A hybrid stochastic-deterministic optimization method for waveform inversion

Tristan van Leeuwen[1], Mark Schmidt[2], Michael Friedlander[2] and Felix Herrmann[1]
[1] Dept. of Earth and Ocean sciences University of British Columbia Vancouver, BC, Canada
[2] Dept. of Computer Science University of British Columbia Vancouver, BC, Canada

January 12, 2011

Abstract

Present-day high quality 3D acquisition can give us lower frequencies and longer offsets with which to invert. However, the computational costs involved in handling this data explosion are tremendous. Therefore, recent developments in full-waveform inversion have been geared towards reducing the computational costs involved. A key aspect of several approaches that have been proposed is a dramatic reduction in the number of sources used in each iteration. A reduction in the number of sources directly translates to less PDE-solves and hence a lower computational cost. Recent attention has been drawn towards reducing the sources by randomly combining the sources in to a few supershots, but other strategies are also possible. In all cases, the full data misfit, which involves all the sequential sources, is replaced by a reduced misfit that is much cheaper to evaluate because it involves only a small number of sources (batchsize). The batchsize controls the accuracy with which the reduced misfit approximates the full misfit. The optimization of such an inaccurate, or noisy, misfit is the topic of stochastic optimization. In this paper, we propose an optimization strategy that borrows ideas from the field of stochastic optimization. The main idea is that in the early stage of the optimization, far from the true model, we do not need a very accurate misfit. The strategy consists of gradually increasing the batchsize as the iterations proceed. We test the proposed strategy on a synthetic dataset. We achieve a very reasonable inversion result at the cost of roughly 13 evaluations of the full misfit. We observe a speed-up of roughly a factor 20.

Introduction

Waveform inversion ultimately aims at producing high-quality velocity models by fitting all available seismic data in a least-squares sense (Tarantola, 1984). We denote the canonical frequency-domain waveform inversion problem as

$$\phi(m) = \sum_{i=1}^{N} \sum_{\omega} ||d_i - PH[m]^{-1} q_i||_2^2, \tag{1}$$

where d_i is a monochromatic shot record corresponding to source q_i, $H[m] = (\omega^2 m + \nabla^2)$ and P samples the wavefield at the receiver locations. While solving the Helmholtz equation can be done efficiently for multiple sources in 2D by employing an LU factorization (Marfurt, 1984), in 3D we have to rely on iterative methods (Erlangga et al., 2006). The cost of evaluating the misfit is then proportional to the number of sources and the number of frequencies. The exponential growth of the number of sources and the number of gridpoints in 3D make waveform inversion prohibitively expensive.
The idea of randomly combining shots into 'supershots' to reduce the costs of acquisition, migration or modeling has been around for quite some time (Beasley et al., 1998; Romero et al., 2000; Ikelle, 2007; Herrmann et al., 2009) and has recently found its way into waveform inversion (Krebs et al., 2009; Haber et al., 2010) (see also other contributions of the authors to these proceedings). The supershots are synthesized from the sequential shots by random superposition. The number of computations can now be significantly reduced at the cost of introducing random cross-talk. The supershots, $\overline{q}$, are related to the sequential shots by

$$\overline{q}_i = \sum_j \alpha_j^{(i)} q_i, \tag{2}$$

where the $\alpha_j^{(i)}$'s are the stacking weights. Similarly, we denote the synthesized data by $\overline{d_i}$. Krebs et al. (2009) propose to draw the weights from a pre-scribed random distribution with zero-mean and unit variance. The framework is quite general, however, and we might consider other encoding strategies. In particular, we will consider letting $\alpha_j^{(i)} = \delta_{ij}$ where i is drawn uniformly from $[1,N]$. This way we randomly select a single source. A notable advantage of this, as opposed to random encoding, is that we can apply it to incomplete data, where not all the sources are sampled by the same receivers. Other possibilities include using a plane wave synthesis for a randomly chosen slowness or using a randomly chosen eigenvector of the residual matrix (Symes, 2010). We denote the modified misfit by

$$\overline{\phi}_K(m) = \sum_{i=0}^{K} \sum_{\omega} ||\overline{d_i} - PH[m]^{-1} \overline{q}_i||_2^2, \tag{3}$$

where K is the batchsize. It is readily verified that $\overline{\phi}_K \rightarrow \phi$ as $K \rightarrow \infty$. For a fixed small batch-size $K \ll N$, the modified misfit can be seen as a 'noisy' (but unbiased) estimate of the true misfit. The optimization of such noisy misfit functions is the subject of *stochastic optimization* and many of the recent developments in randomized FWI can be traced back to this field. In particular, theoretical guarantees can be given that optimization of $\overline{\phi}_K$ will indeed converge to the minimum of ϕ. In the next section, we discuss the optimization algorithm that we use to minimize $\overline{\phi}_K$

Stochastic optimization

We discern two distinct approaches to optimize noisy misfit functions, as introduced above. The *sample average approximation* (SAA) relies on using a fixed batchsize large enough to ensure that the error $\overline{\phi}_K - \phi$ is 'small enough' (Shapiro and Nemirovsky, 2005). Then, one may use any optimization algorithm to minimize $\overline{\phi}_K$. The *stochastic approximation* (SA), on the other hand, uses only a single supershot each iteration of a steepest-descent-like algorithm but changes the supershot at each iteration (Robbins and Monro, 1951; Bertsekas and Tsitsiklis, 1996). SA has been considered for FWI by (Krebs et al., 2009; Moghaddam and Herrmann, 2010).
In the SA approach we are tied to an optimization algorithm that converges slowly, with a theoretical,

sub-linear, rate of $\mathcal{O}(1/k)$ (Nemirovski et al., 2009) (i.e, the misfit at iteration k is of order $1/k$). However, the iterations are cheap since we need only to evaluate the misfit for one source at each iteration. For the SAA approach, on the other hand, we may use any (deterministic) optimization algorithm and achieve a much faster convergence rate. The iterations are much more expensive, though. Clearly, there is a trade-off; for a given computational cost we may either do a lot of SA iterations or a few SAA iterations. For the sake of argument we will assume linear convergence: $\mathcal{O}(c^k)$ (i.e., the misfit at iteration k is of order c^k), where $0 < c \leq 1$ depends on the misfit and optimization method. We note that the convergence rates stated above are derived under particular assumptions on the misfit. We assume that the rates apply to our case if one starts close enough to the true solution. Figure 1 schematically depicts the predicted convergence as function of the computational cost. For the example, we assumed that the SAA iterations are 100 times more expensive than the SA iterations and let $c = \frac{1}{2}$. Clearly, one would prefer to use SA initially and change to SAA after some time. Alternately, we could consider gradually changing between the two methods to obtain the advantages of both methods. The hybrid we propose starts out with a small batchsize and gradually increases the batchsize as the iterations proceed. The idea is simple: far from the true model we do not need an accurate model update and we can still make progress by using only a small batch. Close to the true solution, on the other hand, we want to increase the accuracy to avoid slowing down the convergence. The pseudo-code for our hybrid algorithm, based on standard L-BFGS (cf. Nocedal and Wright, 1999, section 9.1), is given in Algorithm 1.

Algorithm 1 The algorithm has the same basic structure as a typical L-BFGS method. The function `lbfgs` applies the L-BFGS Hessian, calculated from the past n iterations, to the gradient. The linesearch ensures descent. The batchsize at iteration k is $\lfloor K_0 + \gamma k \rfloor$, with a maximum of $K_{\max}$.

while not converged **do**
  $g_k \leftarrow \nabla \overline{\phi}_K(m_k)$ — *% gradient*
  $d \leftarrow \texttt{lbfgs}(-g_k, \{m_i\}_{i=k-n}^{k-1}, \{g_i\}_{i=k-n}^{k-1})$ — *% apply L-BFGS Hessian to get search direction*
  `find` λ s.t. $\phi_K(m_k + \lambda d) < \phi_K(m_k)$ — *% approx. line search*
  $m_{k+1} = m_k + \lambda d$ — *% update model*
  $K = \min(\lfloor K_0 + \gamma k \rfloor, K_{\max})$ — *% increase batch-size*
  $k \leftarrow k+1$
end while

Results

Nonlinear migration

We consider a 'non-linear migration' problem, that is, we start waveform inversion with a very good starting model. We use the model depicted in figure 2(a) as the true model. The 'observed' data for 151 equispaced, co-located, sources and receivers are generated in the frequency domain using a 9-point discretization of the Helmholtz operator with absorbing boundaries on a grid with 10 m spacing. Free surface effects are not included. The wavelet used is a 15 Hz zero-phase Ricker wavelet. As a starting model we use the one depicted in figure 2(b). We invert all the frequencies ([5:1:25] Hz) simultaneously. For the incremental algorithm we use $K_0 = 1, \alpha = 1$. For the full inversion, we simply let $K_0 = 151$. The sources are either randomly permuted sources (i.e., $\alpha_j^{(i)} = \delta_{ij}$) or a random superpositions of the sources, using random ± 1 as the stacking weights. The results are shown in figure 2(c). With the newly proposed approach we reduce the model error significantly with much less PDE solves.

FWI

For the FWI test we use the velocity model depicted in figure 3(a). The observed data were generated with the iWAVE modeling code (developed by the TRIP consortium) for 141 sources with 50 m spacing and 281 receivers with 25 m on a grid with 5 m spacing. As a wavelet we use a 15 Hz zero-phase Ricker wavelet. Free-surface effects were not included.

We perform waveform inversion in the frequency domain with the frequency-domain modeling engine discussed in the previous example. The initial model is depicted in figure 3(b). We employ a multi-scale inversion strategy (Bunks et al., 1995), in 17 frequency bands, starting at 2.5 Hz up to 20 Hz. We window out offsets smaller than 200 m and fix the first 150 m of the model. We apply the hybrid method in each frequency band, using $K_0 = 1, \alpha = 1$. The result for sequential sources (i.e., $\alpha_j^{(i)} = \delta_{ij}$) is shown in figure 3(c). We get a very reasonable result at the cost equivalent to one evaluation of the full misfit per frequency band.Assuming that we would have needed 10 L-BFGS iterations with the full misfit for each frequency band, this would be at least a factor 10 speed-up.

Conclusion and discussion

We have proposed a hybrid stochastic-deterministic optimization method for full waveform inversion. The ultimate goal of the approach is to radically reduce the costs of full waveform inversion by decreasing the number of PDE solves needed to evaluate the misfit. To this end we introduce a *reduced* misfit that evaluates the misfit only for a small batch of sources. The sources may be either randomly synthesized supershots, synthesized plane waves, eigensources or sequential sources. In all cases, the batchsize controls the accuracy with which the reduced misfit function approximates the full misfit. The idea is to gradually increase the number of sources, or the batchsize, used for the inversion as the iterations proceed. The rationale is that far from the true model we can get away with less accuracy, while close to the solution we want better accuracy to speed up the convergence. The results show that the incremental optimization method may indeed give a much better result than the conventional approach for a fixed, relatively small, number of PDE solves.
For the non-linear migration, randomly synthesized supershots seem to yield slightly better results. However, the advantage of using randomly chosen sequential shots is that we may apply this to incomplete data, where not each shot is sampled by the same receivers. In the FWI example, we could not use the random shots because we needed to window the short offsets. This was necessary because the modeling engines used gave different near-field responses.
The need to have a complete acquisition is an important limitation of random source synthesis, and the proposed approach is one way of dealing with this limitation.

Acknowledgments
We thank the BG Group for providing the velocity model; Bill Symes for providing the iWAVE modeling code and Eldad Haber and Sasha Aravkin for numerous insightful discussion on stochastic optimization. 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

Beasley, C.J., Chambers, R.E. and Jiang, Z. [1998] A new look at simultaneous sources. *SEG Technical Program Expanded Abstracts*, **17**(1), 133–135, doi:10.1190/1.1820149.
Bertsekas, D.P. and Tsitsiklis, J. [1996] *Neuro-Dynamic Programming*. Athena Scientific.
Bunks, C., Saleck, F., Zaleski, S. and Chavent, G. [1995] Multiscale seismic waveform inversion. *Geophysics*, **60**(5), 1457–1473.
Erlangga, Y.A., Oosterlee, C.W. and Vuik, C. [2006] A novel multigrid based preconditioner for heterogeneous helmholtz problems. *SIAM Journal on Scientific Computing*, **27**(4), 1471–1492, doi:10.1137/040615195.
Haber, E., Chung, M. and Herrmann, F.J. [2010] An effective method for parameter estimation with pde constraints with multiple right hand sides. Tech. Rep. TR-2010-4, UBC-Earth and Ocean Sciences Department.
Herrmann, F.J., Erlangga, Y.A. and Lin, T. [2009] Compressive simultaneous full-waveform simulation. *Geophysics*, **74**, A35.
Ikelle, L. [2007] Coding and decoding: Seismic data modeling, acquisition and processing. *SEG Technical Program Expanded Abstracts*, **26**(1), 66–70, doi:10.1190/1.2792383.
Krebs, J.R. et al. [2009] Fast full-wavefield seismic inversion using encoded sources. *Geophysics*, **74**(6), WCC177–WCC188, doi:10.1190/1.3230502.
Marfurt, K.J. [1984] Accuracy of finite-difference and finite-element modeling of the scalar and elastic wave equations. *Geophysics*, **49**(5), 533–549, doi:10.1190/1.1441689.
Moghaddam, P.P. and Herrmann, F.J. [2010] Randomized full-waveform inversion: a dimenstionality-reduction approach. *SEG Technical Program Expanded Abstracts*, **29**(1), 977–982, doi:10.1190/1.3513940.
Nemirovski, A., Juditsky, A., Lan, G. and Shapiro, A. [2009] Robust stochastic approximation approach to stochastic programming. *Siam J. Optim.*, **19**(4), 1574–1609.
Nocedal, J. and Wright, S. [1999] *Numerical optimization*. Springer Series in Operations Research, Springer.
Robbins, H. and Monro, S. [1951] Robust stochastic approximation approach to stochastic programming. *Annals*

of Mathematical Statistics, **22**(3), 400–407.
Romero, L.A., Ghiglia, D.C., Ober, C.C. and Morton, S.A. [2000] Phase encoding of shot records in prestack migration. *Geophysics*, **65**(2), 426–436, doi:10.1190/1.1444737.
Shapiro, A. and Nemirovsky, A. [2005] *Continuous Optimization: Current Trends and Applications*, Springer, New York, chap. On Complexity of Stochastic Programming Problems.
Symes, W. [2010] Source synthesis for waveform inversion. *SEG Expanded Abstracts*, **29**(1), 1018–1022.
Tarantola, A. [1984] Inversion of seismic reflection data in the acoustic approximation. *Geophysics*, **49**(8), 1259–1266.

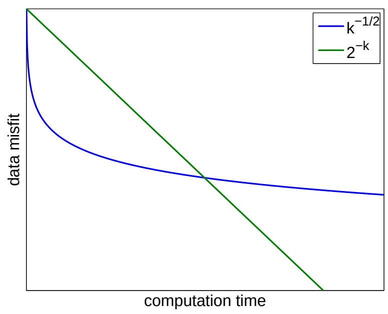
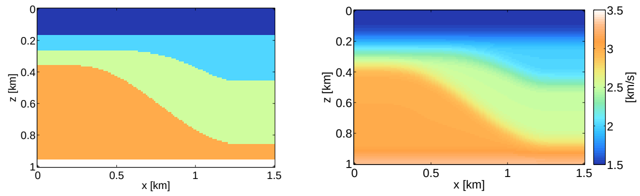
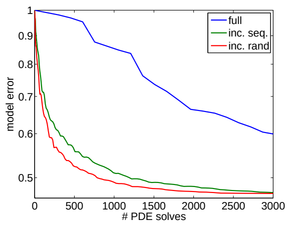


Figure 1. *Theoretical convergence rates of SA and SAA as a function of computation time, assuming that the SAA iterations are 100 times more expensive than the SA iterations.*

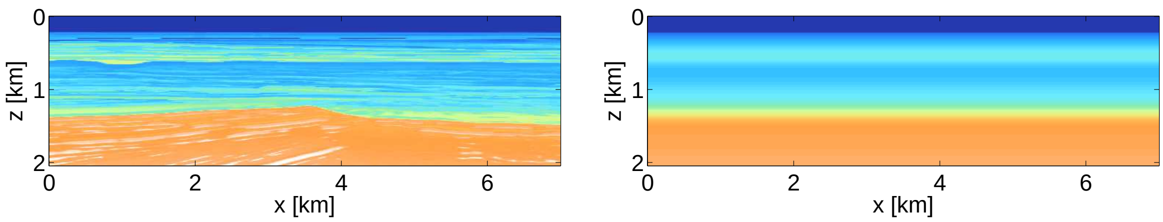


(a) (b)

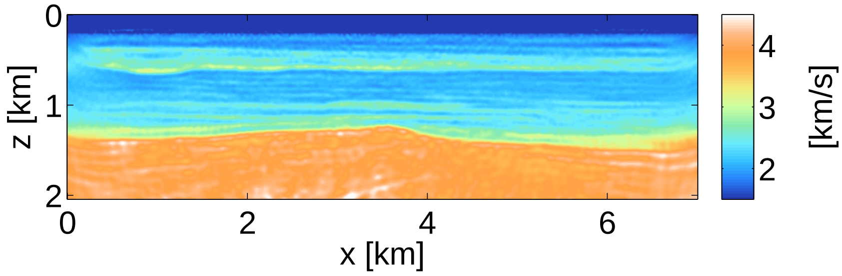


(c)

Figure 2. *(a,b) True and initial model used for nonlinear migration test. (c) Error between true and reconstructed model as a function of the number of PDE solves for different approaches: full and incremental with sequential and random sources. In this example, we obtain an error of 60 % with only 5 % of the PDE solves compared to the full optimization.*



(a) (b)



(c)

Figure 3. *(a,b) True and initial model used for FWI test. (c) Reconstructed model after multiscale FWI in 17 frequency bands from 2.5 to 20 Hz using the batching algorithm. The result was obtained at a computational cost equivalent to 1 evaluation of the full misfit per frequency band. We used different modeling engines for the synthetic and the inversion.*