A stochastic L-BFGS approach for full waveform inversion

Gabriel Fabien-Ouellet, Erwan Gloaguen, Bernard Giroux, INRS*

Summary

Speeding-up convergence rates and reducing the computational burden of Full Waveform Inversion (FWI) is increasingly important as we move toward large-scale 3D multi-parameter inversion. To this end, second-order optimization algorithms like L-BFGS or the truncated Newton method allow a much faster convergence rate at minimal computational costs. In the same fashion, stochastic source subsampling approaches have been shown to reduce the computational cost of FWI. In this study, we propose to combine these two strategies and present how the L-BFGS algorithm can be used along with the stochastic source subsampling strategy, or what we call the stochastic L-BFGS algorithm.

Introduction

The advances in high performance computing over the last decades have allowed the application of FWI to larger and larger 3D problems, and to more complex physics, going from acoustic to the more complex anisotropic (visco)elastic wave propagation (Fabien-Ouellet et al., 2017, Komatitsch et al., 2002). Still, large-scale multi-parameter FWI remains computationally challenging, preventing its widespread adoption. Hence, reducing the computing times remains an important issue to broaden the applicability of FWI.

Many strategies have been proposed to decrease the computational burden of FWI. One such strategy is the use of second-order descent algorithms, like the Newton method, which has been shown to dramatically improve the convergence rate of FWI and its resolution (Pratt et al., 1998). More precisely, inexact Newton methods like the limited memory Broyden-Fletcher-Goldfarb-Shannol (L-BFGS) or the truncated Newton method retain the better convergence rates of the full Newton method, without its prohibitively high computing cost. In effect, inexact Newton methods speed-up convergence at minimal costs, thus reducing the required number of iterations and the overall computing time of FWI. In addition, inexact Newton methods are particularly important for multi-parameter FWI, as second order information helps to decouple different parameter classes (Virieux et al., 2017).

Another successful strategy to mitigate the computing requirements of FWI is to use data subsampling, be it in the form of random sources encoding (Krebs et al., 2009) or of stochastic source subsampling, two methods that show similar performance (van Leeuwen and Herrmann, 2013). These methods take their roots in the stochastic optimization theory. In stochastic optimization, the descent direction is obtained by calculating the gradient on a random subset of the data, which reduces the cost of the computation. This method is advantageous for optimization problems of large size and large datasets (Bottou, 2010), like FWI.

Most stochastic optimization algorithms used in FWI are based on first order gradient descent methods (van Leeuwen et al., 2011). As shown by Castellanos et al. (2015), the difficulty of introducing second order approximations stems from the error introduced by the random subsampling in the Hessian approximation. To be able to apply stochastic second order descent algorithms, a strategy to reduce this error must be adopted. In this study, we show how this can be achieved for the L-BFGS method. We first present the theory of L-BFGS with random sources subsampling and then show a performance comparison between the proposed algorithm and the standard stochastic descent method with the Marmousi model.

Problem definition

Full waveform inversion is formulated as a minimization problem: find the Earth parameters $\boldsymbol{m}$ that minimize a measure of the discrepancy between the modelled and the recorded data, $\boldsymbol{d}$. This measure is given by the cost function, often taken as the l^2 norm of the residuals:

$$\mathrm{X}_{\Omega}(\boldsymbol{m}) = \frac{1}{2}\frac{\sum_{\mathrm{i}\in\Omega}(\boldsymbol{S_i v_i}(\boldsymbol{m}) - \boldsymbol{d_i})^T(\boldsymbol{S_i v_i}(\boldsymbol{m}) - \boldsymbol{d_i})}{\sum_{\mathrm{i}\in\Omega}\boldsymbol{d_i}^T\boldsymbol{d_i}} \qquad (1)$$

where Ω represents a source ensemble and $\boldsymbol{v_i}$ is the modelled particle velocities due to source i, sampled at the recorder's location by the sampling operator $\boldsymbol{S}$. The cost function is normalized by the sum of the squared amplitude of the data to scale appropriately for different sources subset. In what follows, X_{Ω} will be used to designate the cost function on a source subset Ω and X will be used for the cost function on the complete set of sources.

The particle velocities $\boldsymbol{v_i}$ obey the wave equation, which must be solved numerically for an arbitrarily heterogeneous Earth (Virieux et al., 2011). Solving the wave equation, or what we call forward modeling, represents the main computational cost of FWI. Time-domain finite differences (FDTD) remains the method of choice for large 3D elastic FWI. FDTD requires one complete forward modeling per source, which is why reducing the number of modelled shot points during inversion is advantageous. As solving the wave equation remains challenging even with the computational resources of today, the cost function is usually optimized with local line search algorithms of the following form:

$$\boldsymbol{m}_{k+1} = \boldsymbol{m}_k - \alpha \boldsymbol{H}_k \nabla \mathrm{X} \qquad (2)$$

where ∇X is the misfit gradient, calculated at the cost of approximately two forward modelling per source owing to the adjoint method (Plessix, 2006), α is the step size and $\boldsymbol{H}_k$ is an approximation of the inverse Hessian matrix, i.e. $(\boldsymbol{H}_k \approx \nabla^2 \mathrm{X}^{-1})$. For the simplest line search algorithm, the steepest descent or gradient descent, the approximation of the inverse Hessian is discarded, i.e. $\boldsymbol{H}_k \rightarrow \boldsymbol{I}$. Although this approach has the merit of being the most parsimonious in terms of forward modelling, it suffers from slow convergence.

Brossier et al. (2009) show that a much better convergence can be attained at no additional forward modelling costs with the L-BFGS algorithm (Nocedal and Wright, 2006). Furthermore, this method does not require the complete storage of $\boldsymbol{H}_k$, which is prohibitive for large models. Instead, the product $\boldsymbol{H}_k \nabla \mathrm{X}$ can be computed by storing n vector pairs of parameter changes, $\boldsymbol{s}_k = \boldsymbol{m}_{k+1} - \boldsymbol{m}_k$, and gradient changes, $\boldsymbol{y}_k = \nabla X_{k+1} - \nabla X_k$. The inverse Hessian preconditioning is then obtained with the two-loops recursion (Algorithm 1), which involves only vector products. This has a negligible cost compared to the forward/adjoint modelling. To ensure that the approximate inverse Hessian remains positive definite and that the step direction remains a descent direction, the step length α is chosen to respect the Wolfe conditions, namely sufficient decrease of the cost function and its curvature. A simple line search implementing those two conditions is presented in Algorithm 2. The sufficient decrease and curvature conditions appear at lines 5 and 7 respectively.

Algorithm 1: Two-loops recursion

```
Inputs: $\nabla \mathrm{X}$, $\boldsymbol{H_{pre}}$
$\boldsymbol{q} \leftarrow \nabla \mathrm{X}$
for i=k-1…k-n do
    $\rho_i \leftarrow (\boldsymbol{y}_i^T \boldsymbol{s}_i)^{-1}$
    $\gamma_i \leftarrow \rho_i \boldsymbol{s}_i^T \boldsymbol{q}$
    $\boldsymbol{q} \leftarrow \boldsymbol{q} - \gamma_i \boldsymbol{y}_i$
end for
$\boldsymbol{q} \leftarrow \boldsymbol{H_{pre}} \boldsymbol{q}$
for i=k-n…k-1 do
    $\beta \leftarrow \rho_i \boldsymbol{y}_i^T \boldsymbol{q}$
    $\boldsymbol{q} \leftarrow \boldsymbol{q} + \boldsymbol{s}_i(\gamma_i - \beta)$
end for
Outputs: $\boldsymbol{q}$=$\boldsymbol{H}\nabla \mathrm{X}$
```

Stochastic formulation

In the traditional form, the source ensemble Ω is constant throughout the inversion and taken as the complete ensemble of sources positions. On the other hand, the stochastic approach uses a random subset of the sources that changes at every iteration, with a number of sources usually much smaller than the complete ensemble. In what follows, this is achieved by doing a random draw on shot gathers, at each iteration, with a constant probability distribution over shots. Because the acquired seismic data is highly redundant by design, a small source subsample can be used to estimate the value of the cost function and its gradient. This subsampling introduces some noise, but on average, the expectation of the cost function should converge to the true value along iterations, which justifies the stochastic gradient descent (SGD) algorithm (equation (2) with $\boldsymbol{H}_k \rightarrow \boldsymbol{I}$ and $\nabla \mathrm{X} \rightarrow \nabla \mathrm{X}_{\Omega_k}$).

The noise introduced by the stochastic subsampling is more problematic in the case of the L-BFGS algorithm. In particular, the gradient change vector $\boldsymbol{y}_k = \nabla X_{\Omega_{k+1}} - \nabla X_{\Omega_k}$ will be dominated by the sampling noise if $\Omega_{k+1} \neq \Omega_{k+1}$. Hence a naïve implementation of L-BFGS using the previously defined $\boldsymbol{y}_k$ will be unstable, and in most cases, will diverge. As shown by Schraudolph et al. (2007) for online learning, we can circumvent this problem by using the same source subset in the evaluation of the vector pairs $\boldsymbol{s}_k$ and $\boldsymbol{y}_k$. To efficiently implement this solution, we propose the stochastic SL-BFGS algorithm (Algorithm 3). A single iteration of this algorithm contains two parameters updates. After the first update (line 5), the gradient of the updated model is computed with the same subset of sources (line 6), which can be used to update the $\boldsymbol{s}_k$ and $\boldsymbol{y}_k$ vectors (line 9). The Wolfe line search can proceed simultaneously, at virtually no cost because the gradient of the updated model is already computed. Note

Algorithm 2: Wolfe line search

```
1.  Inputs: $\boldsymbol{m}, \boldsymbol{p}, \nabla X_0, X_0$
2.  $\alpha \leftarrow 1,\ \tau \leftarrow 0.6, c_1 \leftarrow 10^{-4}, c_2 \leftarrow 0.9$
3.  while stop_criteria do
4.      Compute $X, \nabla X, \boldsymbol{H_{pre}}$ with $\boldsymbol{m} \leftarrow \boldsymbol{m} + \alpha \boldsymbol{p}$
5.      if $X > X_0 + c_1 \boldsymbol{p}^{\boldsymbol{T}} \nabla X_0$
6.          $\alpha \leftarrow \tau\alpha$
7.      else if $\boldsymbol{p}^{\boldsymbol{T}} \nabla X < c_2 \boldsymbol{p}^{\boldsymbol{T}} \nabla X_0$
8.          $\alpha \leftarrow \alpha/\tau$
9.      else
10.         break
11. end while
12. Outputs: $\alpha, \nabla X, \boldsymbol{H_{pre}}$
```

that, with high probability, α=1 for L-BFGS and the line search does not require any new computations. Finally, the gradient of the updated model is used to update the model a second time (line 10), without updating $\boldsymbol{s}_k$ and $\boldsymbol{y}_k$, nor performing a line search. In our experience, this strategy allows two model updates using only two gradient computations, i.e. the step size $\alpha = 1$ respects Wolfe conditions most of the time. Note also that this algorithm can use a preconditioning matrix $\boldsymbol{H}_{\boldsymbol{pre}}$, for example the diagonal approximation of Shin et al. (2001).

Numerical experiment

To evaluate the performance of the SL-BFGS algorithm, we performed an acoustic FWI experiment with the classical Marmousi model (Versteeg, 1994). For seismic modelling and gradient calculations, we used the FDTD code of Fabien-Ouellet et al. (2017). The Marmousi model is discretized on a grid with cells of 20x20 m^2. We use the full aperture data, with shots and receivers every 20 meters, for a total of 460 shot points. The source is a Ricker wavelet with a central frequency of 7.5 Hz. The source signature and the density are considered fixed in this experiment, and no noise is added to the data. This is to keep to a minimum the number of factors that can impact FWI, as we want to focus on the convergence of different stochastic algorithms.

We compared the performance of the SL-BFGS with the stochastic gradient descent SGD algorithm. For SL-BFGS, we use a memory length n of 8. The stochastic gradient descent algorithm is identical to Algorithm 2, with the two-loops recursion (lines 5 and 10) replaced by a simple preprocessing of the gradient, $\boldsymbol{p}_k \leftarrow \boldsymbol{H}_{\Omega_k} \nabla X_{\Omega_k}$. Hence, each iteration step of the two algorithms should have more or less the same cost, with two gradient calculations per

Algorithm 3: Stochastic L-BFGS

```
1.  Inputs: $\boldsymbol{m}_0, \boldsymbol{d}$
2.  while stop_criteria do
3.      Draw $\boldsymbol{\Omega}_k$ from $\boldsymbol{d}$
4.      Compute $X_{\Omega_k}, \nabla X^0_{\Omega_k}, \boldsymbol{H}^0_{\boldsymbol{pre}\,\Omega_k}$
5.      $\boldsymbol{p}^0_k \leftarrow$ two-loops recursion$(\nabla X^0_{\Omega_k}, \boldsymbol{H}^0_{\boldsymbol{pre}\,\Omega_k})$
6.      $\{\alpha, \nabla X^1_{\Omega_k}, \boldsymbol{H}^1_{\boldsymbol{pre}\,\Omega_k}\} \leftarrow$Wolfe search$(\boldsymbol{p}^0_k, \nabla X^0_{\Omega_k}, X_{\Omega_k})$
7.      if $k>n$
8.          Discard $\{\boldsymbol{s}_{k-n}, \boldsymbol{y}_{k-n}\}$
9.      $\boldsymbol{s}_k \leftarrow \alpha \boldsymbol{p}^0_k$, $\boldsymbol{y}_k \leftarrow \nabla X^1_{\Omega_k} - \nabla X^0_{\Omega_k}$
10.     $\boldsymbol{p}^1_k \leftarrow$ two-loops recursion$(\nabla X^1_{\Omega_k}, \boldsymbol{H}^1_{\boldsymbol{pre}\,\Omega_k})$
11.     $\boldsymbol{m}_{k+1} \leftarrow \boldsymbol{m}_k + \alpha(\boldsymbol{p}^0_k + \boldsymbol{p}^1_k)$
12.     $k \leftarrow k+1$
13. end while
```

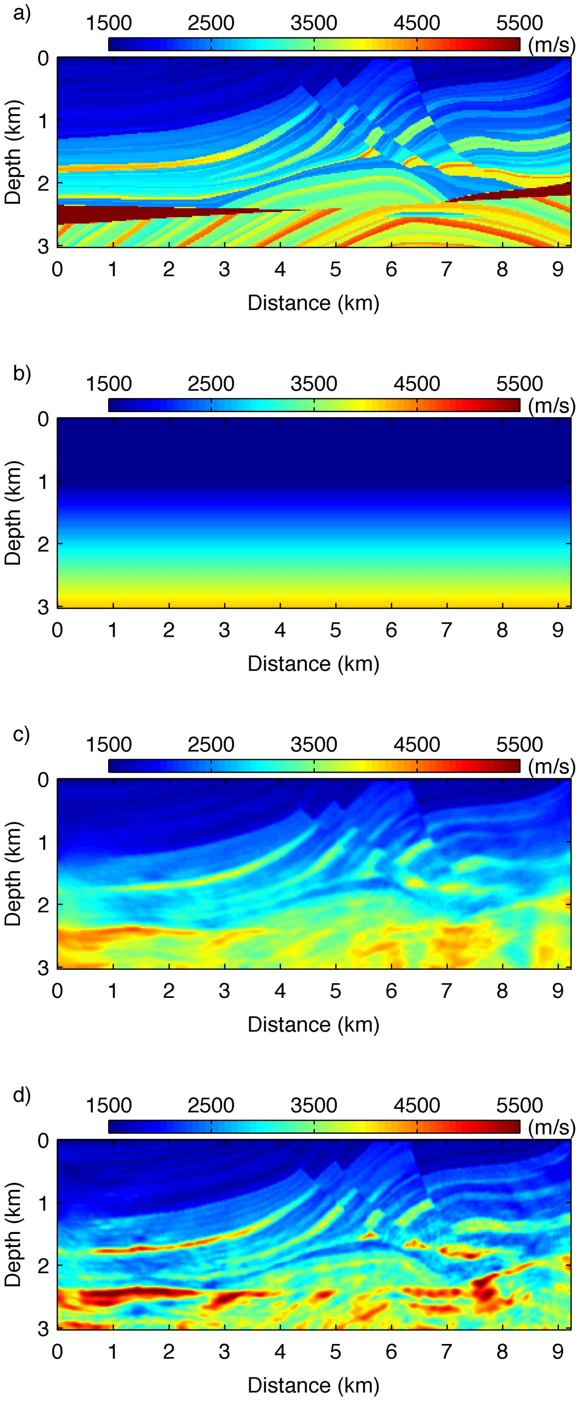

Figure 1: True Marmousi model (a), initial model (b), SGD inverted model (c) and stochastic L-BFGS inverted model (d).

iteration. We used the hierarchical inversion strategy of Bunks et al. (1995) and inverted sequentially for increasing discrete frequencies (2, 3, 5, 8, 12 and 16 Hz). In our time domain code, those frequencies are computed with the Fast Fourier Transform. For each frequency, 40 iterations are performed. We started with a linearly increasing *P*-wave velocity model starting from 1500 m/s to 4200 m/s (Figure1 b).

The inverted models obtained with SGD and SL-BFGS are presented in Figure 1 c) and d) respectively. Comparing the results with the true model (Figure 1 a), we see that the model above 2 km is very well reconstructed in both cases. Below 2 km, the inversion is more challenging due to poorer illumination, but the velocity magnitude is better reconstructed with the stochastic SL-BFGS algorithm. Overall, the resolution of the model obtained with SL-BFGS is higher than the SGD inversion. This is due to the better convergence of SL-BFGS that can take advantage of the curvature information.

To better compare the convergence of both algorithms, the cost function value is plotted against the number of iterations in Figure 2. At the lowest frequency of 2 Hz, both algorithms behave similarly and lead to more or less the same decrease in the cost function, with a faster decrease for SGD. However, from 5 Hz and higher, SGD performance degrades rapidly and the cost function stays above 10 %. The SL-BFGS convergence stays much more constant across frequency bands and reaches a plateau below 10% in all cases. This is the main reason why the model obtained by L-BFGS shows a much better solution: higher frequencies have converged, contrary to SGD.

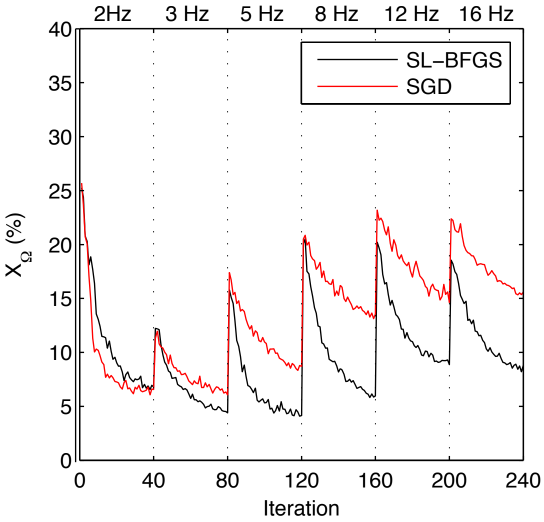

Figure 2: Cost function value as a function of iteration number for SGD and SL-BFGS. Increasing frequency bands are shown on the top axis.

Conclusions

We proposed a modification of the classical L-BFGS algorithm that supports the stochastic random subsampling of sources. The random subsampling allowed a drastic reduction of the computing time over the complete dataset with the Marmousi model: each iteration of the complete dataset would have required 460 shots, whereas we used a batch size of 20 shots with SL-BFGS. This represents a mere 5% of the cost of traditional L-BFGS for the same number of iterations. The second order information included in SL-BFGS allowed an improved convergence over SGD, at virtually no further computing costs. In summary, the stochastic L-BFGS algorithm allows a much faster convergence than SGD, at a fraction of the cost of the non-stochastic version.

Acknowledgements

This work was supported by the Vanier Canada Graduate Scholarships.

EDITED REFERENCES

Note: This reference list is a copyedited version of the reference list submitted by the author. Reference lists for the 2017 SEG Technical Program Expanded Abstracts have been copyedited so that references provided with the online metadata for each paper will achieve a high degree of linking to cited sources that appear on the Web.

REFERENCES

Bottou, L., 2010, Large-scale machine learning with stochastic gradient descent: Proceedings of COMPSTAT, 177–186, http://doi.org/10.1007/978-3-7908-2604-3_16.

Brossier, R., S. Operto and J. Virieux, 2009, Seismic imaging of complex onshore structures by 2D elastic frequency-domain full-waveform inversion: Geophysics, **74**, no. 6, WCC105-WCC118, http://doi.org/10.1190/1.3215771.

Bunks, C., F. M. Saleck, S. Zaleski and G. Chavent, 1995, Multiscale seismic waveform inversion: Geophysics, **60**, 1457–1473, http://doi.org/10.1190/1.1443880.

Castellanos, C., L. Metivier, S. Operto, R. Brossier and J. Virieux, 2015, Fast full waveform inversion with source encoding and second-order optimization methods: Geophysical Journal International, **200**, no. 2, 718–742, http://doi.org/10.1093/gji/ggu427.

Fabien-Ouellet, G., E. Gloaguen and B. Giroux, 2017, Time-domain seismic modeling in viscoelastic media for full waveform inversion on heterogeneous computing platforms with OpenCL: Computers & Geosciences, **100**, 142–155, http://doi.org/10.1016/j.cageo.2016.12.004.

Komatitsch, D., J. Ritsema and J. Tromp, 2002, The spectral-element method, Beowulf computing, and global seismology: Science, **298**, 1737–42, http://doi.org/10.1126/science.1076024.

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**, no. 6, WCC177–WCC188, http://doi.org/10.1190/1.3230502.

Nocedal, J. and S. Wright, 2006, Numerical optimization, Springer Science & Business Media.

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, http://doi.org/10.1111/j.1365-246X.2006.02978.x.

Pratt, R. G., C. Shin and G. Hicks, 1998, Gauss-Newton and full Newton methods in frequency-space seismic waveform inversion: Geophysical Journal International, **133**, 341–362, http://doi.org/10.1046/j.1365-246X.1998.00498.x.

Schraudolph, N. N., J. Yu and S. Günter, A Stochastic Quasi-Newton Method for Online Convex Optimization: Proceedings of The 11th International Conference on Artificial Intelligence and Statistics, 436-443.

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, http://doi.org/10.1046/j.1365-2478.2001.00279.x.

van Leeuwen, T., A. Y. Aravkin and F. J. Herrmann, 2011, Seismic Waveform Inversion by Stochastic Optimization: International Journal of Geophysics, **2011**, 1–18, http://doi.org/10.1155/2011/689041.

van Leeuwen, T. and F. J. Herrmann, 2013, Fast waveform inversion without source-encoding: Geophysical Prospecting, **61**, 10–19, http://doi.org/10.1111/j.1365-2478.2012.01096.x.

Versteeg, R., 1994, The Marmousi experience: Velocity model determination on a synthetic complex data set: The Leading Edge, **13**, 927–936, http://doi.org/10.1190/1.1437051.

Virieux, J., A. Asnaashari, R. Brossier, L. Métivier, A. Ribodetti and W. Zhou, 2014, An introduction to full waveform inversion: Encyclopedia of Exploration Geophysics, R1-1–R1-40, https://doi.org/10.1190/1.9781560803027.entry6.

Virieux, J., H. Calandra and R. E. Plessix, 2011, A review of the spectral, pseudo-spectral, finite-difference and finite-element modelling techniques for geophysical imaging: Geophysical Prospecting, **59**, no. 5, 794-813, http://doi.org/10.1111/j.1365-2478.2011.00967.x.