Fast waveform inversion without source encoding

Tristan van Leeuwen and Felix Herrmann
Dept. of Earth and Ocean Sciences, University of British Columbia, 6339 Stores rd, Vancouver, BC, V6T1Z4

ABSTRACT

Randomized source encoding has recently been proposed as a way to dramatically reduce the costs of full waveform inversion. The main idea is to replace all sequential sources by a small number of simultaneous sources. This introduces random crosstalk in the model updates and special stochastic optimization strategies are required to deal with this. Two problems arise with this approach: *i)* source encoding can only be applied to fixed-spread acquisition setups, and *ii)* stochastic optimization methods tend to converge very slowly, relying on averaging to get rid of the cross-talk. Although the slow convergence is partly offset by the low iteration cost, we show that conventional optimization strategies are bound to outperform stochastic methods in the long run. In this paper we argue that we don't need randomized source encoding to reap the benefits of stochastic optimization and we review an optimization strategy that combines the benefits of both conventional and stochastic optimization. The method uses a gradually increasing batch of sources. Thus, iterations are very cheap initially and this allows the method to make fast progress in the beginning. As the batch size grows, the method behaves like conventional optimization, allowing for fast convergence. Numerical examples suggest that the stochastic and hybrid method perform equally well with and without source encoding and that the hybrid method outperforms both conventional and stochastic optimization. The method does not rely on source encoding techniques and can thus be applied to non fixed-spread data.

INTRODUCTION

Recently, several authors have investigated the use of simultaneous sources in waveform inversion. The main drive for this development is twofold; *i)* the exponential growth of the number of sources (and receivers) in 3D acquisition, and *ii)* the lack of a 3D modelling kernel that can efficiently deal with multiple right-hand-sides. Following earlier developments in acquisition, modelling and migration, Krebs et al. (2009) proposed to use randomly synthesized sources in FWI. The main idea is to replace all the sequential sources by a small number of synthesized simultaneous sources, and thus signicicantly reduce the computational cost per iteration. This introduces noisy cross-talk in the gradient updates and special *stochastic* optimization techniques can be used to deal with this. The assumption underlying these methods is that the cross-talk noise has zero mean. The cross-talk is then supressed during the iterations by choosing a different encoding at each iteration and averaging over past iterations. While such methods usually show fast initial progress, the need to average out the crosstalk makes then slow to convergence later on. These methods have also been reported to be more sensitive to noise in the data. The use of such stochastic

techniques in the context of waveform inversion are reviewed by Moghaddam and Herrmann (2010); van Leeuwen et al. (2011a), while Haber et al. (2010); van den Doel and Asher (2011) discuss such methods for more general inverse problems. Alternative, deterministic, ways to synthesize simultaneous sources are discussed by Symes (2010); Habashy et al. (2010).
The main problem of all these approaches is that they require fixed-spread acquisition. Also, source-dependent data-weighting and on-the-fly source-estimation are not possible. To remedy some of these shortcommings Choi and Alkhalifah (2011) and Routh et al. (2011) proposed a modified misfit criterion that is less sensitive to artifacts introduced by missing data. This criterion is essentially stackpower in the data-domain. A disadvantage of this approach is the insensitivity to amplitude information.
In this paper, we argue that we need not use simultaneous sources to reap the benefits of stochastic optimization. We can in principle use a different, randomly chosen, sequential source at each iteration. While this gets rid of the source-crosstalk, it might introduce other artifacts and we do not sample all the data efficiently. Moreover, the stochastic optimization methods still exhibit slow convergence. We review a hybrid stochastic-deterministic optimization method (Friedlander and Schmidt, 2011) that promises the same fast initial convergence as stochastic methods without the slow convergence later on. The main idea of this method is to use a slowly increasing batch of (sequential) sources. The method is very flexible and can be incorporated easily in any existing optimization code.
The paper is organized as follows. First, we review different ways of approximating the misfit using only a small number of sources. We then discuss optimization strategies that are suitable for these approximate misfits. Finally, we give some stylized numerical examples to illustrate the main ideas.

APPROXIMATING THE MISFIT

The *full* misfit is given by

$$\Phi[\mathbf{m}] = \frac{1}{N}\sum_{i=1}^{N}\phi_i[\mathbf{m}], \quad \phi_i[\mathbf{m}] = ||\mathcal{F}[\mathbf{m}]\mathbf{q}_i - \mathbf{d}_i||_2^2, \tag{1}$$

where $\mathbf{m}$ denotes the model, $\mathcal{F}[\mathbf{m}]$ the modelling operator, $\mathbf{q}_i$ the i-th source and $\mathbf{d}_i$ the corresponding shot record. This notation applies to both frequency and time domain. We'll assume that the cost of evaluating the misfit is proportional to N and aim at reducing the cost by reducing the effective N. In particular, we consider two different methods: *i)* use only a subset (batch) of the terms in the sum, *ii)* replace the sum over all the residual terms by a sum over a smaller number of randomized residuals. The quality of the approximation will be measured in terms of the error in the gradient.

Batching

This approximation to the full misfit relies on selecting a batch of sources:

$$\Phi[\mathbf{m}] \approx \Phi_B[\mathbf{m}] = \frac{1}{K}\sum_{i\in B}\phi_i[\mathbf{m}], \tag{2}$$

where B is the indexset and $K = |B|$ denotes it's size. The error depends on the strategy used to choose the batch. The *worst-case* scenario is investigated by writing the error

$\mathbf{e}_B = \nabla\Phi - \nabla\Phi_B$ explicitly as

$$\mathbf{e}_B = \left(\frac{N-K}{NK}\right)\sum_{i\in B}\nabla\phi_i + \frac{1}{N}\sum_{i\notin B}\nabla\phi_i. \tag{3}$$

This gives us the worst-case error bound:

$$||\mathbf{e}_B||_2^2 \leq 4\left(\frac{N-K}{N}\right)^2 \max_i ||\nabla\phi_i||_2^2. \tag{4}$$

The expected error is investigated by choosing the batch randomly without replacement. Friedlander and Schmidt (2011) show that

$$\mathsf{E}\{||\mathbf{e}_B||_2^2\} = \left(\frac{N-K}{N(N-1)}\right)\frac{N}{K}\max_i ||\nabla\phi_i - \nabla\Phi||_2^2 \tag{5}$$

The asymptotic behaviour of these error bounds is illustrated in figure 1 (a).

Source encoding

A simultaneous source is a weighted stack of the sources:

$$\widetilde{\mathbf{q}} = \sum_{i=1}^{N} w_i\mathbf{q}_i, \tag{6}$$

and the corresponding data is denoted by $\widetilde{\mathbf{d}}$. Such encodings with random weights have been considered for modeling (Ikelle, 2007; Herrmann et al., 2009; Neelamani et al., 2010), acquisition (Beasley et al., 1998; Berkhout, 2008), migration (Romero et al., 2000; Dai et al., 2010) and waveform inversion (Krebs et al., 2009; Moghaddam and Herrmann, 2010; Haber et al., 2010; van Leeuwen et al., 2011a). The encoded misfit is then given by

$$\widetilde{\phi}[\mathbf{m}] = ||\mathcal{F}[\mathbf{m}]\widetilde{\mathbf{q}} - \widetilde{\mathbf{d}}||_2^2 \tag{7}$$

If we choose the random weight vector $\mathbf{w} = (w_1, \ldots, w_K)^T$ such that $\mathsf{E}\{\mathbf{w}\mathbf{w}^T\} = N^{-1}$, we find:

$$\Phi[\mathbf{m}] = \mathsf{E}\left\{\widetilde{\phi}[\mathbf{m}]\right\}, \tag{8}$$

Upon replacing the expectation by a sample average we find

$$\Phi \approx \Phi_W = \frac{1}{K}\sum_{i=1}^{K}\widetilde{\phi}_i, \tag{9}$$

where the index i now runs over different realizations of the random vector $\mathbf{w}$.
As explained by Haber et al. (2010), this approximation is an instance of randomized trace estimation. The approximation error is given by:

$$||\mathbf{e}_W||_2^2 = \mathcal{O}(1/K). \tag{10}$$

The constants depend on the properties of the matrix and the random variables. A more detailed study is presented by Avron and Toledo (2010). The theoretical asymptotic behaviour of the error is illustrated in figure 1 (a). This suggests that the batching strategy provides a more efficient way to approximate the misfit with as few sources as possible. It is important to realize, however, that the constants will make a difference in the approximation error in practice. We will present some numerical examples later on to support our assertion that the batching strategy provides a more efficient way to approximate the misfit.
Interestingly enough, we may link this approach to the batching approach by taking $\mathbf{w}_i$ to be a random unit vector. This would correspond to using a single, randomly chosen source. The difference is that here we draw the sources *with* replacement, whereas in the batching case we draw *without* replacement.
Using random unit vectors would also alow us to apply these ideas to marine data. However, the random unit vectors are less efficient in the framework of trace estimation since they only sample single source, whereas other encoding strategies sample the full data (albeit at the cost of introducing the cross-talk). We illustrate this behaviour by showing $K^{-1}\sum_{i=1}^{K}\mathbf{w}_i{\mathbf{w}_i}^T$ for various batch sizes in figure 2. We want the covariance matrix to be as close to the identity matrix as possible.

OPTIMIZATION STRATEGIES

In the following exposition we assume the misfit to be of the form

$$\Phi[\mathbf{m}] = \frac{1}{K}\sum_{i=1}^{K}\phi_i[\mathbf{m}], \tag{11}$$

where ϕ_i can be either the misfit for a sequential or a simultaneous source.
The basic scheme for any iterative optimization method is

$$\mathbf{m}_{k+1} = \mathbf{m}_k + \gamma_k\mathbf{s}_k, \tag{12}$$

where γ_k is an appropriately chosen steplength and $\mathbf{s}_k$ is the search direction. An important property of an optimization method is its convergence rate. This gives a theoretical rate of decay of the misfit in terms of the iteration count. Such rates are derived under certain assumptions on the misfit. While typical assumptions (such as convexity and positive definiteness of the Hessian) are not valid in general for our problem, we may at least assume they are valid when we start 'close' to a (local) minimum.
A regular steepest descent method (i.e., $\mathbf{s}_k = -\nabla\Phi[\mathbf{m}_k]$), for example, will have a linear convergence rate:

$$||\Phi[\mathbf{m}_k] - \Phi[\mathbf{m}_*]||_2^2 = \mathcal{O}(c^k), \quad 0 \leq c < 1. \tag{13}$$

The question now is how the approximations discussed above will affect the convergence. Friedlander and Schmidt (2011) show that if $\mathbf{s}_k = -\nabla\Phi[\mathbf{m}_k] + \mathbf{e}_k$:

$$||\Phi[\mathbf{m}_k] - \Phi[\mathbf{m}_*]||_2^2 = \mathcal{O}(\max\{c^k, ||\mathbf{e}_k||^2\}), \quad 0 \leq c < 1. \tag{14}$$

This tells us that our convergence will only be as fast as we can bring down the error of our approximation. From the above discussion we may already argue that the batching

strategy is more favourable since we have explicit control over the error, whereas the random encoding strategy gives us a 'fixed' decay of $\mathcal{O}(1/K)$. We investigate different strategies in more detail below.
Again, the constants play an important role in practice. These constants rely heavily on the details of the problem and the method. Like the steepest descent method, both non-linear CG and L-BFGS also have a linear convergence rate, but are expected to perform much better in practice. We will present some numerical results in a later section to illustrate the advertised behaviour of the different optimization strategies.

Stochastic optimization

Here, the search direction is based on the negative gradient of a single (randomly chosen) term of the misfit:

$$\mathbf{s}_k = -\nabla\phi_i[\mathbf{m}_k], \tag{15}$$

Such methods are known as *stochastic gradient descent* or *incremental gradient* methods and are popular in the machine learning community (cf. Robbins and Monro, 1951; Bertsekas and Tsitsiklis, 1996; Bertsekas, 1997). There is no theory to support the use of second order information in such schemes, at least not in the sense of using the curvature related to $\nabla\phi_i$. Convergence can be proven under some assumptions on the steplengths and if we have an *unbiased estimate* of the gradient, i.e., $\mathsf{E}\{\nabla\phi_i\} = \nabla\Phi$. The convergence rate for such methods is sub-linear:

$$||\Phi[\mathbf{m}_k] - \Phi[\mathbf{m}_*]||_2^2 = \mathcal{O}(1/k). \tag{16}$$

This may be intuitively understood as follows. The error in the gradient is controlled by the number of iterations since we are implicitly summing many different randomly sampled gradients. As discussed above, the error decays as $\mathcal{O}(1/K)$, which explains the convergence rate. These type of methods can be made more stable by accumulating information on past iterations (Polyak and Juditsky, 1992). Results with such approaches are presented by (Krebs et al., 2009; Haber et al., 2010; van Leeuwen et al., 2011a).

Hybrid optimization

In this case we choose a batch of the terms in the misfit for our update (cf. Bertsekas and Tsitsiklis, 1996; Bertsekas, 1997; Friedlander and Schmidt, 2011):

$$\mathbf{s}_k = -\left(\frac{1}{|B_k|}\sum_{i\in B_k}\nabla\phi_i[\mathbf{m}_k]\right). \tag{17}$$

We construct the batch for the next iteration by adding elements to the current batch. Thus we can directly use the results discussed above to control the error in the gradient. In particular, if we choose batching strategy such that

$$\mathsf{E}\{||\mathbf{e}_k||_2^2\} = \mathcal{O}(c^k), \tag{18}$$

Friedlander and Schmidt (2011) show that the hybrid method has a linear convergence rate. In figure 1 (b) we plot the theoretical convergence rates for the conventional, stochastic and

hybrid approach as a function of the cumulative cost. The hybrid method seems to combine the best of both conventional and stochastic methods as it exhibits fast convergence initially and does not drop down to the sublinear rate of the stochastic method in the end. Some experiments with this method are also reported by van Leeuwen et al. (2011b).

Example

To illustrate the advertised behaviour of the above outlined optimization strategies we consider a stylized example based on the Marmousi model. The true and initial model are depicted in figure 3. The data are generated for 151 equispaced sources, 301 equispaced receivers and 6 frequencies between 5 and 25 Hz (irregularly sampled).
The error between the actual and sampled gradient for the initial model for different strategies is depicted in figure 4 (a). The error exhibits the predicted behaviour (cf. figure 1 (a)). The stochastic method is implemented as a steepest-descent method with constant steplength. The steplength is chosen manually by trial and error and we use either unit vectors or Gaussian weights for the encoding. Our implementation of the Hybrid method is based on L-BFGS with an Armijo line-search. The search direction at iteration k is of the form:

$$\mathbf{s}_k = -Y_k \left(\frac{1}{|B_k|} \sum_{i \in B_k} \nabla \phi_i[\mathbf{m}_k] \right), \tag{19}$$

where the 'L-BFGS inverse Hessian', Y_k, is estimated from past gradients and model-updates. Eventhough the misfit is changing from one iteration to the next, Y_k is estimated from only a limited number (four, in this case) of past iterations so we may argue that it still provides a reasonable estimate of the inverse of the Hessian. As a reference, we also apply L-BFGS to the full misfit.
The convergence in terms of the model is depicted in figure 4 (b). The convergence in terms of the misfit is depicted in figure 4 (c). We calculated these full misfits just for the purpose of making these plots since the full misfit is not calculated at every iteration in either the stochastic or hybrid approach. These plots show that both the batching approach and the stochastic-gradient descent out-perform the full approach. Surprisingly, the SGD approach does nearly as well as the batching approach, in this case. We have to remember, however, that we picked the optimal steplength by trial and error. The batching approach allows for a linesearch, and is hence preferable in practice. Also, these results show that the encoding strategy makes little difference. Representative model reconstructions, each with roughly the same reconstruction error, for each method are depicted in figure 5. The L-BFGS approach needed 100 iterations, while the both the SGD and batching approach needed only 50 iterations to reach this error level.

CONCLUSIONS

- We can use random unit vectors for source encoding and apply ideas from stochastic optimization to marine data. The numerical example indicates that unit-vector encoding can be as efficient as simultaneous source encoding for FWI. The underlying assumption is that all terms in the misfit contribute equally.

- The batching strategy is a more efficient way to approximate the misfit with a small number of terms because it does not rely on Monte Carlo sampling to reduce the error. Moreover, the batching strategy allows for preconditioning and line-search strategies, whereas the SGD approach does not.
- The incremental gradient and Hybrid approaches can be applied in many more settings such as ray-based traveltime fitting and online processing of data.

ACKNOWLEDGMENTS

The authors thank Eldad Haber, Michael Friedlander and Aleksandr Aravkin for fruitfull discussions 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

Avron, H., and S. Toledo, 2010, Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix: Submitted for publication, 1–16.

Beasley, C. J., R. E. Chambers, and Z. Jiang, 1998, A new look at simultaneous sources: SEG Technical Program Expanded Abstracts, 133–135.

Berkhout, A., 2008, Changing the mindset in seismic acquisition: The Leading Edge, **27**, 924–938.

Bertsekas, D., and J. Tsitsiklis, 1996, Neuro-dynamic programming: Athena Scientific.

Bertsekas, D. P., 1997, A new class of incremental gradient methods for least squares problems: SIAM J. on Optimization, **7**, no. 4, 913–926.

Choi, Y., and T. Alkhalifah, 2011, Application of encoded multi-source waveform inversion to marine-streamer acquisition based on the global correlation: Presented at the EAGE Expanded Abstracts.

Dai, W., C. Boonyasiriwat, and G. Schuster, 2010, 3d multi-source least-squares reverse-time migration: SEG Expanded abstracts, 66–70.

Friedlander, M., and M. Schmidt, 2011, Hybrid Deterministic-Stochastic Methods for Data Fitting: Submitted for publication.

Habashy, T. M., A. Abubakar, G. Pan, and A. Belani, 2010, Full-waveform seismic inversion using the source-receiver compression approach: SEG Technical Program Expanded Abstracts, 1023–1028.

Haber, E., M. Chung, and F. Herrmann, 2010, An effective method for parameter estimation with pde constraints with multiple right hand sides: Technical Report TR-2010-4, UBC-Earth and Ocean Sciences Department.

Herrmann, F. J., Y. A. Erlangga, and T. T. Y. Lin, 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, 66–70.

Krebs, J. R., J. E. Anderson, D. Hinkley, A. Baumstein, S. Lee, R. Neelamani, and M.-D. Lacasse, 2009, Fast full wave seismic inversion using source encoding: geophysics, **28**, 2273–2277.

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

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

Polyak, B., and A. Juditsky, 1992, Accelaration of stochastic approximation by averaging: SIAM Journal on control and optimization, **30**, 838–855.

Robbins, H., and S. Monro, 1951, A stochastic approximation method: Annals of mathematical statistics, **22**, 400–407.

Romero, L. A., D. C. Ghiglia, C. C. Ober, and S. A. Morton, 2000, Phase encoding of shot records in prestack migration: Geophysics, **65**, 426.

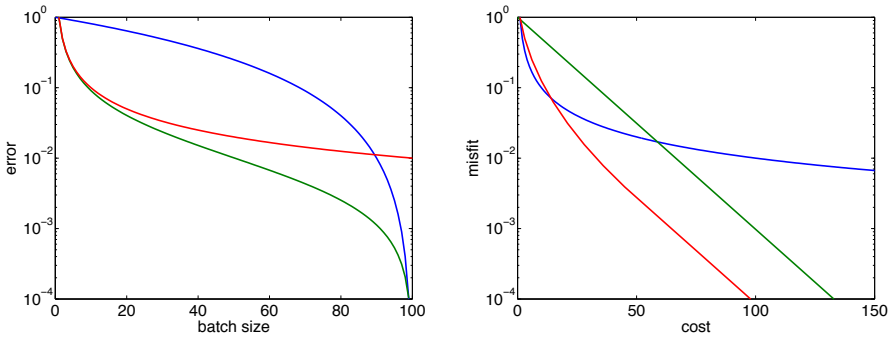
Routh, P., J. Krebs, S. Lazaratos, A. Baumstein, I. Chikichev, S. Lee, N. Downey, D. Hinkley, and J. Anderson, 2011, Full-wavefield inversion of marine streamer data with the encoded simultaneous source method: Presented at the EAGE Expanded Abstracts.

Symes, W., 2010, Source synthesis for waveform inversion: SEG Technical Program Expanded Abstracts, 1018–1022.

van den Doel, K., and U. Asher, 2011, Adaptive and stochastic algorithms for eit and dc resistivity problems with piecewise constant solutions and many measurements: submitted.

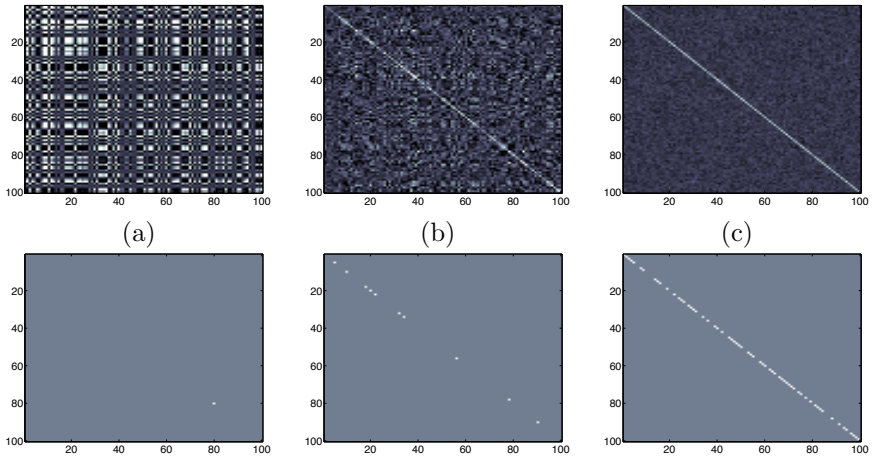
van Leeuwen, T., A. Aravkin, and F. Herrmann, 2011a, Seismic waveform inversion by stochastic optimization: Internation Journal of Geophysics.

van Leeuwen, T., M. Schmidt, F. M.P., and F. Herrmann, 2011b, A hybrid stochastic-deterministic method for waveform inversion: Presented at the EAGE Expanded Abstracts.



(a) (b)

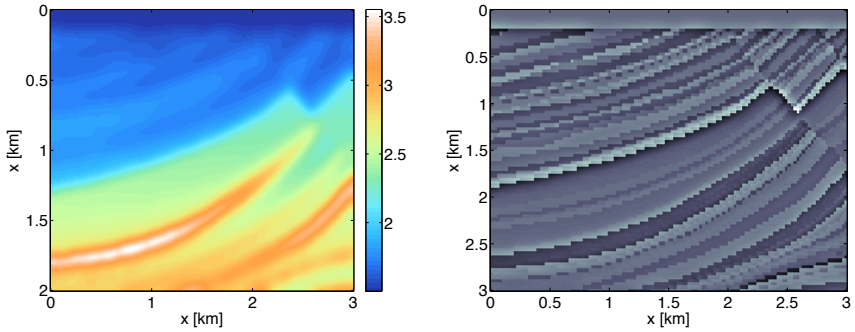
Figure 1: (a) Asymptotic behaviour of the error between the approximate and true gradient: worst-case batching (blue), average batching (green), source encoding (red). (b) Asymptotic convergence rate for different optimization strategies: conventional (green), stochastic (blue) and hybrid (red).



(a) (b) (c)

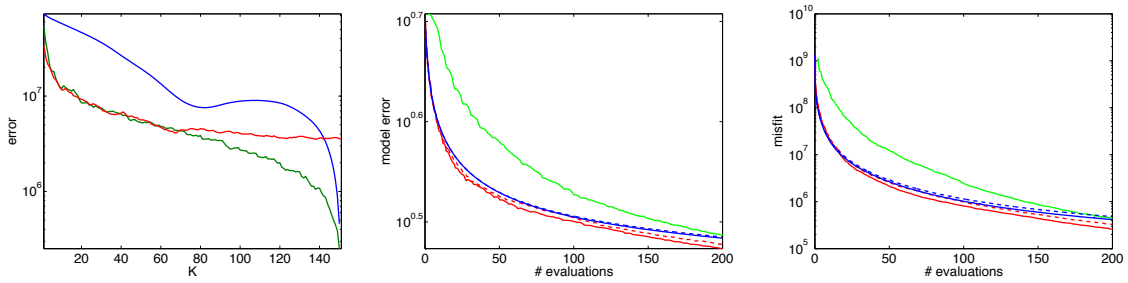
(d) (e) (f)

Figure 2: Covariance matrix: $K^{-1}\sum_{i=1}^{K}\mathbf{w}_i{\mathbf{w}_i}^T$ for random Gaussian vectors (top) and random unit vectors (bottom) for $K = \{1, 10, 100\}$ (left to right)



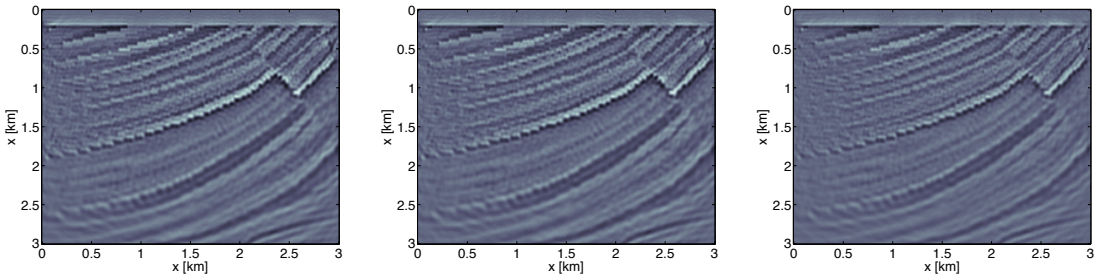
(a) (b)

Figure 3: Background model (a) and reflectivity (b) used for example.



(a) (b) (c)

Figure 4: (a) Gradient error for batching strategy with natural order (blue) and random order (green) and source encoding (red). The latter two are averages over 5 different realizations. This behaviour confirms the predicted behaviour of the error depicted in figure 1 (a). (b) model error and (c) misfit for L-BFGS (green), Hybrid (red) with unit-vector (solid) and Gaussian (dash) encoding, and SGD (blue) with unit-vector (solid) and Gaussian (dash) encoding.



(a) (b) (c)

Figure 5: reconstructions for (a) L-BFGS (approx. 100 evaluations), (b) Hybrid (approx. 50 evaluations) and (c) SGD (approx. 50 evaluations)