_1 constraint.

being less sensitive to unmodeled components in the data, by considering each Gauss-Newton subproblem as a compressive-sensing type of sparse recovery problem. Compared to conventional linear sparse inversion problems, full-waveform inversion is significantly more challenging because it is nonlinear, and therefore, it is not clear how sparsity promotion could benefit the inversion. By means of carefully selected examples, we have attempted to classify under which conditions sparsity constraints on the model updates improve the inversion results. We found that for earth models that are sparse in the curvelet domain, improved inversion results can be obtained as long as the model updates are also sparse with locations of significant coefficients persisting amongst the different model updates. We verified this empirical observation on quadratic problems with sparse spikes and on two realistic synthetic data sets for which we obtained improved results when imposing



(a) **(b)**

Figure 5.11: Sample shot comparison (black wiggle is one of the true observe shot at 60km, while background is simulated shot record with initial model or FWI result): (a) initial model shot record; (b) FWI result shot record.

sparsity on the updates rather than on the model itself.

Our examples exhibited that nonlinear inversions with constraints on the model itself, even when the one-norm of this model is known, do not necessarily lead to accurate results. Ad hoc relaxation of the constraints helped, but still led to erroneous results; however, results from the modified Gauss-Newton method with one-norm constraints greatly improved the results while relying on automatic choices for the constraints for each model update. Empirical observations demonstrated that sparse recovery techniques from the well-understood compressive-sensing framework at least partially carry over to nonlinear full-waveform problems for earth models and model updates that permit sparse representations in some transformed domains. While the results on the blind Gulf of Mexico salt model leave room for improvement, the proposed method at least demonstrates that promoting curvelet-domain sparsity improves the results and reduces the reliance on labor intensive data processing, parameter selections, and hand picking of the top and bottom of the salt.

Chapter 6

Conclusions

6.1 Seismic imaging

In the first of this thesis (chapter 2), we combined randomized dimensionally reduction and a curvelet-based sparsity-promoting promoting inversion algorithm to create high-resolution images at a reduced computational cost while also mitigating subsampling related artifacts. Randomized dimensionality reduction saved computation costs by replacing all the sources with subsets of randomly selected sequential shots or simultaneous shots, requiring fewer wave-equation solves. This process creates subsampling related artifacts, such as source-cross talk and non-uniform illumination. We mitigated these subsampling related artifacts using sparsity-promoting inversion, exploiting the structure of the imaging results in the curvelet domain, which efficiently represents geological models. Our approach allowed us to carry out seismic imaging with subsampled sources at the cost of approximately one reverse-time migration with all the sources, obtaining significantly better results. We made our case by providing examples, as well as convergence rate and model error, and performing a case study of the BG compass model dataset, generated from information of real well-log data.

6.2 Full-waveform inversion

In the second portion of this thesis (chapter 3, chapter 4 and chapter 5), we introduced a modified Gauss-Newton algorithm to solve the full-waveform inversion problem. Again, we utilized the randomized dimensionality reduction technique in order to avoid the high costs of computing the Gauss-Newton updates; we also used sparsity-promoting inversion to mitigate the corresponding subsampling related artifacts. As a result, we discovered that we could compute the modified Gauss-Newton updates at the approximate cost of one gradient direction with all sources, and also, we found the final result provides a better resolution compared to alternative methods without promoting sparsity on the updates. Finally, we discovered our method has another advantage. The most intuitive way to regularize the velocity model is the directly impose constraint upon the velocity model itself; however, by utilizing the modified Gauss-Newton method, we have observed that imposing constraint on the individual updates can also regularize the model. We explored different situations in which the Gauss-Newton method can be applied by using carefully selected examples, such as phase-retrieval problem, the BG compass model and the Chevron Gulf of Mexico blind full-waveform inversion test.

6.3 Future extensions

6.3.1 Density variation

Including density in the full-waveform inversion or seismic imaging will increase its accuracy because real data has contributions from density variations. To invert for density, we can borrow ideas from joint-sparse recovery van den Berg and Friedlander (2009); Miao (2014), which is designed to recover an unknown sparse matrix from sets of compressed measurements. According to Gardner et al. (1974), density and velocity models from the same area often share same structures, meaning both of them will share similar curvelet support, as shown in Figure 6.1. Figure 6.1a and Figure 6.1b show a synthetic BG Compass model, generated constrained by real well-log information. Figure 6.1c and Figure 6.1d show the synthesis curvelet coefficients of Figure 6.1a and Figure 6.1b. We observed from the experiment that about 93% non-zero coefficients overlap, allowing us to

invert density and velocity together with joint-sparsity promotion.

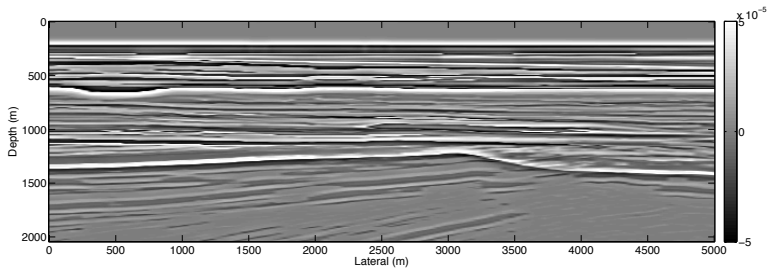
Joint-sparse recovery: In (van den Berg and Friedlander, 2009), a multiple-measurement-vector (MMV) problem can be formed as

$$\underset{\mathbf{X}}{\text{minimize}} \|\mathbf{X}\|_{p,q} \quad \text{subject to} \quad \mathscr{A}(\mathbf{X}) = \mathbf{b}, \tag{6.1}$$

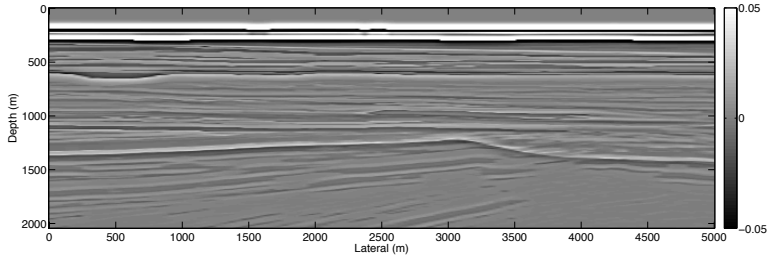
where $\mathscr{A}$ is a sampling operator acting on the columns of the matrix $\mathbf{X}$. The measurement $\mathbf{b}$ is a vector. The $\ell_{p,q}$ norm of $\mathbf{X}$ is defined as $\|\mathbf{X}\|_{p,q} = (\sum_{j=1}^{n} \|\mathbf{X}_{j,:}\|_q^p)^{1/p}$, in which $\mathbf{X}_{j,:}$ is j^{th} row of $\mathbf{X}$. According to van den Berg and Friedlander (2009), joint sparsity via $\ell_{1,2}$ norm can provide better recovery compared to sparsity promotion on $\mathbf{X}$ organized as a long vector (this corresponds to the $\ell_{1,1}$ norm), when $\mathbf{X}$ has nonzero entries in a few rows (ie. $\mathbf{X}$ is row-sparse). To test the improved recovery by joint sparsity promotion, we sample a sparse 512×2 matrix $\mathbf{X}$ with a sampling operator $\mathscr{A}$, which contains two random 120×512 Gaussian matrices. There are only 20 non-zeros in each column of $\mathbf{X}$; 16 non-zeros in these two columns are at the same location. We generate the measurement vector $\mathbf{b}$ by $\mathbf{b} = \mathscr{A}(\mathbf{X}) + \mathbf{e}$, where $\mathbf{e}$ is a random noise vector. From the recovered results in Figure 6.2, we can easily observe the advantage of joint-sparsity promotion when $\mathbf{X}$ is row-sparse.

6.3.2 Time-domain approach

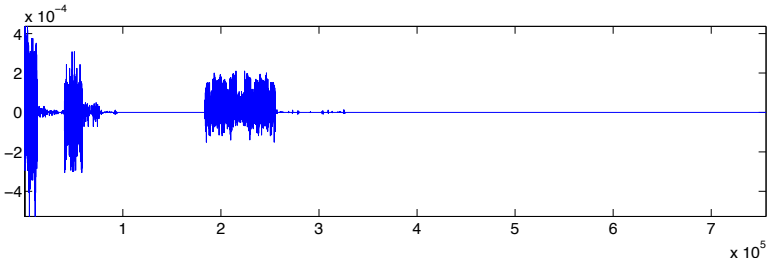
In order to produce accurate inversion results, we must select significant data along the time axis. For example, full-waveform inversion is mostly driven by refraction waves, containing long wavelength information which builds up background velocity. Additionally, seismic imaging is mainly driven by the reflected wave, which contains information of the reflectivity position. This cannot be easily done in the frequency domain because we must simulate data of all frequencies which are computationally intractable; however, in the time domain it is relatively easy to remove irrelevant data. Figure 6.3 shows the actions of three different operator $\mathscr{B}$ and their corresponding Gauss-Newton search directions. Figure 6.3a, Figure 6.3c and Figure 6.3e are full data, refraction data and reflection data, respectively. The Gauss-Newton search direction of the refraction wave contains low-frequency background



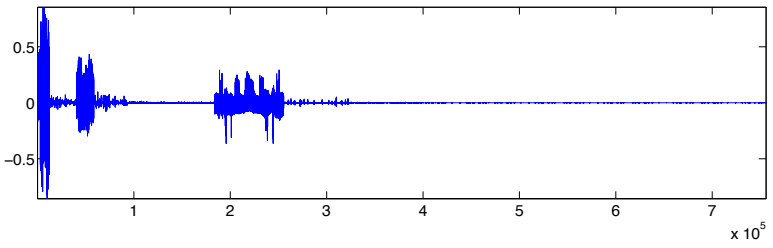
(a)



(b)

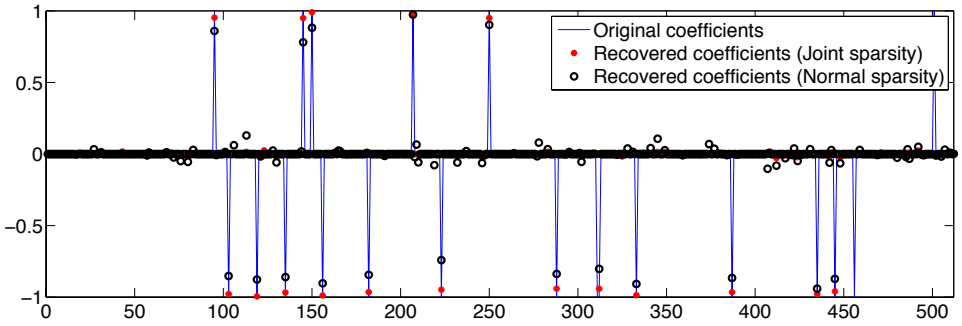


(c)

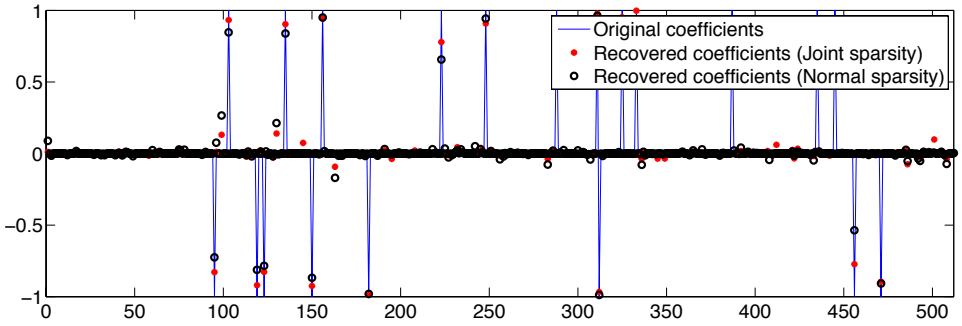


(d)

Figure 6.1: Curvelet domain sparsity. **(a:)** True velocity perturbation. **(b:)** True density perturbation. **(c:)** Curvelet synthesis of **a**. **(d:)** Curvelet synthesis of **b**.



(a)



(b)

Figure 6.2: Joint-recovery from MMV. **(a)** First row of **X**. **(b)** Second row of **X**.

velocity components; while the direction of the reflected waves contain more information of the reflector positions. By sequentially selecting and utilizing the different data, we should be able to construct a correct background velocity model before introducing accurate details to the solution.

6.3.3 Towards 3D

It is more expensive to solve 3D than 2D wave-equations. The seismic wave has to propagate exponentially more grid points in the 3D model, meaning that the cost will grow rapidly with its size. Additionally, exclusive to the frequency domain, the 3D Helmholtz matrix has more off-diagonal entries than the 2D Helmholtz matrix because of the differing wave-equation stencils. Therefore, the 3D Helmholtz matrix is more dense, making it more expensive or even impossible to invert. This

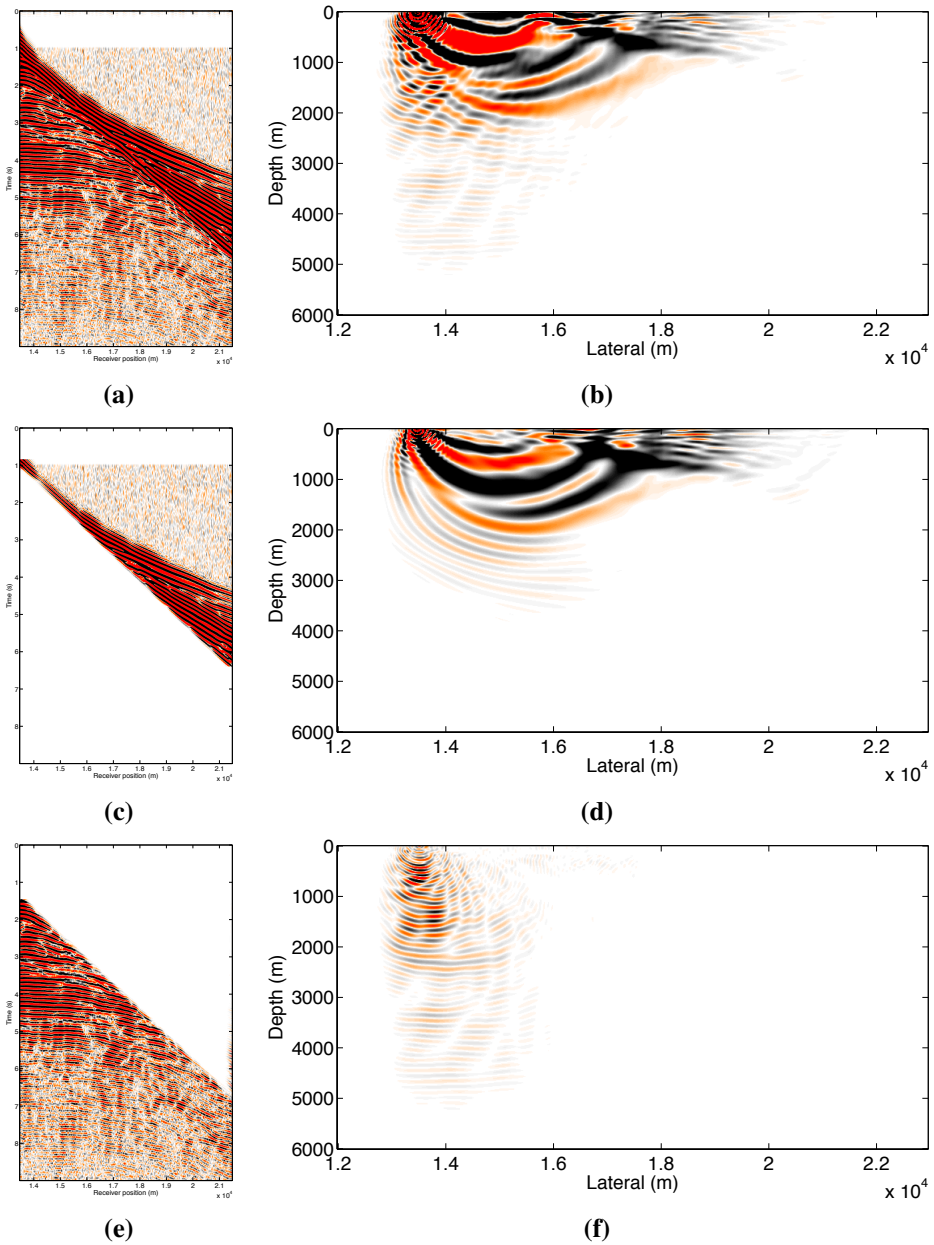


Figure 6.3: Gauss-Newton search directions. **(a)** full data residual. **(b)** direction from **(a)**; **(c):** muted data residual (keeping refraction wave); **(d)** direction from **(c)**; **(e):** muted data residual (keeping reflection wave); **(f)** direction from **(e)**;

creates a situation in which we can use the randomized dimensionality reduction techniques proposed in this thesis to reduce costs even more significantly in the 3D than the 2D problems. Additionally, 3D geometry provides us more space to subsample the data in two dimensions on the surface; therefore, dimensionality reduction techniques could be more efficient for 3D problems than 2D problems.

Bibliography

Abubakar, A., and P. M. van den Berg, 2002, The contrast source inversion method for location and shape reconstructions: Inverse Problems, **18**, 495. → pages 71

Andersson, F., M. V. de Hoop, H. F. Smith, and G. Uhlmann, 2008, A multi-scale approach to hyperbolic evolution equations with limited smoothness: Communications in Partial Differential Equations, **33**, 988–1017. → pages 26

Aoki, N., and G. T. Schuster, 2009, 568, *in* Fast least squares migration with a deblurring filter: 2829–2833. → pages 2

Aravkin, A., T. van Leeuwen, and F. Herrmann, 2011, Robust full-waveform inversion using the student's t-distribution: SEG, Expanded Abstracts, **30**, 2669–2673. → pages 39

Askan, A., V. Akcelik, J. Bielak, and O. Ghattas, 2007, Full waveform inversion for seismic velocity and anelastic losses in heterogeneous structures: Bulletin of the Seismological Society of America, **97**, 1990–2008. → pages 69

Avron, H., and S. Toledo, 2011, Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix: J. ACM, **58**, no. 2, 8:1–8:34. → pages 13

Ayeni, G., 2010, Seismic reservoir monitoring with permanent encoded seismic arrays: SEG, Expanded Abstracts, **29**, 4221–4226. → pages 7

Bauschke, H. H., P. L. Combettes, and D. R. Luke, 2002, Phase retrieval, error reduction algorithm, and fienup variants: a view from convex optimization: JOSA A, **19**, 1334–1345. → pages 70

Berg, E. v., and M. P. Friedlander, 2008, Probing the Pareto frontier for basis pursuit solutions: SIAM Journal on Scientific Computing, **31**, 890–912. → pages x, 16, 17, 18, 37

Bertsekas, D., and J. Tsitsiklis, 1996, Neuro-Dynamic Programming (Belmont, MA: Athena Scientific). → pages 8, 10, 12, 34, 38

Betrsekas, D. P., and J. N. Tsitsiklis, 2000, Gradient convergence in gradient methods with errors: Siam Journal of Optimization, **10**, 627–642. → pages 51

Beylkin, G., 1984, The inversion problem and applications of the generalized

radon transform: Communications on Pure and Applied Mathetmatics, **37**, 579–599. → pages 55

Birgin, E. G., J. Martinez, and M. Raydan, 2010, Nonmonotone spectral projected gradient methods on convex sets: Siam J. Optim., **10**, 1196–1211. → pages 54

Boonyasiriwat, C., and G. T. Schuster, 2010, 3d multisource full-waveform inversion using dynamic random phase encoding: SEG Technical Program Expanded Abstracts, **29**, 1044–1049. → pages 48

Bourgeois, A., M. Bourget, P. Lailly, M. Poulet, P. Ricarte, and R. Versteeg, 1991, Marmousi data and model, *in* The Marmousi experience: EAGE, **5-9**. → pages 17

Bunks, C., F. Saleck, S. Zaleski, and G. Chavent, 1995, Multiscale seismic waveform inversion: Geophysics, **60**, 1457–1473. → pages 3, 33, 38, 60, 88

Burke, J., 1990, Numerical optimization: Course notes for math 516 (University of Washington).: http://www.math.washington.edu/ burke/crs/516/index.html. → pages 36, 57

Burke, J. V., 1992, A robust trust region method for constrained nonlinear programming problems: SIAM Journal on Optimization, **2**, 325–347. → pages 74, 77

Candès, E., J. Romberg, and T. Tao, 2006, Stable signal recovery from incomplete and inaccurate measurements: Comm. Pure Appl. Math., **59**, 1207–1223. → pages 2, 8, 33

Candes, E. J., 2006, Compressive sampling: Presented at the Proceedings of the International Congress of Mathematicians. → pages 52

Candes, E. J., and L. Demanet, 2004, The curvelet representation of wave propagators is optimally sparse: Communications on Pure and Applied Mathematics, **58**, 1472–1528. → pages 46, 49

Candès, E. J., L. Demanet, D. L. Donoho, and L. Ying, 2006, Fast discrete curvelet transforms: Multiscale Modeling and Simulation, **5**, no. 3, 861–899. → pages 16, 37, 46, 49, 71

Claerbout, J. F., 1985, Imaging the earth's interior: Blackwell Scientific Publications, Inc. → pages 1

Crase, E., A. Pica, M. Noble, J. McDonald, and A. Tarantola, 1990, Robust elastic nonlinear waveform inversion: Application to real data: Geophysics, **55**, 527–538. → pages 68

Daubechies, I., M. Defrise, and C. de Mol, 2005, An iterative thresholding algorithm for linear inverse problems with a sparsity constrains: CPAM, 1413–1457. → pages 36

de Hoop, M. V., and S. Brandsberg-Dahl, 2000, Maslov asymptotic extension of generalized radon transform inversion in anisotropic elastic media: a least-squares approach.: Inverse problems, **16**, 519–562. → pages 55

Demanet, L., and G. Peyré, 2011, Compressive wave computation: Found. Comput. Math., **11**, no. 3, 257–303. → pages 28, 61

Demanet, L., and L. Ying, 2007, Curvelets and wave atoms for mirror-extended images: Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series. → pages 37

Doel, K. v., and U. Ascher, 2011, Adaptive and stochastic algorithms for eit and dc resistivity problems with piecewise constant solutions and many measurements. → pages 37

Donoho, D. L., 2006, Compressed sensing: IEEE Trans. Inform. Theory, **52**, 1289–1306. → pages 2, 8, 33, 49, 52

Douma, H., and M. V. de Hoop, 2007, Leading-order seismic imaging using curvelets: Geophysics, **72**, S231–S248. → pages 26

Fei, T. W., Y. Luo, and G. T. Schuster, 2010, De-blending reverse-time migration: SEG, Expanded Abstracts, **29**, 3130–3134. → pages 7

Fichtner, A., 2011, Full Seismic Waveform Moldelling and Inversion: Springer. → pages 39

Friedlander, M. P., and M. Schmidt, 2011, Hybrid deterministic-stochastic methods for data fitting: Tech. rep., Department of Computer Science, University of British Columbia, Vancouver. (revised September 2011, http://arxiv.org/pdf/1104.2373v2). → pages 28

Gardner, G., L. Gardner, and A. Gregory, 1974, Formation velocity and density-the diagnostic basics for stratigraphic traps: Geophysics, **39**, 770–780. → pages 96

Gauthier, O., J. Virieux, and A. Tarantola, 1986, Two-dimensional nonlinear inversion of seismic waveforms: Numerical results: Geophysics, **51**, 1387–1403. → pages 69

Gilbert, J. C., and J. Nocedal, 1992, Global convergence properties of conjugate gradient methods for optimization: SIAM Journal on optimization, **2**, 21–42. → pages 3, 68

Gratton, S., A. S. Lawless, and N. K. Nichols, 2007, Approximate gaussnewton methods for nonlinear least squares problems: SIAM, **18**, no. 1, 106–132. → pages 4, 68

Gray, S. H., 1997, True amplitude seismic migration: A comparison of three approaches: Geophysics, **62**, 929–936. → pages 1

Guitton, A., G. Ayeni, and G. Gonzales, 2010, A preconditioning scheme for full waveform inversion: SEG, Expanded Abstracts, **29**, 1008–1012. → pages 28, 39

Guitton, A., and D. J. Verschuur, 2004, Adaptive subtraction of multiples using the l^1-norm: Geophys Prospect, **52**, 27–27. → pages 1

Habashy, T. M., A. Abubakar, G. Pan, and A. Belani, 2010, Full-waveform

seismic inversion using the source-receiver compression approach: SEG, Expanded Abstracts, **29**, 1023–1028. → pages 8

Haber, E., M. Chung, and F. J. Herrmann, 2010a, An effective method for parameter estimation with PDE constraints with multiple right hand sides: Technical Report TR-2010-4, UBC-Earth and Ocean Sciences Department. (revised August 2011, http://slim.eos.ubc.ca/Publications/Public/Journals/Haber2010emp.pdf). → pages 8, 10, 12, 15, 33, 34, 38

———, 2010b, An effective method for parameter estimation with pde constraints with multiple right hand sides: Technical Report TR-2010-4, UBC-Earth and Ocean Sciences Department. → pages 48

Hansen, P. C., 1997, Rank-Deficient and Discrete Ill-posed Problems: SIAM. → pages 16

———, 1998, Rank-deficient and discrete ill-posed problems: numerical aspects of linear inversion: Siam, **4**. → pages 69, 71

Hennenfent, G., and F. J. Herrmann, 2006a, Seismic denoising with non-uniformly sampled curvelets: Computing in Science and Engineering, **8**, no. 3. → pages 46, 49

———, 2006b, Seismic denoising with nonuniformly sampled curvelets: Computing in Science & Engineering, **8**, 16–25. → pages 92

Hennenfent, G., E. van den Berg, M. P. Friedlander, and F. J. Herrmann, 2008a, New insights into one-norm solvers from the Pareto curve: Geophysics, **73**, no. 4, A23–A26. → pages 17, 37, 54

———, 2008b, New insights into one-norm solvers from the Pareto curve: Geophysics, **73**, A23–A26. → pages 69, 74, 79

Herrmann, F. J., 2003, Multifractional splines: application to seismic imaging: Proceedings of SPIE Technical Conference on Wavelets: Applications in Signal and Image Processing X, SPIE, SPIE, 240–258. → pages 5

———, 2010, Randomized sampling and sparsity: Getting more information from fewer samples: Geophysics, **75**, WB173–WB187. → pages 13

Herrmann, F. J., C. R. Brown, Y. A. Erlangga, and P. P. Moghaddam, 2009a, Curvelet-based migration preconditioning and scaling: Geophysics, **74**, A41–A46. → pages 3, 10, 55

Herrmann, F. J., A. J. Calvert, I. Hanlon, M. Javanmehri, R. Kumar, T. van Leeuwen, X. Li, B. Smithyman, E. T. Takougang, and H. Wason, 2013, Frugal full-waveform inversion: from theory to a practical algorithm: The Leading Edge, **32**, 1082–1092. → pages 91

Herrmann, F. J., Y. A. Erlangga, and T. Lin, 2009b, Compressive simultaneous full-waveform simulation: Geophysics, **74**, A35. → pages 2, 8, 13, 15

Herrmann, F. J., and X. Li, 2011a, Efficient least-squares imaging with sparsity

promotion and compressive sensing.: Tech. rep., University of British Columbia, Vancouver. → pages 36, 38, 61

———, 2011b, Efficient least-squares migration with sparsity promotion: Presented at the , EAGE, EAGE Technical Program Expanded Abstracts. → pages 14, 15, 36, 38, 49

———, 2012, Efficient least-squares imaging with sparsity promotion and compressive sensing: Geophysical Prospecting, **60**, 696–712. → pages 2, 69

Herrmann, F. J., X. Li, A. Aravkin, and T. van Leeuwen, 2011a, A modified, sparsity promoting, Gauss-Newton algorithm for seismic waveform inversion.: Proc. SPIE, **2011**, no. 81380V. → pages 20, 25, 36

Herrmann, F. J., X. Li, A. Y. Aravkin, and T. van Leeuwen, 2011b, A modified, sparsity promoting, Gauss-Newton algorithm for seismic waveform inversion: , 81380V–81380V–14. → pages 69

Herrmann, F. J., P. P. Moghaddam, and C. C. Stolk, 2008a, Sparsity- and continuity-promoting seismic image recovery with curvelet frames: Applied and Computational Harmonic Analysis, **24**, no. 2, 150–173. → pages 3, 55

———, 2008b, Sparsity- and continuity-promoting seismic imaging with curvelet frames: Journal of Applied and Computational Harmonic Analysis, **24**, 150–173. (doi:10.1016/j.acha.2007.06.007). → pages 13, 26, 36

Hestenes, M. R., and E. Stiefel, 1952, Methods of Conjugate Gradients for Solving Linear Systems: Journal of Research of the National Bureau of Standards, **49**, 409–436. → pages 4, 68

Horst, R., P. Pardalos, and N. Van Thoai, 2000, Introduction to global optimization: Springer. Nonconvex Optimization and Its Applications. → pages 79

Hutchinson, M., 1990, A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines: J. Commun. Statist. Simul, **19**, 433–450. → pages 13

Jo, C. H., C. Shin, and J. H. Suh, 1996a, An optimal 9-point, finite-difference, frequency-space, 2-D scalar wave extrapolator: Geophysics, **61**, 529–537. → pages 18

———, 1996b, An optimal 9-point, finite-difference, frequency-space, 2-d scalar wave extrapolator: Geophysics, **61**, 529–537. → pages 59

Kim, S. J., K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, 2007, An interior-point method for large-scale l1-regularized least squares: IEEE J. Sel. Top. Signal. Process., 606–617. → pages 53

Krebs, J. R., J. E. Anderson, D. Hinkley, R. Neelamani, S. Lee, A. Baumstein, and M.-D. Lacasse, 2009a, Fast full-wavefield seismic inversion using encoded sources: Geophysics, **74**, WCC177–WCC188. → pages 8, 12, 34

———, 2009b, Fast full-wavefield seismic inversion using encoded sources:

Geophysics, **74**, WCC177–WCC188. → pages 48
Kühl, H., and M. D. Sacchi, 2003, Least-squares wave-equation migration for avp/ava inversion: Geophysics, **68**, 262273. → pages 2, 3
Kumar, V., 2009, Incoherent noise suppression and deconvolution using curvelet-domain sparsity: mastersmasters, University of British Columbia. → pages 3, 92
Lailly, P., 1983, The seismic inverse problem as a sequence of before stack migrations: Proc. Conf. Inverse Scattering, Theory and Applications. → pages 1, 60
Li, X., A. Y. Aravkin, T. van Leeuwen, and F. J. Herrmann, 2012, Fast randomized full-waveform inversion with compressive sensing: Geophysics, **77**, A13–A17. → pages 3, 4, 68, 69, 86, 87, 89
Li, X., and F. J. Herrmann, 2010a, Full-waveform inversion from compressively recovered model updates: SEG, Expanded Abstracts, **29**, 1029–1033. → pages 8, 20, 25, 74
———, 2010b, Full waveform inversion from compressively recovered model updates: SEG Expanded Abstracts, **29**, 1029–1033. → pages 48
Lin, T. T., and F. J. Herrmann, 2013, Robust estimation of primaries by sparse inversion via one-norm minimization: Geophysics, **78**, R133–R150. → pages 69, 74
Lin, T. T. Y., and F. J. Herrmann, 2007, Compressed wavefield extrapolation: Geophysics, **72**, no. 5, SM77–SM93. → pages 15
Liu, D., and J. Nocedal., 1989, On the limited memory method for large scale optimization: Mathematical Programming B, **45**, no. 3, pp. 503–528. → pages 3
Mallat, S. G., 2009, A Wavelet Tour of Signal Processing: the Sparse Way: Academic Press. → pages 8, 33, 36
Mansour, H., H. Wason, T. T. Lin, and F. J. Herrmann, 2011, Simultaneous-source marine acquisition with compressive sampling matrices: Technical Report TR-2011-02, UBC-Earth and Ocean Sciences Department. (revised October 2011, for publication in this issue). → pages 28
Miao, L., 2014, Efficient seismic imaging with spectral projector and joint sparsity: masters, University of British Columbia. → pages 96
Moghaddam, P. P., and F. J. Herrmann, 2010a, Randomized full-waveform inversion: a dimenstionality-reduction approach: , SEG, 977–982. → pages 34
———, 2010b, Randomized full-waveform inversion: a dimenstionality-reduction approach: SEG Technical Program Expanded Abstracts, **29**, 977–982. → pages 48
Montanari, A., 2010, Graphical models concepts in compressed sensing: Arxiv preprint arXiv:1011.4328, **abs/1011.4328**. → pages 28

Mora, P., 1987, Nonlinear two-dimensional elastic inversion of multioffset seismic data: Geophysics, **52**, 1211–1228. → pages 68

Morton, S. A., and C. C. Ober, 1998, Faster shot-record depth migrations using phase encoding: SEG Technical Program Expanded Abstracts, SEG, 1131–1134. → pages 7

Mulder, W., and R. Plessix, 2004, How to choose a subset of frequencies in frequency-domain finite-difference migration: Geoph. J. Int., **158**, 801–812. → pages 8

Mtivier, L., R. Brossier, J. Virieux, and S. Operto., 2013, Full waveform inversion and the truncated newton method: SIAM, **35**, B401–B437. → pages 3, 4, 68

Natterer, F., 2001, The mathematics of computerized tomography: Society for Industrial Mathematics. → pages 12

Neelamani, R. N., C. E. Krohn, J. R. Krebs, J. K. Romberg, M. Deffenbaugh, and J. E. Anderson, 2010, Efficient seismic forward modeling using simultaneous random sources and sparsity: Geophysics, **75**, WB15–WB27. → pages 2, 8, 13

Nemeth, T., C. Wu, and G. T. Schuster, 1999, Least-squares migration incomplete reflection data: Geophysics, **64**, 208–221. → pages 2, 3

Nemirovski, A., A. Juditsky, G. Lan, and A. Shapiro, 2009a, Robust stochastic approximation approach to stochastic programming: SIAM Journal on Optimization, **19**, 1574–1609. → pages 8, 10, 12, 34

———, 2009b, Robust stochastic approximation approach to stochastic programming: Siam J. Optim., **19**, 1574–1609. → pages 50

Nocedal, J., and S. J. Wright, 2006a, Least-squares problems: Springer. → pages 70, 72

———, 2006b, Numerical optimization, 2nd ed.: Springer Verlag. → pages 38

Paige, C. C., and M. A. Saunders, 1982, Lsqr: An algorithm for sparse linear equations and sparse least squares: ACM TOMS, **8**, 43–71. → pages 11, 16, 38

Plessix, R., and W. Mulder, 2004, Frequency-domain finite difference amplitude-preserving migration: Geoph. J. Int., **157**, 975–987. → pages 2, 7, 9

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**, 495–503. → pages 60

Pratt, R., C. Shin, and G. Hicks, 1998a, Gauss-Newton and full Newton methods in frequency-space waveform inversion: Geoph. J. Int., **133**, 341–362. → pages 3, 4, 67, 68

Pratt, R., Z. Song, P. Williamson, and M. Warner, 1996, Two-dimensional velocity models from wide-angle seismic data by wavefield inversion: Geophysical Journal International, **124**, 232–340. → pages 60

Pratt, R. G., C. Shin, and G. Hicks, 1998b, Gauss-newton and full newton methods in frequency-space seismic waveform inversion: Geophysical Journal

International, **133**, 341362. → pages 67
Rickett, J. E., 2003, Illumination-based normalization for wave-equation depth migration: Geophysics, **68**, no. 4, 1371–1379. → pages 7
Romberg, J., 2009, Compressive sensing by random convolution: SIAM Journal on Imaging Sciences, **2**, 1098–1128. → pages 15
Romero, L. A., D. C. Ghiglia, C. C. Ober, and S. A. Morton, 2000, Phase encoding of shot records in prestack migration: Geophysics, **65**, no. 2, 426–436. → pages 7
Routh, P., J. Krebs, S. Lazaratos, A. Baumstein, S. Lee, Y. H. Cha, I. Chikichev, N. Downey, D. Hinkley, and J. Anderson, 2011, Encoded simultaneous source fullwavefield inversion for spectrally shaped marine streamer data: SEG, Expanded Abstracts, **30**, 2433–2438. → pages 39
Ruszczyński, A., 2006, Nonlinear optimization: Princeton University Press. Nonlinear optimization, No. v. 13. → pages 79
Shapiro, A., 2003, Monte carlo sampling methods, *in* Stochastic Programming, Volume 10 of Handbooks in Operation Research ad Management Science: North-Holland. → pages 50
Shapiro, A., D. Dentcheva, and D. Ruszczynski, 2009, Lectures on stochastic programming: Modeling and theory: SIAM, Philadelphia. → pages 8
Shapiro, A., and A. Nemirovsky, 2005, On complexity of stochastic programming problems, *in* Continuous Optimization: Current Trends and Applications: Springer, New York. → pages 50
Shin, C., S. Jang, and D.-J. Min, 2001, Improved amplitude preservation for prestack depth migration by inverse scattering theory: Geophysical Prospecting, **49**, 592–606. → pages 68
Sirgue, L., and R. G. Pratt, 2004, Efficient waveform inversion and imaging: A strategy for selecting temporal frequencies: Geophysics, **69**, 231–248. → pages 8
Smith, H. F., 1998, A Hardy space for Fourier integral operators.: J. Geom. Anal., **8**, 629–653. → pages 28, 61
Stolk, C. C., and W. W. Symes, 2003, Smooth objective functionals for seismic velocity inversion: Inverse Problems, **19**, 73–89. → pages 55
Strohmer, T., and R. Vershynin, 2009, A randomized Kaczmarz algorithm with exponential convergence: Journal of Fourier Analysis and Applications, **15**, 262–278. → pages 12, 28
Symes, W., 2008a, Approximate linearized inversion by optimal scaling of prestack depth migration: Geophysics, **73**, R23–R35. → pages 55
———, 2010, Source synthesis for waveform inversion: SEG, Expanded Abstracts, **29**, 1018–1022. → pages 8
Symes, W. W., 2008b, Migration velocity analysis and waveform inversion:

Geophysical Prospecting, **56**, 765–790. → pages 8
Tarantola, A., 1984a, Inversion of seismic reflection data in the acoustic approximation: Geophysics, **49**, 1259–1266. → pages 3, 67
———, 1984b, Inversion of seismic reflection data in the acoustic approximation: Geophysics, **49**, 1259–1266. → pages 46, 60
———, 1986, A strategy for nonlinear elastic inversion of seismic reflection data: Geophysics, **51**, 1893–1903. → pages 68
ten Kroode, A., D.-J. Smit, and A. Verdel, 1998, A microlocal analysis of migration: Wave Motion, **28**. → pages 55
Tibshirani, R., 1997, Regression shrinkage and selection via the lasso: J. Royal. Statist. Soc B., **58**, 267–288. → pages 16
Tu, N., A. Y. Aravkin, T. van Leeuwen, and F. J. Herrmann, 2013, Fast least-squares migration with multiples and source estimation: Presented at the EAGE. → pages 2, 68, 69
van den Berg, E., and M. P. Friedlander, 2008a, Probing the pareto frontier for basis pursuit solutions: SIAM Journal on Scientific Computing, **31**, 890–912. → pages 53, 56, 58
———, 2008b, Probing the pareto frontier for basis pursuit solutions: SIAM Journal on Scientific Computing, **31**, 890–912. → pages 69, 70, 74, 79, 89
———, 2009, Joint-sparse recovery from multiple measurements: Technical report, Department of Computer Science. → pages 96, 97
Van Den Doel, K., U. Ascher, and E. Haber, 2012, The lost honour of l2-based regularization: Large Scale Inverse Problems, Radon Ser. Comput. Appl. Math, **13**, 181–203. → pages 81, 82
van Leeuwen, T., A. Aravkin, and F. J. Herrmann, 2011a, Seismic waveform inversion by stochastic optimization: International Journal of Geophysics, **2011**. → pages 11, 12, 33, 34, 38
van Leeuwen, T., and F. J. Herrmann, 2011, Fast waveform inversion without source encoding.: Tr-2011-06, the University of British Columbia, University of British Columbia, Vancouver. → pages 39
van Leeuwen, T., M. Schmidt, M. Friedlander, and F. J. Herrmann, 2011b, A hybrid stochastic-deterministic optimization method for waveform inversion: EAGE, Expanded Abstracts. → pages 28, 29, 39
Virieux, J., and S. Operto, 2009, An overview of full-waveform inversion in exploration geophysics: Geophysics, **74**, 127–152. → pages 3, 67
Vogel, C. R., and M. E. Oman, 1996, Iterative methods for total variation denoising: SIAM J. Sci. Comput., **17**, no. 1, 227–238. → pages 71
Warner, M., A. Ratcliffe, T. Nangoo, J. Morgan, A. Umpleby, N. Shah, V. Vinje, I. Štekl, L. Guasch, C. Win, et al., 2013, Anisotropic 3d full-waveform inversion: Geophysics, **78**, R59–R80. → pages 68