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

Felix J. Herrmann, Xiang Li, Aleksandr Y. Aravkin and Tristan van Leeuwen
Dept. of Earth and Ocean sciences, Universirty of British columbia

ABSTRACT

Images obtained from seismic data are used by the oil and gas industry for geophysical exploration. Cutting-edge methods for transforming the data into interpretable images are moving away from linear approximations and high-frequency asymptotics towards Full Waveform Inversion (FWI), a nonlinear data-fitting procedure based on full data modeling using the wave-equation. The size of the problem, the nonlinearity of the forward model, and ill-posedness of the formulation all contribute to a pressing need for fast algorithms and novel regularization techniques to speed up and improve inversion results.

In this paper, we design a modified Gauss-Newton algorithm to solve the PDE-constrained optimization problem using ideas from stochastic optimization and compressive sensing. More specifically, we replace the Gauss-Newton subproblems by randomly subsampled, ℓ_1 regularized subproblems. This allows us us significantly reduce the computational cost of calculating the updates and exploit the compressibility of wavefields in Curvelets.

We explain the relationships and connections between the new method and stochastic optimization and compressive sensing (CS), and demonstrate the efficacy of the new method on a large-scale synthetic seismic example.

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 $[5 - 100]\,\text{Hz}$. Current acquisition practice is moving towards recording 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 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 (1984), 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 (1984). 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 (2006) into our framework that offer a compressible representation for wavefields, which improves FWI.

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:

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

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 $\boldsymbol{\mathcal{F}}[\mathbf{m}; \mathbf{q}_i]$ is the forward operator for the i^{th} source. The vector of unknown medium parameters is denoted by $\mathbf{m}$. The forward operator $\boldsymbol{\mathcal{F}}$ acts linearly on the sources $\mathbf{q}_i$; that is

$$\boldsymbol{\mathcal{F}}[\mathbf{m}; a\mathbf{q}_i + b\mathbf{q}_j] = a\boldsymbol{\mathcal{F}}[\mathbf{m}; \mathbf{q}_i] + b\boldsymbol{\mathcal{F}}[\mathbf{m}; \mathbf{q}_j] \ . \tag{2}$$

Formulation (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 (1) as

$$\phi(\mathbf{m}) = \frac{1}{2} \|D - \boldsymbol{\mathcal{F}}[\mathbf{m}; Q]\|_F^2 \ , \tag{3}$$

where $\|\cdot\|_F$ is the Frobenius norm.
We will pause to make several important observations about (3). First, the forward operator $\boldsymbol{\mathcal{F}}$ 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 (3) is nonlinear and non-convex. It is, however, convex-composite, meaning that we can write

$$\phi(\mathbf{m}) = \rho(\boldsymbol{\mathcal{G}}(\mathbf{m})), \tag{4}$$

with ρ convex, and $\boldsymbol{\mathcal{G}}$ smooth (differentiable). Here, $\rho(X) = \frac{1}{2}\|X\|_F$, and $\boldsymbol{\mathcal{G}}(\mathbf{m}) = D - \boldsymbol{\mathcal{F}}[\mathbf{m}; Q]$. This structure allows for natural design and analysis of algorithms to solve (3). The natural approach to FWI motivated by this structure is the Gauss-Newton method, which involves iterative linearization of $\boldsymbol{\mathcal{G}}(\mathbf{m})$ and solution of least-squares problems of the form

$$\underset{\delta\mathbf{m}}{\text{minimize}} \|\delta D - \nabla\boldsymbol{\mathcal{F}}[\mathbf{m}, Q]\delta\mathbf{m}\|_F^2 , \tag{5}$$

where $\delta D = D - \boldsymbol{\mathcal{F}}[\mathbf{m}, Q]$ and $\nabla\boldsymbol{\mathcal{F}}$ is the "Jacobian tensor" of $\boldsymbol{\mathcal{F}}$, 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.

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. (2009); Moghaddam and Herrmann (2010); Boonyasiriwat and Schuster (2010); Li and Herrmann (2010); Haber et al. (2010). We follow the same approach, but focus the exposition on the Gauss-Newton subproblem, setting the stage for further modifications. We replace (5) by

$$\underset{\delta\mathbf{m}}{\text{minimize}} \|\delta DW - \nabla\boldsymbol{\mathcal{F}}[\mathbf{m}, QW]\delta\mathbf{m}\|_F^2 , \tag{6}$$

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:

$$\widetilde{\phi}(\mathbf{m}; W) = \frac{1}{2}\|DW - \boldsymbol{\mathcal{F}}[\mathbf{m}; QW]\|_F^2. \tag{7}$$

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

$$\mathsf{E}\{\widetilde{\phi}(\mathbf{m}; W)\} = \phi(\mathbf{m}). \tag{8}$$

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. Felix J. Herrmann (2011); Herrmann and Li. (2011) 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 (2006). 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:

$$\underset{\delta\mathbf{x}}{\text{minimize}}\, \|\delta D W - \nabla \boldsymbol{\mathcal{F}}[\mathbf{m}, QW] C^H \delta\mathbf{x}\|_F^2 \quad \text{s.t.} \quad \|\delta\mathbf{x}\|_1 \leq \tau, \tag{9}$$

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 (9) is a way to regularize the GN-subproblem to take advantage of wavefield sparsity. Because of the convex composite structure of (7), the solution to (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.

Outline of the paper

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 and . In section , 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 . 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 , and conclusions follow in section .

STOCHASTIC OPTIMIZATION

Stochastic optimization deals with optimization problems of the form

$$\underset{\mathbf{m}}{\text{minimize}} \big\{ \phi(\mathbf{m}) = \mathsf{E}\{\widetilde{\phi}(\mathbf{m}; W)\} \big\}. \tag{10}$$

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).

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. (2009). 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.

Stochastic Approximation

A second alternative is to apply specialized stochastic optimization methods to problem (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 ($K = \mathcal{O}(1)$), 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.

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

$$\underset{\mathbf{x}}{\text{minimize}} \|\mathbf{x}\|_0 \quad \text{s.t.} \quad A\mathbf{x} = \mathbf{b}, \tag{11}$$

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

$$\underset{\mathbf{x}}{\text{minimize}} \|\mathbf{x}\|_1 \quad \text{s.t.} \quad A\mathbf{x} = \mathbf{b}\ . \tag{12}$$

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^HA), 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

$$\underset{\mathbf{x}}{\text{minimize}} \|\mathbf{x}\|_1 \quad \text{s.t.} \quad \|A\mathbf{x} - \mathbf{b}\|_2 \leq \sigma\ , \tag{13}$$

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

$$\underset{\mathbf{x}}{\text{minimize}} \|A\mathbf{x} - \mathbf{b}\|_2^2 + \lambda\|\mathbf{x}\|_1 \tag{14}$$

and

$$\underset{\mathbf{x}}{\text{minimize}} \|A\mathbf{x} - \mathbf{b}\|_2^2 \quad \text{s.t.} \quad \|\mathbf{x}\|_1 \leq \tau \tag{15}$$

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 (13, 14, 15) all coincide. However, the values of these parameters are not known ahead of time. Therefore, most algorithms that solve (13) do some sort of continuation either in λ (see Kim et al. (2007)) or in τ (see van den Berg and Friedlander (2008)). 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 (13), the LASSO-based approach offers two advantages: the Spectral Projected Gradient method can be used to quickly solve (15) for very large linear systems, and the continuation in τ can be naturally derived using the graph of the value function for (15). The SPG method for (15) is detailed in ((van den Berg and Friedlander, 2008,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. (2008), we use it as a subroutine in the current convex-composite formulation to solve the modified Gauss-Newton LASSO subproblems.

MODIFIED GAUSS-NEWTON METHOD FOR SAA APPROACH

Recall the dimensionality reduced misfit function defined in (7):

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

This problem has the same convex composite structure (see (4)) as the full misfit, and we exploit this structure to design an algorithm for solving (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

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

the quantity $\delta D^k W$ is the dimensionality-reduced linearized data residual $DW - \boldsymbol{\mathcal{F}}[\mathbf{m}^k; QW]$, 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 $\nabla\boldsymbol{\mathcal{F}}[\mathbf{m}^k; QW]^H \nabla\boldsymbol{\mathcal{F}}[\mathbf{m}^k; QW]$ 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

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

  where B is a positive-definite scaling matrix and L is a discrete LaplacianSymes (2008); Herrmann et al. (2008, 2009). 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, $\nabla\phi = \nabla\boldsymbol{\mathcal{F}}^H\delta D$, can be computed by correlating two wavefields (see (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

$$\delta\mathbf{m} = -(\nabla\boldsymbol{\mathcal{F}}^H[\mathbf{m}^k;Q]\nabla\boldsymbol{\mathcal{F}}[\mathbf{m}^k;Q])^{-1}\nabla\phi. \tag{16}$$

  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 (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

$$\begin{aligned} \delta\mathbf{x}^k &= \arg\min_{\delta\mathbf{x}} \|\delta D^k W - \nabla\boldsymbol{\mathcal{F}}[\mathbf{m}^k;QW]C^H\delta\mathbf{x}\|_F^2 \quad \text{s.t.} \quad \|\delta\mathbf{x}\|_1 \le \tau_k \\ \delta\mathbf{m}^k &= C^H\delta\mathbf{x}^k\ , \end{aligned} \tag{17}$$

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 (17):

$$\begin{aligned} v_k(\tau_k) &= \min_{\delta\mathbf{x}} \|\delta D^k W - \nabla\boldsymbol{\mathcal{F}}[\mathbf{m}^k;QW]C^H\delta\mathbf{x}\|_F^2 \quad \text{s.t.} \quad \|\delta\mathbf{x}\|_1 \le \tau_k \\ &= \|\delta D^k W - \nabla\boldsymbol{\mathcal{F}}[\mathbf{m}^k;QW]\delta\mathbf{m}^k\|_F^2\ . \end{aligned} \tag{18}$$

The value function is carefully studied in van den Berg and Friedlander (2008) , 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 (17) has a corresponding BPDN problem (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 .

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 $\nabla\boldsymbol{\mathcal{F}}$ in the seismic application. Nonetheless, the point here is that the LASSO problem (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 $\rho(\boldsymbol{\mathcal{G}}(\mathbf{m}))$, the directional derivative $\rho'(\mathbf{m}, \delta\mathbf{x})$ exists and satisfies

$$\rho'(\mathbf{m}; \delta\mathbf{x}) \leq \rho(\boldsymbol{\mathcal{G}}(\mathbf{m}) + \nabla\boldsymbol{\mathcal{G}}^H \delta\mathbf{x}) - \rho(\boldsymbol{\mathcal{G}}(\mathbf{m})) \tag{19}$$

from (Burke, 1990,Lemma 1.3.1). Because the subproblem (17) minimizes $\rho(\boldsymbol{\mathcal{G}}(\mathbf{m})+\nabla\boldsymbol{\mathcal{G}}^H\delta\mathbf{x})$ over the one norm ball of radius τ_k, we know that $\delta\mathbf{m}^k$ satisfies

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

with equality only if we have stationarity. Combining this with (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 1. If we make the additional assumptions

1. The sequence $\{\tau_k\}$ is bounded, and
2. $\nabla\boldsymbol{\mathcal{F}}[\mathbf{m}; QW]$ 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 1.

Algorithm 1 GN-Method for FWI with Sparse Updates

initialize $\mathbf{m}$, $k \leftarrow 0$, $\Delta_1 \leftarrow 1$, ϵ, c.
while $\Delta_k > \epsilon$ **do**
  $k \leftarrow k + 1$
  Compute residual $\delta D^k W = D^k W - \boldsymbol{\mathcal{F}}[\mathbf{m}^k; QW]$
  $\delta\mathbf{x}^k \leftarrow \arg\min_{\delta\mathbf{x}} \left\{ \begin{array}{l} \|\delta D^k W - \nabla\boldsymbol{\mathcal{F}}[\mathbf{m}^k; QW]C^T\delta\mathbf{x}\|_F^2 \\ \text{s.t.} \quad \|\delta\mathbf{x}\|_1 \leq \tau_k \end{array} \right\}$
  $\delta\mathbf{m}^k = C^T\delta\mathbf{x}^k$
  $\Delta_k = \|\delta D^k W - \nabla\boldsymbol{\mathcal{F}}[\mathbf{m}^k; Q]\delta\mathbf{m}^k\|_F^2 - \|\delta D^k W\|_F^2$
  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

PRACTICAL IMPLEMENTATION

Modified Gauss-Newton approach

We propose slight modifications to Algorithm 1. 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 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:

$$\tau_k \approx (\alpha - 1)v_k(0)/v_k'(0) \ . \tag{20}$$

A closed-form expression for v' is computed in (van den Berg and Friedlander, 2008,Theorem 2.1); in our context (20) is given by

$$v_k'(0) = \frac{(\alpha - 1)\|\delta D^k W^k\|_2}{\|C\nabla\boldsymbol{\mathcal{F}}[\mathbf{m}^k; W^k]^H \delta D^k W^k\|_\infty} \ . \tag{21}$$

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 2.

Algorithm 2 GN-Method for FWI in Practice

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

Modeling operator

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

$$H[\omega; \mathbf{m}]U(\omega) = Q(\omega), \tag{22}$$

where H is a 9-point, mixed-grid discretizationJo et al. (1996) 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:

$$\text{Solve} \quad H[\omega, \mathbf{m}]U(\omega) = Q \tag{23}$$

$$\text{Solve} \quad H[\omega, \mathbf{m}]^H \delta U(\omega) = \omega^2 \mathsf{diag}(\delta\mathbf{m})U(\omega)\ . \tag{24}$$

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

$$\text{Solve} \quad H[\omega, \mathbf{m}]U(\omega) = Q \tag{25}$$

$$\text{Solve} \quad H[\omega, \mathbf{m}]^H V(\omega) = P^H \delta D(\omega) \tag{26}$$

$$\text{Compute} \quad \delta\mathbf{m} = \sum_\omega \omega^2 \mathsf{diag}(UV^H)\ . \tag{27}$$

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

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.

RESULTS

We test the proposed method on a part of the BG-Compass synthetic benchmark model. The velocity is depicted in figure . 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 . 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 . 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.

DISCUSSION

In this paper, we designed a modified Gauss-Newton algorithm for seismic waveform inversion using ideas from stochastic optimization and compressive sensing. Stochastic opti-

mization 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 imagesHerrmann and Li. (2011) 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 multiscale 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 "curvelets 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".

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.

ACKNOWLEDGMENTS

We thank the BG Group for providing the Compass velocity model. 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, BP, Chevron, ConocoPhillips, Petrobras, Total SA, and WesternGeco.

REFERENCES

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

Beylkin, G., 1984, The inversion problem and applications of the generalized radon transform: Communications on Pure and Applied Mathetmatics, **37**, 579–599.

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

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.

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

Burke, J., 1990, Numerical optimization: Course notes for math 516 (university of washington).: http://www.math.washington.edu/ burke/crs/516/index.html.

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

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.

Candès, E. J., L. Demanet, D. L. Donoho, and L. Ying, 2006, Fast discrete curvelet transforms: Multiscale Modeling and Simulation, **5**, 861–899.

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.

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

Donoho, D. L., 2006, Compressed sensing: IEEE Transactions on Information Theory, **52**, 1289–1306.

Felix J. Herrmann, X. L., 2011, Efficient least-squares migration with sparsity promotion: Presented at the , EAGE, EAGE Technical Program Expanded Abstracts.

Haber, E., M. Chung, and F. J. Herrmann, 2010, An effective method for parameter estima-

tion with pde constraints with multiple right hand sides: Technical Report TR-2010-4, UBC-Earth and Ocean Sciences Department.

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

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**, A23–26.

Herrmann, F. J., C. R. Brown, Y. A. Erlangga, and P. P. Moghaddam, 2009, Curvelet-based migration preconditioning and scaling: Geophysics, **74**, A41.

Herrmann, F. J., and X. Li., 2011, Efficient least-squares imaging with sparsity promotion and compressive sensing.: Tech. rep., University of British Columbia, University of British Columbia, Vancouver.

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

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

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.

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

Lailly, P., 1983, The seismic inverse problem as a sequence of before stack migrations: Proc. Conf. Inverse Scattering, Theory and Applications.

Li, X., and F. J. Herrmann, 2010, Full waveform inversion from compressively recovered model updates: SEG Expanded Abstracts, **29**, 1029–1033.

Moghaddam, P. P., and F. J. Herrmann, 2010, Randomized full-waveform inversion: a dimenstionality-reduction approach: SEG Technical Program Expanded Abstracts, **29**, 977–982.

Nemirovski, A., A. Juditsky, G. Lan, and A. Shapiro, 2009, Robust stochastic approximation approach to stochastic programming: Siam J. Optim., **19**, 1574–1609.

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.

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.

Shapiro, A., 2003, Monte carlo sampling methods, *in* Stochastic Programming, Volume 10 of Handbooks in Operation Research ad Management Science: North-Holland.

Shapiro, A., and A. Nemirovsky, 2005, On complexity of stochastic programming problems, *in* Continuous Optimization: Current Trends and Applications: Springer, New York.

Smith, H. F., 1998, A Hardy space for Fourier integral operators.: J. Geom. Anal., **8**, 629–653.

Stolk, C. C., and W. W. Symes, 2003, Smooth objective functionals for seismic velocity inversion: Inverse Problems, **19**, 73–89.

Symes, W., 2008, Approximate linearized inversion by optimal scaling of prestack depth migration: Geophysics, **73**, R23–R35.

Tarantola, A., 1984, Inversion of seismic reflection data in the acoustic approximation: Geophysics, **49**, 1259–1266.

ten Kroode, A., D.-J. Smit, and A. Verdel, 1998, A microlocal analysis of migration: Wave Motion, **28**, 149–172.

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.

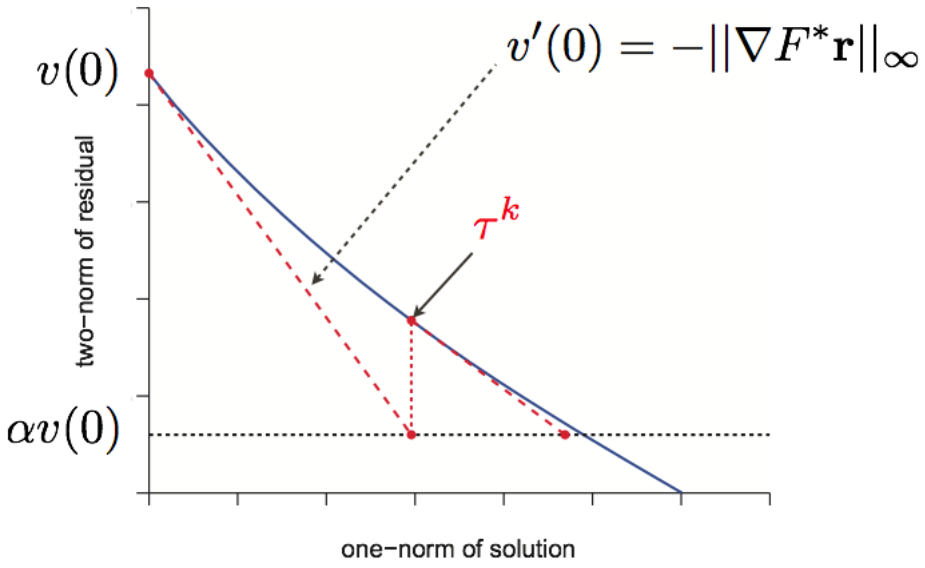

Figure 1: Schematic depiction of pareto curve used to select τ for the GN subproblems.

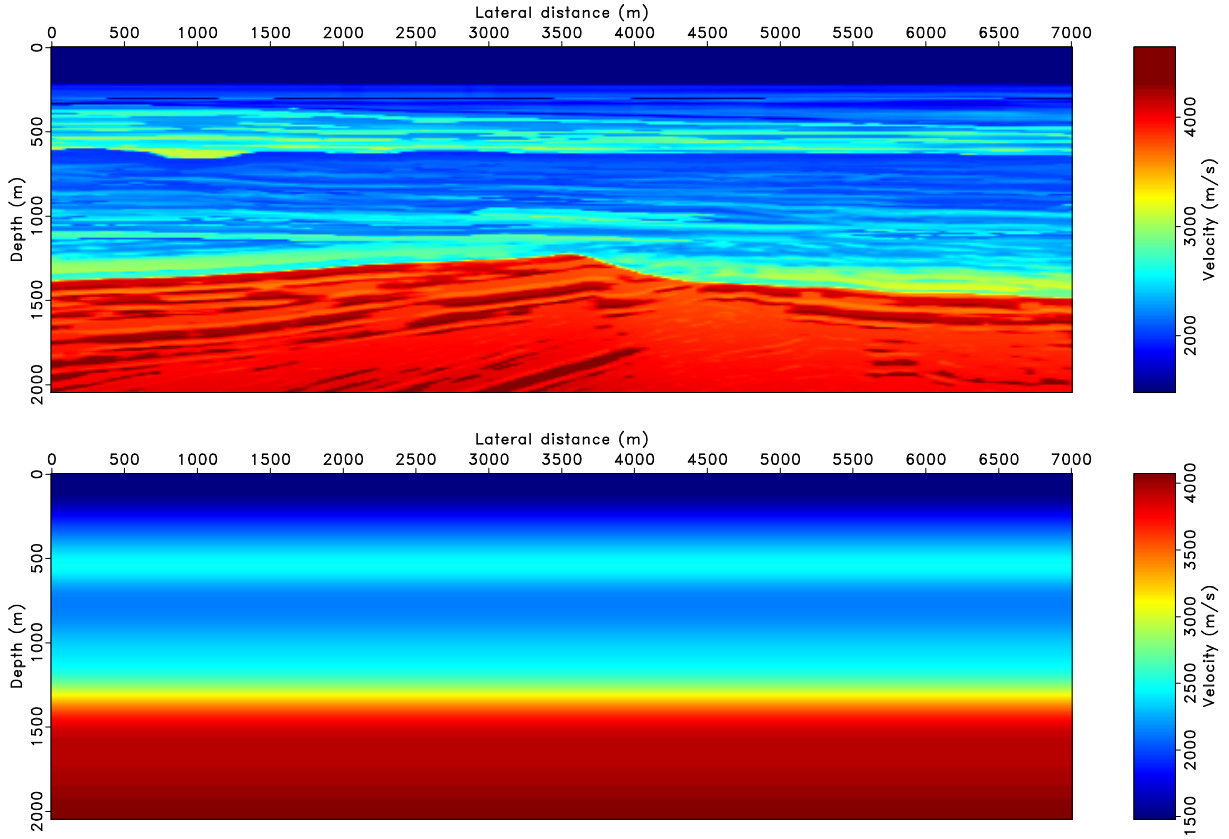

Figure 2: Compass Benchmark model with velocities ranging from 1480 - 4500 m/s (top) and initial model used for the inversion (bottom). Note the total lack of lateral variation in the initial model.

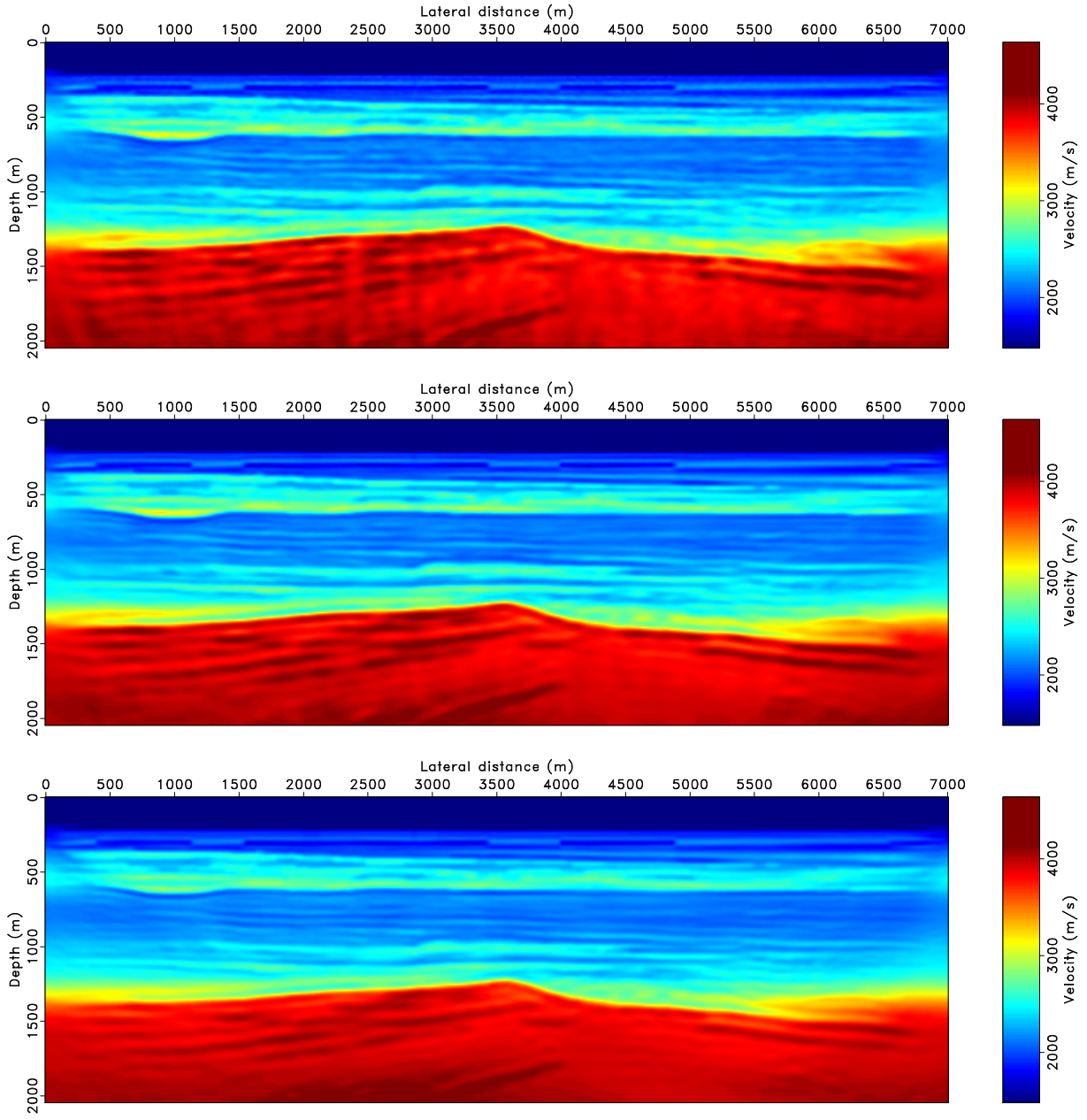

Figure 3: Inversion result for the modified Gauss-Newton algorihtm without (top) and with (middle) renewals. The result obtained with a standard Quasi-Newton approach is depicted in (c). The latter approach does not include dimensionaility reduction techniques.

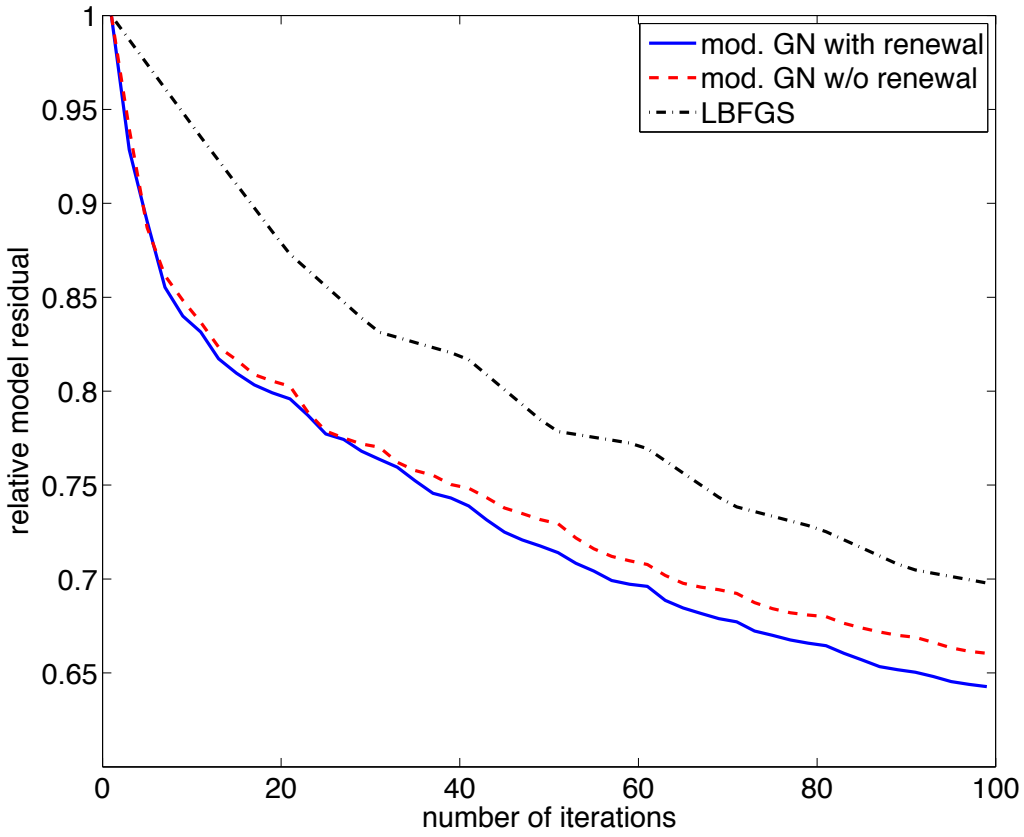

Figure 4: Convergenve in terms of the reconstruction error of the modified GN approach (with and without renewals) and the Quasi-Newton approach.