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

Xiang Li*, Ernie Esser†, and Felix J. Herrmann*
*University of British Columbia, Earth and Ocean Sciences
†Deceased March 8th, 2015

Abstract

Full-waveform inversion can be formulated as a nonlinear least-squares optimization problem. This non-convex problem can be extremely computationally expensive because it requires repeatedly solving large linear systems that correspond to discretized partial differential equations. Randomized subsampling techniques allow us to work with small subsets of (monochromatic) source experiments, reducing the computational cost. However, this subsampling weakens subsurface illumination and introduces subsampling related incoherent artifacts. These subsampling-related artifacts—in conjunction with local minima that are known to plague full-waveform inversion—motivate us to come up with a technique to "regularize" this problem. Following earlier work, we take advantage of the fact that curvelets represent subsurface models and perturbations parsimoniously. At first impulse promoting sparsity on the model directly seems the most natural way to proceed, but we will demonstrate that in certain cases it can be advantageous to promote sparsity on the Gauss-Newton updates instead. While constraining the one-norm of the descent directions does not change not change the underlying full-waveform inversion objective, the constrained model updates remain descent directions, remove subsampling-related artifacts and improve the overall inversion result. We empirically observe this phenomenon in situations where the different model updates occur at roughly the same locations in the curvelet domain. We further investigate and analyze this phenomenon, where nonlinear inversions benefit from sparsity-promoting constraints on the updates, by means of a set of carefully selected examples including phase retrieval and full-waveform inversion. In all cases, we observe a faster decay of the residual and model error as a function of the number of iterations.

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, 1984; Pratt et al., 1998a). In the last twenty years, numerous first-order gradient-based methods have been developed to solve this inversion problem, including gradient descent (Pratt et al., 1998a; M 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 Métivier 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 (Métivier 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 (Herrmann and Li, 2012; 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., 2011) 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, 2008; Lin and Herrmann, 2013; Hennenfent et al., 2008). 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, 2008), 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.

Our paper 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.

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.

The unconstrained least-squares objective

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

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

where the vector $\mathbf{d}$ represents the observed data that we want to fit with the nonlinear forward modeling operator $\boldsymbol{\mathcal{F}}$, parameterized by the discrete and vectorized model $\mathbf{m}$. Without loss of generality, we assume the source function to be known and fixed throughout this paper.

FWI is challenging for the following reasons. First, evaluations of the forward modeling operator $\boldsymbol{\mathcal{F}}[\mathbf{m}]$ 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 paper, 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 $\boldsymbol{\mathcal{F}}$, which are assumed to be differentiable functions with respect to the model $\mathbf{m}$.

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

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

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.

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 $\boldsymbol{\mathcal{F}}$ 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, 2006):

$$\delta\mathbf{m}_k = \arg\min_{\delta\mathbf{m}} \|\delta\mathbf{d}_k - \nabla\boldsymbol{\mathcal{F}}[\mathbf{m}_k]\delta\mathbf{m}\|_2^2. \tag{1}$$

In this expression, $\nabla\boldsymbol{\mathcal{F}}[\mathbf{m}_k]$ represents the Jacobian evaluated at the model iterate of the k^{th} iteration $\mathbf{m}_k$. The vector $\delta\mathbf{d}_k = \mathbf{d} - \boldsymbol{\mathcal{F}}[\mathbf{m}_k]$ contains the corresponding data residual. As outlined in Algorithm 1 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 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 1 is ill conditioned.

Algorithm 1 Gauss-Newton method for unconstrained objective (**LS**).

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** $\|\mathbf{d} - \boldsymbol{\mathcal{F}}[\mathbf{m}_k]\|_2^2 \geq \xi$ **do**
3. $\quad\delta\mathbf{m}_k = \arg\min_{\delta\mathbf{m}} \|\delta\mathbf{d}_k - \nabla\boldsymbol{\mathcal{F}}[\mathbf{m}_k]\delta\mathbf{m}\|_2$ `// 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. $\widetilde{\mathbf{m}} \longleftarrow \mathbf{m}_k$

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

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

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 1 by the constrained least-squares problem of Equation 2.

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 2 below. Solutions of the ℓ_1-norm constrained Gauss-Newton subproblems of the type

$$\delta\mathbf{m}_k = \mathbf{S}^{\mathrm{H}} \arg\min_{\delta\mathbf{x}} \|\delta\mathbf{d}_k - \nabla\boldsymbol{\mathcal{F}}[\mathbf{m}_k]\mathbf{S}^{\mathrm{H}}\delta\mathbf{x}\|_2^2 \quad \text{subject to} \quad \|\delta\mathbf{x}\|_{\ell_1} \leq \tau_k \tag{3}$$

lie at the heart of this approach (Li and Herrmann, 2010). Contrary to Equation 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 3) using a root-finding algorithm on the Pareto tradeoff curve (van den Berg and Friedlander, 2008; Hennenfent et al., 2008; Lin and Herrmann, 2013). For our purpose, it is sufficient to solve for each Gauss-Newton subproblem Equation 3 with the sparsity level set to $\tau_k = \frac{\|\delta\mathbf{d}_k\|_2}{\|\mathbf{S}\nabla\boldsymbol{\mathcal{F}}^{\mathrm{T}}[\mathbf{m}_k]\delta\mathbf{d}\|_\infty}$ 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, 2008; Hennenfent et al., 2008; 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 2, we summarize the details of the modified Gauss-Newton method.

Algorithm 2 Modified Gauss-Newton method with sparse update for unconstrained LS objective function.

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** $\|\mathbf{d} - \boldsymbol{\mathcal{F}}[\mathbf{m}]\|_2^2 \geq \xi$ **do**
3. $\quad \tau_k = \|\delta\mathbf{d}_k\|_2 / \|\mathbf{S}\nabla\boldsymbol{\mathcal{F}}^{\mathrm{H}}[\mathbf{m}_k]\delta\mathbf{d}_k\|_\infty$
4. $\quad \delta\mathbf{m} = \arg\min_{\delta\mathbf{x}} \|\delta\mathbf{d} - \nabla\boldsymbol{\mathcal{F}}[\mathbf{m}_k]\mathbf{S}^{\mathrm{H}}\delta\mathbf{x}\|_2^2$ subject to $\|\delta\mathbf{x}\|_{\ell_1} \leq \tau_k$ `// 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**

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 1, or do exploit sparsity by either constraining the model as in **LS**ℓ_1 and Equation 2 or descent directions as in

Equation 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 are global minima coincide and non-convex problems that may have local minima. For convex problems, we show that Algorithm 2 converges to the solution while Gauss-Newton applied to ℓ_1-norm constrained objective function (Algorithm 1 with line 3 defined by Equation 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.

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 1 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 2. By setting $\boldsymbol{\mathcal{F}}[\mathbf{m}] := \mathbf{A}\mathbf{m}$ 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 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 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 1) and correct ℓ_1-norm constrained (Algorithm 1 with line 3 replaced by Equation 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 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 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 1d or the circular-shaped ℓ_2-norm ball as in Figure 1e). This behavior, where the descent directions are constrained, is consistent with theoretical findings of Burke (1992), who states that Algorithm 2 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 2 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.

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 by now for now 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

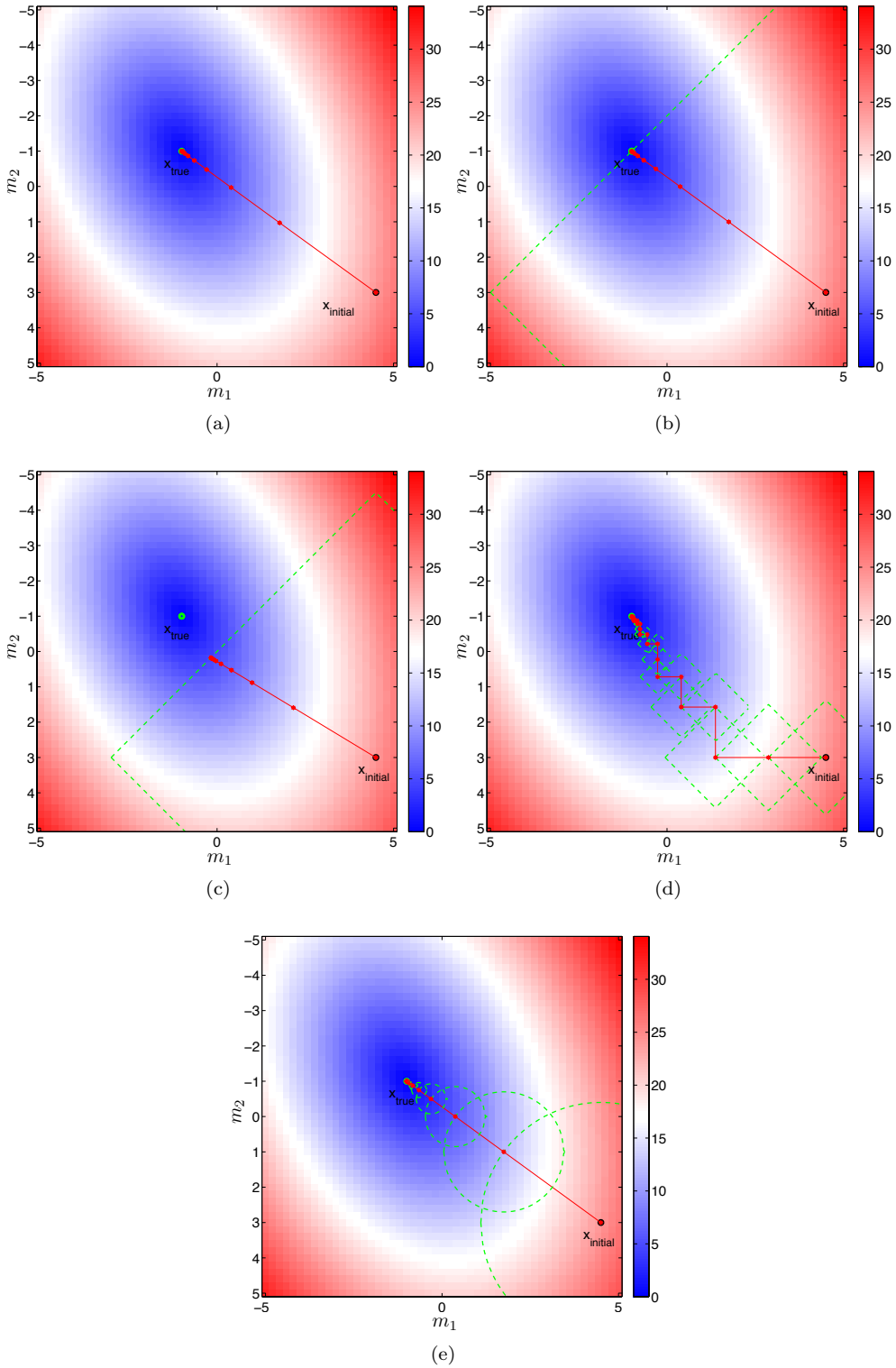


Figure 1: Solution path of different methods for convex problem that has a unique solution: (a) Algorithm 1; (b) Algorithm 1 with line 3 defined by Equation 2 (correct ℓ_1 constraint); (c) same as Figure 1b but with wrong ℓ_1 constraint; (d) Algorithm 2 but with ℓ_1 constraint on the updates; (d) Algorithm 2 but with ℓ_2 constraint.

objective that lie on the line $-4 + 2m_1 + 4m_2 = 0$, and the solutions paths in Figure 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 2b), in which case the solution corresponds to the unconstrained least-squares solution, or too small (Figure 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 2d), which yields a sparse solution where one of the two model parameters is zero.

As illustrated in Figure 2e, solution paths for our modified Gauss-Newton method with ℓ_1-norm constraints on the descent directions also find the sparse solution 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 2. Figure 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 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, 2008; Hennenfent et al., 2008) 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.

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 setting $\boldsymbol{\mathcal{F}}[\mathbf{m}] := (\mathbf{Am})^T\mathbf{Am}$ with $\boldsymbol{\nabla\mathcal{F}}[\mathbf{m}] = (\mathbf{Am})^T$. In this formulation, the data is no longer linear but quadratic in the model parameters. As we can see from Figure 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 3a); constrained with a ℓ_1-norm constraint τ that is too large (Figure 3b, $\tau > \tau_{\text{true}}$); correct (Figure 3c, $\tau = \tau_{\text{true}}$) or gradually relaxed (Figure 3d) from a small τ to a large τ. In all cases (Figures 3a—3c), no sparse solutions (denoted by the green dot) are found

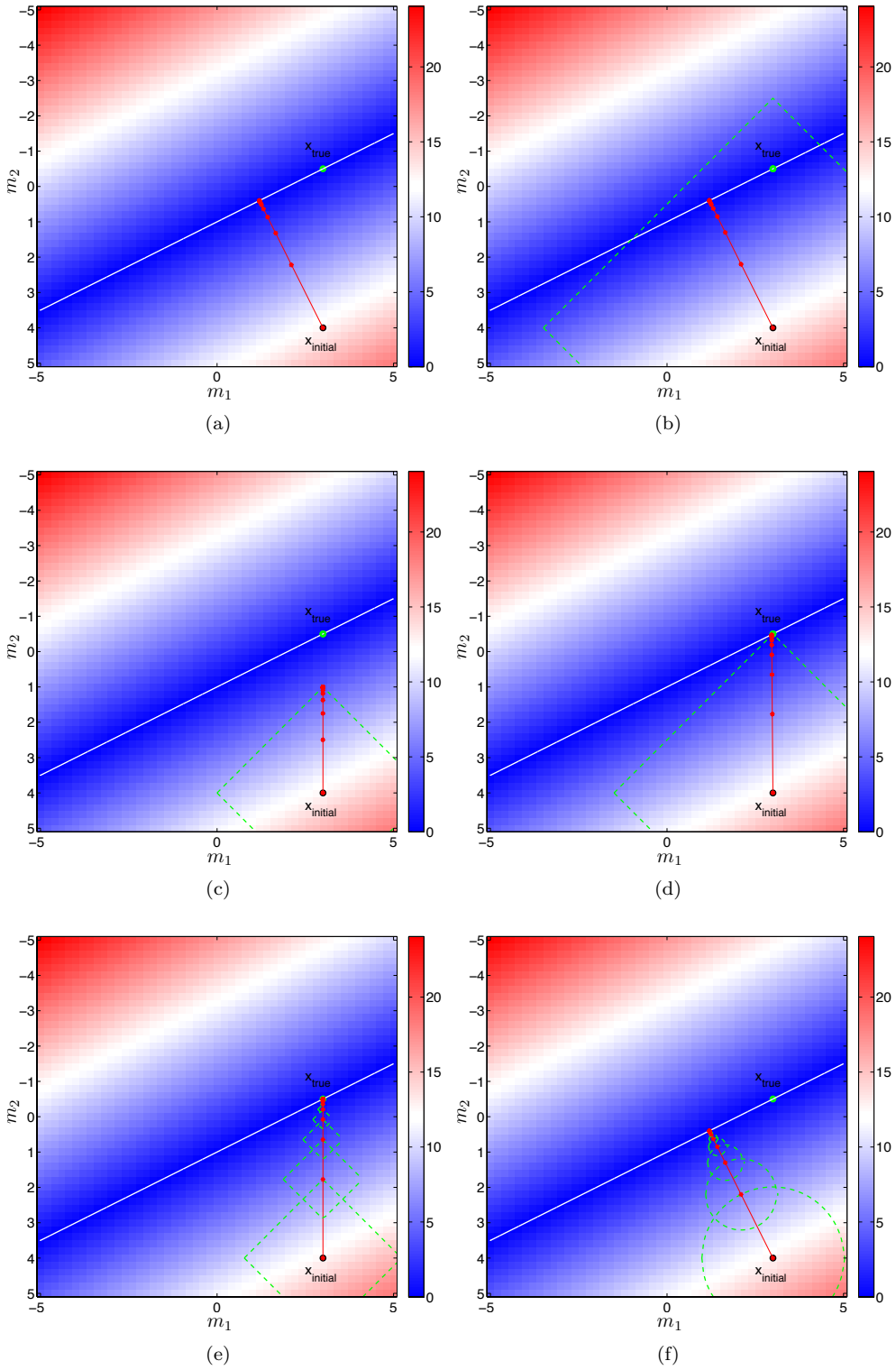


Figure 2: Solution path of different methods for a linear problem that multiple solutions: (a) Algorithm 1; (b) Algorithm 1 with line 3 defined by Equation 2 with wrong constraint ($\tau > \tau_{true}$); (c) same as Figure 1b with wrong constraint ($\tau < \tau_{true}$) ; (d) same as Figure 1b but with the right constraint ($\tau = \tau_{true}$); (e) Algorithm 2; (f) Algorithm 2 but with ℓ_2 constraint on the updates.

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 3e), Algorithm 2 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 3e and 3f). We explain the relative success of the modified Gauss-Newton compared to imposing (the correct) ℓ_1-norm constraint on the model (Figure [#fig:spnlpc]) 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 is now 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.

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 $\mathbf{A}$, $\boldsymbol{\mathcal{F}}[\mathbf{m}] := \text{diag}(\mathbf{Am})(\mathbf{Am})$ and $\boldsymbol{\nabla\mathcal{F}}[\mathbf{m}] = \text{diag}(\mathbf{Am})\mathbf{A}$ our task is to recover the model parameters from data collected in the vector $\mathbf{d} = \boldsymbol{\mathcal{F}}[\mathbf{m}_{\text{true}}]$. 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 $\boldsymbol{\mathcal{F}}[\mathbf{m}]$ notoriously difficult to 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 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 4b and 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.

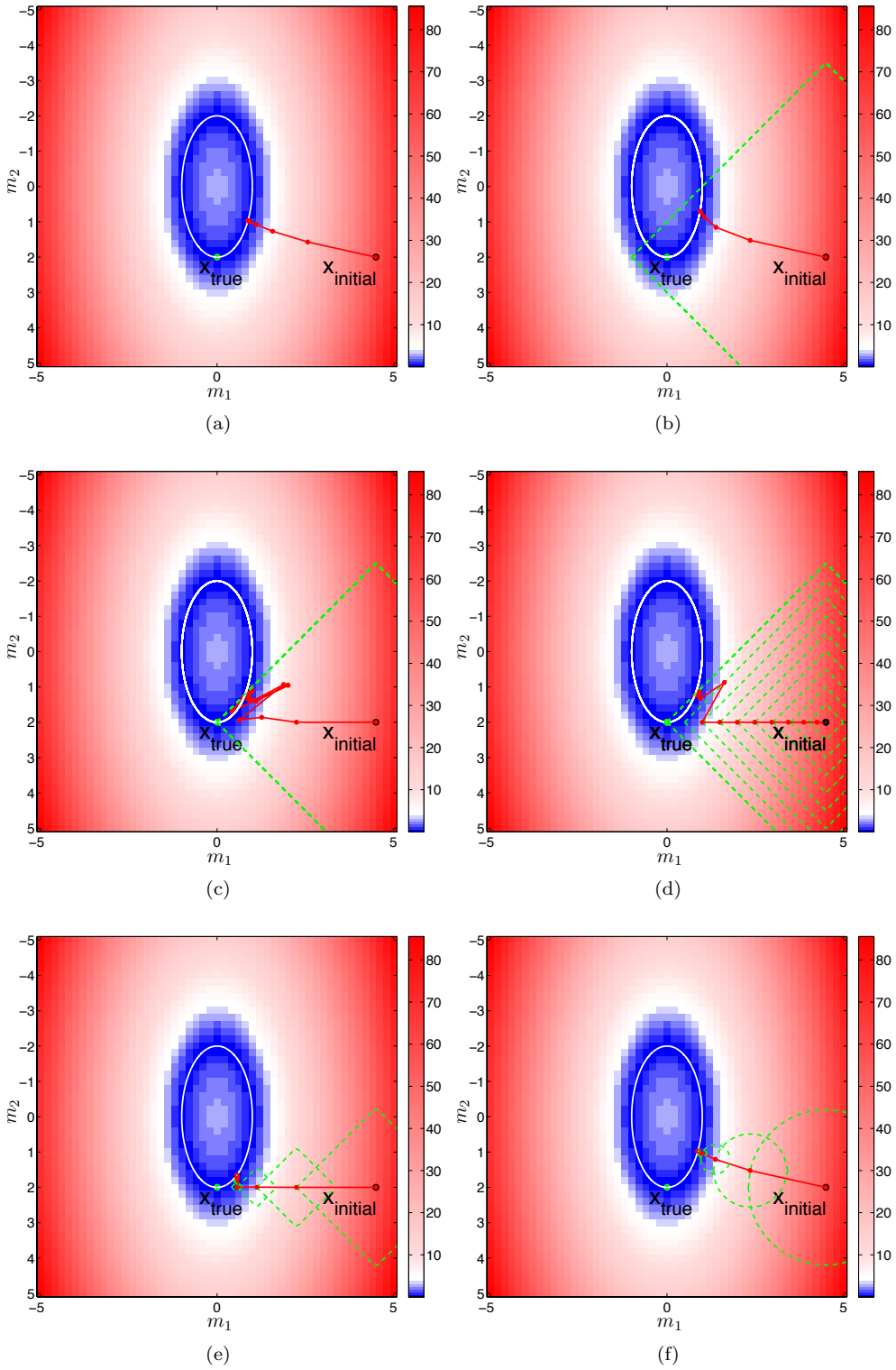


Figure 3: Solution path of different methods for a nonlinear problem that multiple solutions: (a) Algorithm 1; (b) Algorithm 1 with line 3 defined by Equation 2 with wrong constraint ($\tau > \tau_{true}$); (c) same as Figure 1b but with the right constraint; (d) same as Figure 1b but with the gradually relaxed constraint; (e) Algorithm 2; (f) Algorithm 2 with ℓ_2 constraint on the updates.

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 **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, 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, 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 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 4c and 4f).

Our observations are also confirmed by plots for the relative ℓ_2 norms for the residuals and model errors as a function of the number of Gauss-Newton iterations included in Figure 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 2 continues to make progress towards the solution bringing both the relative data residual and model errors down as Algorithm 2 progresses.

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.

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 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

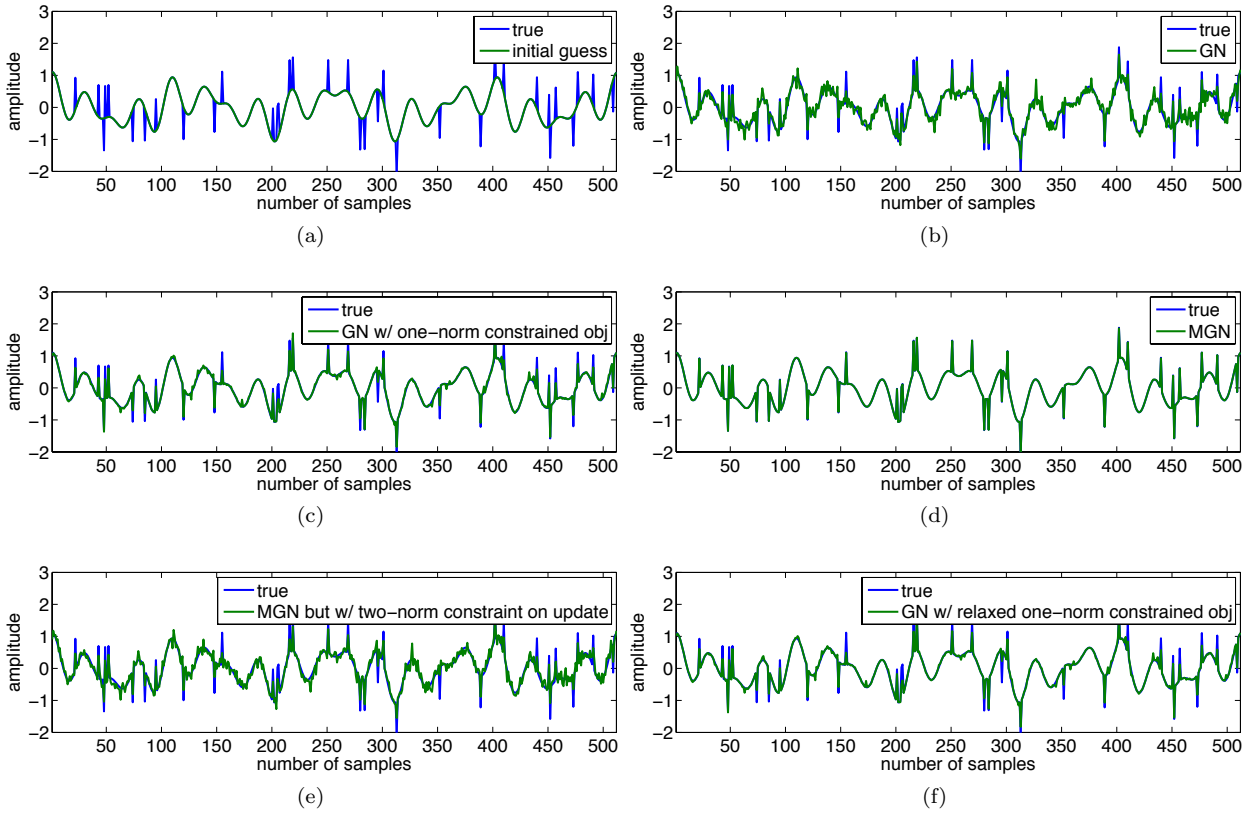


Figure 4: Results for phase retrieval example from difference methods: (a) true solution and initial guess; (b) solution of Algorithm 1; (c) solution of Algorithm 1 with line 3 defined by Equation 2; (d) solution of Algorithm 2; (e) solution of Algorithm 2 but with ℓ_2 constraint on the updates; (f) same as Figure 4c but with gradually relaxed ℓ_1 constrained objective function;

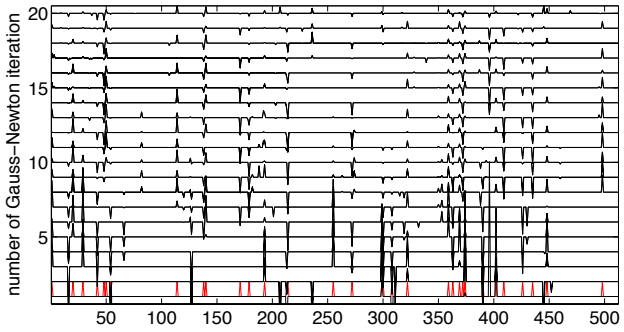
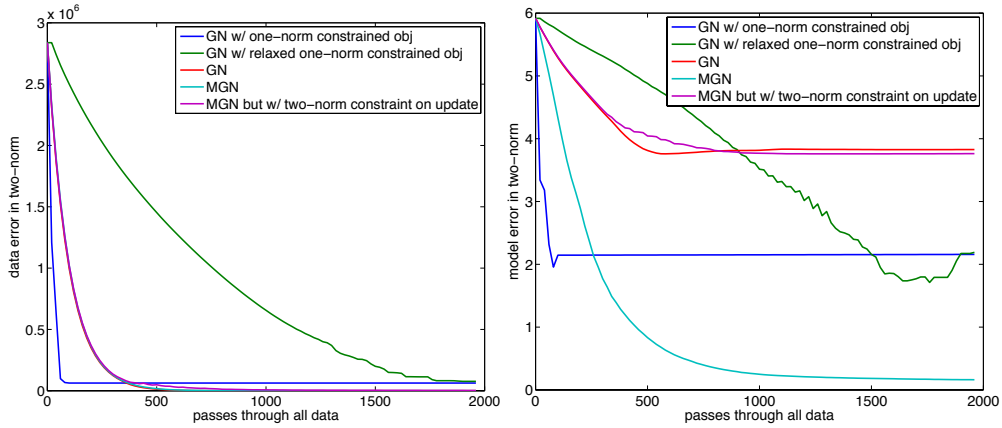


Figure 5: All modified Gauss-Newton updates for the phase retrieval problem.



(a) (b)

Figure 6: Data misfit and relative model error for phase retrieval example: (a) data misfit; (b) relative model error.

optimization problem (Li et al., 2012):

$$\delta\mathbf{m}_k = \mathbf{S}^H \arg\min_{\delta\mathbf{x}} \|\delta\underline{\mathbf{D}}_k - \nabla\boldsymbol{\mathcal{F}}[\mathbf{m}_k, \underline{\mathbf{Q}}_k]\mathbf{S}^H\delta\mathbf{x}\|_F^2 \quad \text{subject to} \quad \|\delta\mathbf{x}\|_{\ell_1} \leq \tau_k. \tag{4}$$

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) $\{\delta\underline{\mathbf{D}}_k = \underline{\mathbf{D}} - \boldsymbol{\mathcal{F}}[\mathbf{m}_k, \underline{\mathbf{Q}}_k]\}$ and sources $\underline{\mathbf{Q}}_k$ that are also randomly selected. We denote these randomly sub-sampled quantities by the underbar. To accelerate the converge, 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 3. 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 {–based–} 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 Golf 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.

BG COMPASS model

The BG COMPASS model (Figure 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.

Algorithm 3 Modified Gauss-Newton with curvelet-domain sparsity promotion and randomization.

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** $\|\delta\underline{\mathbf{D}}_k\|_2 \geq \xi$ **do**
4. $\tau_k = \|\delta\underline{\mathbf{D}}_k\|_F / \|\mathbf{S}\nabla\boldsymbol{\mathcal{F}}^{\mathrm{H}}[\mathbf{m}_k, \underline{\mathbf{Q}}_k]\delta\underline{\mathbf{D}}_k\|_\infty$
5. Solve Equation 4
6. $\mathbf{m}_{k+1} = \mathbf{m}_k + \alpha\mathbf{S}^{\mathbf{H}}\delta\mathbf{x}_k$ `// update with linesearch`
7. **end**

All inversions based on our frequency-domain methods start from 5 Hz with a heavily smoothed velocity model without lateral variations (Figure 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 3) for each frequency band. For each modified Gauss-Newton subproblem, we use roughly 10 iterations of SPGℓ_1 (van den Berg and Friedlander, 2008). Figure 7 contains our results for the inverted velocity by the different algorithms, namely unconstrained **LS** is plotted in Figure 7c; constrained **LS**ℓ_1 with correct sparsity constraint level as in Figure 7e and wrong sparsity constraint level as in Figure 7d; our modified Gauss-Newton method is plotted in Figure 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 7f. As before, replacing the ℓ_1-norm constraint by a ℓ_2-norm constraint leads to noisy and inferior results (Figure 7g). These observations are also reflected in the behavior of the relative data error (plotted in Figure 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 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 9.

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 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 10a, are included in Figure 10b for unconstrained Gauss-Newton; in Figure 10c for modified Gauss-Newton with ℓ_2; and in Figure 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, 2006) 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.

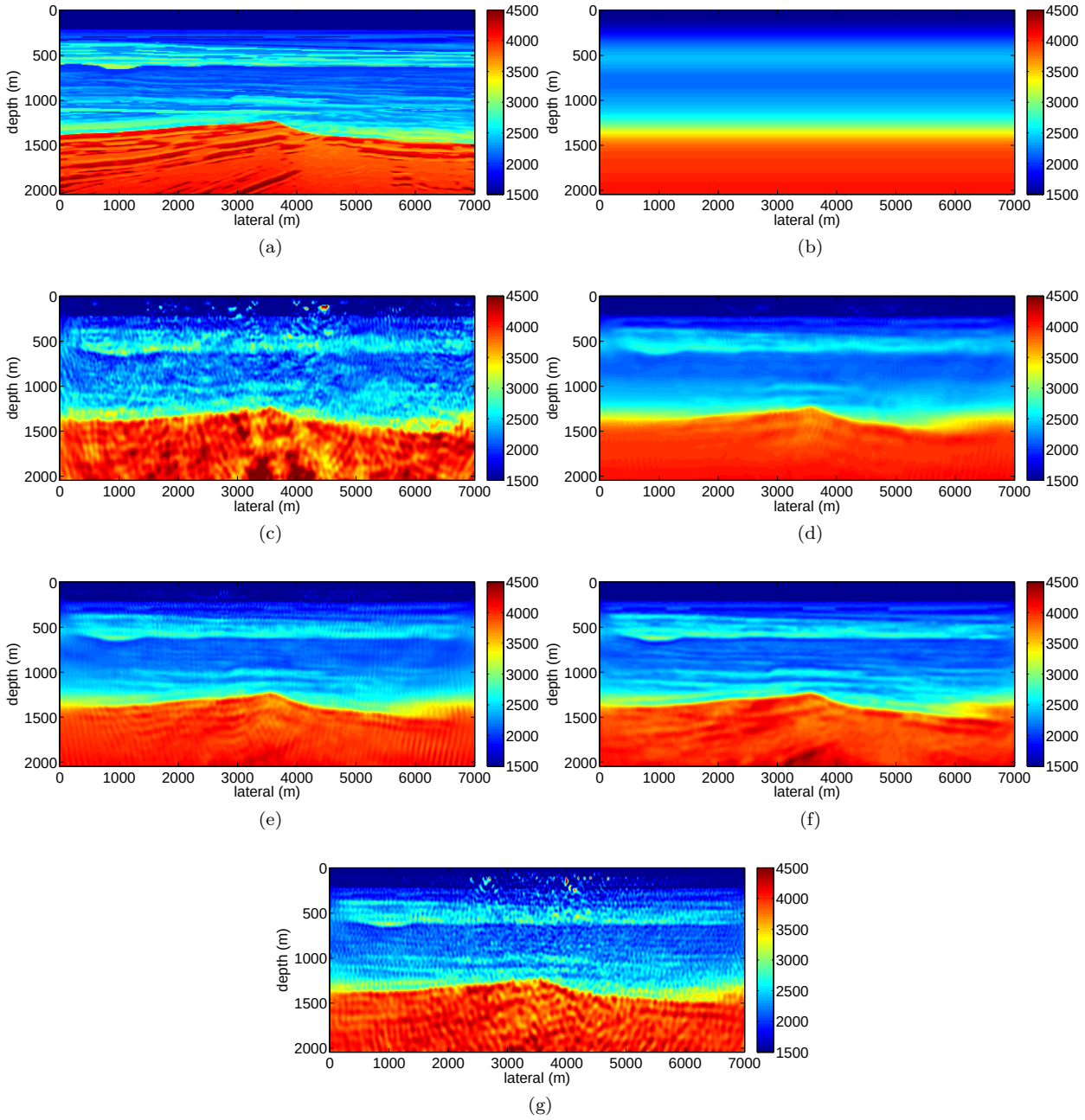
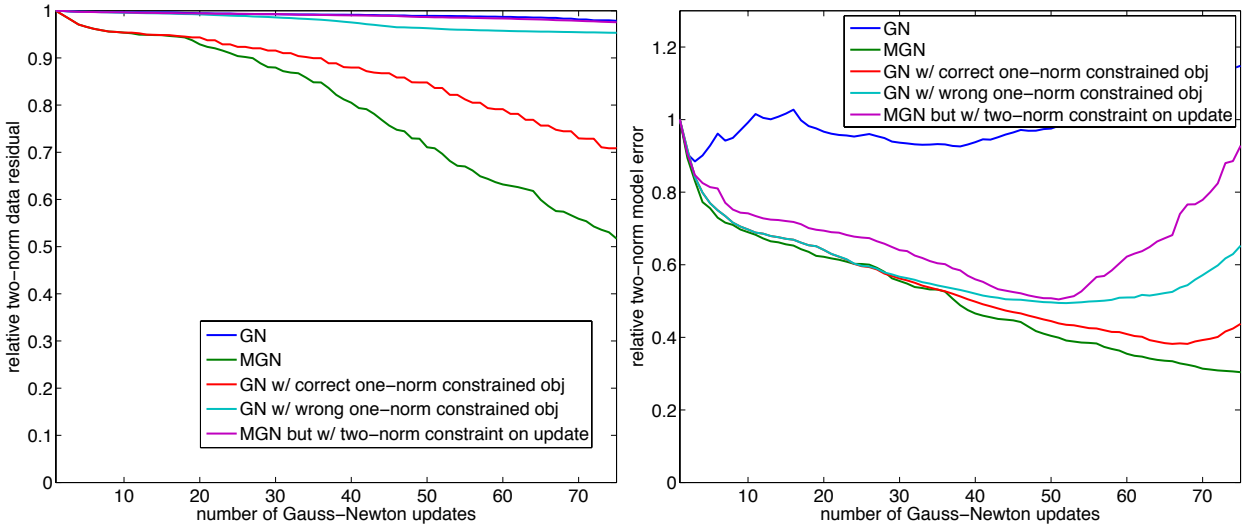


Figure 7: FWI result from BG COMPASS model data set: (a) truth model used to generate observed data; (b) starting model for FWI; (c) Gauss-Newton result with unconstrained objective function; (d) Gauss-Newton result with ℓ_1 constrained objective function; ($\tau < \tau_{true}$); (e) Gauss-Newton result with ℓ_1 constrained objective function; ($\tau = \tau_{true}$); (f) modified Gauss-Newton result with ℓ_1 constraint on the updates; (g) same as (f) but with ℓ_2 constraint on the updates.



(a) (b)

Figure 8: Data misfit and relative model error for BG model example: (a) data fitting residual; (b) relative model error.

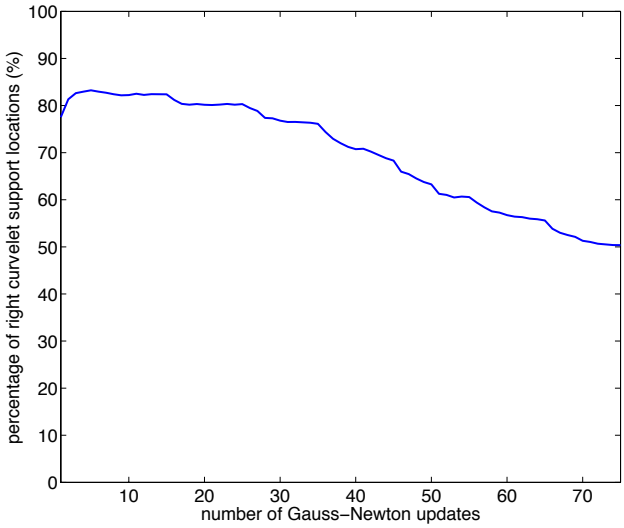


Figure 9: Percentage of curvelet coefficients that are at the right support positions.

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 10 are important for the following reasons. First, the results are obtained automatically without human intervention by running Algorithm 3 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 10a. One the other hand, imposing constraints on the norm of the curvelet coefficients of the model updates improves the inversion results (Figure 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 11.

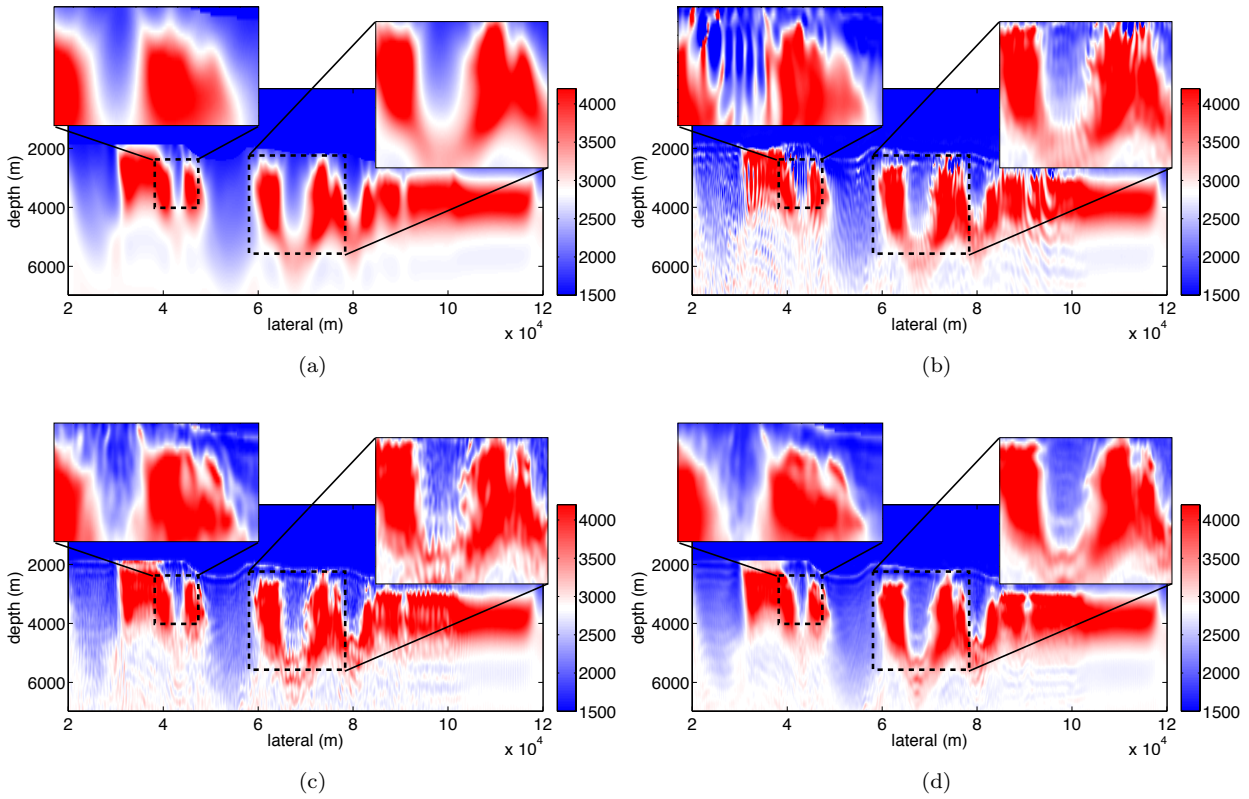
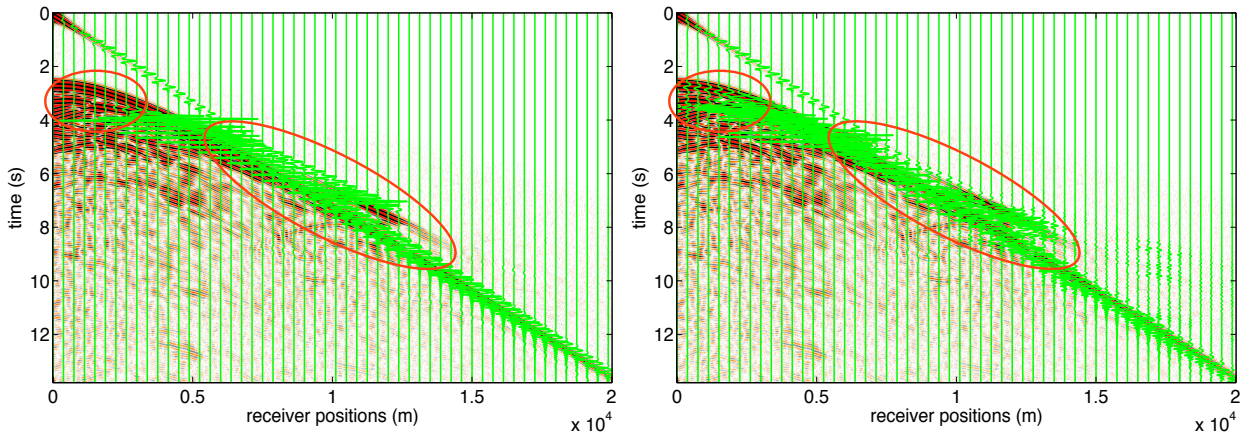


Figure 10: FWI result from Chevron Gulf of Mexico data set: (a) ray based tomography starting model for FWI; (b) inverted result with modified Gauss-Newton with ℓ_1 constraint; (c) inverted result with modified Gauss-Newton with ℓ_2 constraint. (d) smoothed (c).

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 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



(a) (b)

Figure 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.

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 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 Golf 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.

Acknowledgements

We would like to recognize the extraordinary contributions Ernie Esser made to the scientific accomplishments of this paper. Unfortunately, Ernie was not able to see the final product. We will miss him dearly. We are also grateful to Charles Jones from BG for providing us with the BG Compass velocity model. We would also like to thank the authors of SPOT, SPGℓ_1, and CurveLab. This work was in part financially supported by the Natural Sciences and Engineering Research Council of Canada 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 Corporation, Petrobras, PGS, Sub Salt Solutions, WesternGeco, and Woodside.

References

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

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.

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.

Bunks, C., F. Saleck, S. Zaleski, and G. Chavent, 1995, Multiscale seismic waveform inversion: Geophysics, **60**, 1457–1473.

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

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.

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.

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

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

Gratton, S., A. S. Lawless, and N. K. Nichols, 2007, Approximate gauss–newton methods for nonlinear least squares problems: SIAM, **18**, no. 1, 106–132.

Hansen, P. C., 1998, Rank-deficient and discrete ill-posed problems: numerical aspects of linear inversion: Siam, **4**.

Hennenfent, G., and F. J. Herrmann, 2006, Seismic denoising with nonuniformly sampled curvelets: Computing in Science & Engineering, **8**, 16–25.

Hennenfent, G., E. van den Berg, M. P. Friedlander, and F. J. Herrmann, 2008, New insights into one-norm solvers from the Pareto curve: Geophysics, **73**, A23–A26.

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.

Herrmann, F. J., and X. Li, 2012, Efficient least-squares imaging with sparsity promotion and compressive sensing: Geophysical Prospecting, **60**, 696–712.

Herrmann, F. J., X. Li, A. Y. Aravkin, and T. van Leeuwen, 2011, A modified, sparsity promoting, Gauss-Newton algorithm for seismic waveform inversion: Presented at the Proc. SPIE.

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.

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

Kumar, V., 2009, Incoherent noise suppression and deconvolution using curvelet-domain sparsity: masters-masters, University of British Columbia.

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.

Li, X., and F. J. Herrmann, 2010, Full-waveform inversion from compressively recovered model updates: , SEG, 1029–1033.

Lin, T. T., and F. J. Herrmann, 2013, Robust estimation of primaries by sparse inversion via one-norm minimization: Geophysics, **78**, R133–R150.

M, W., R. A, N. T, and et al., 2013, Anisotropic 3d full-waveform inversion: Geophysics, **78**, R59–R80.

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

Métivier, L., R. Brossier, J. Virieux, and S. Operto., 2013, Full waveform inversion and the truncated newton

method: SIAM, **35**, B401–B437.
Nocedal, J., and S. J. Wright, 2006, Least-squares problems: Springer.
Pratt, R., C. Shin, and G. Hicks, 1998a, Gauss-Newton and full Newton methods in frequency-space waveform inversion: Geophysical Journal International, **133**, 341–362.
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**, 341Ð362.
Ruszczyński, A., 2006, Nonlinear optimization: Princeton University Press. Nonlinear optimization, No. v. 13.
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.
Tarantola, A., 1984, Inversion of seismic reflection data in the acoustic approximation: Geophysics, **49**, 1259–1266.
———, 1986, A strategy for nonlinear elastic inversion of seismic reflection data: Geophysics, **51**, 1893–1903.
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.
van den Berg, E., and M. P. Friedlander, 2008, Probing the pareto frontier for basis pursuit solutions: SIAM Journal on Scientific Computing, **31**, 890–912.
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.
Virieux, J., and S. Operto, 2009, An overview of full-waveform inversion in exploration geophysics: Geophysics, **74**, 127–152.
Vogel, C. R., and M. E. Oman, 1996, Iterative methods for total variation denoising: SIAM J. Sci. Comput., **17**, no. 1, 227–238.