Time compressively sampled full-waveform inversion with stochastic optimization

Mathias Louboutin, and Felix J. Herrmann*
Seismic Laboratory for Imaging and Modeling (SLIM), University of British Columbia

SUMMARY:

Time-domain Full-Waveform Inversion (FWI) aims to image the subsurface of the earth accurately from field recorded data and can be solved via the reduced adjoint-state method. However, this method requires access to the forward and adjoint wavefields that are meet when computing gradient updates. The challenge here is that the adjoint wavefield is computed in reverse order during time stepping and therefore requires storage or other type of mitigation because storing the full time history of the forward wavefield is too expensive in realistic 3D settings. To overcome this challenge, we propose an approximate adjoint-state method where the wavefields are subsampled randomly, which drastically the amount of storage needed. By using techniques from stochastic optimization, we control the errors induced by the subsampling. Examples of the proposed technique on a synthetic but realistic 2D model show that the subsampling-related artifacts can be reduced significantly by changing the sampling for each source after each model update. Combination of this gradient approximation with a quasi-Newton method shows virtually artifact free inversion results requiring only 5% of storage compared to saving the history at Nyquist. In addition, we avoid having to recompute the wavefields as is required by checkpointing.

INTRODUCTION

Time-domain Full-Waveform Inversion (FWI) is a well known and widely used method for acoustic inversion as its marching in time structure allows easy and memory-efficient implementation and fast solutions. The requirement to access the forward wavefield in a reverse order in time led to numerous techniques including checkpointing and optimal checkpointing strategies (Symes, 2007,Griewank and Walther (2000)), recomputing the wavefield when needed from a partial save of the time history, or boundary methods that use integrals to compute the wavefield from its full history saved at the boundary of the domain (Plessi, 2006; Clapp, 2009).

All approaches that aim to address the above problem are balancing memory requirements and computational cost of recovering the wavefield for the unknown part of the history. The method we are proposing in addition will also balance memory requirements with the accuracy of the gradient calculations. We motivate this strategy from recent insights in stochastic optimization that prove that gradients need not be accurate especially in the beginning of an iterative inversion procedure (van Leeuwen and Herrmann, 2014). However, our approach differs slightly because we do not approximate the objective. Instead, we approximate the gradient by randomly subsampling the time history of the wavefield prior to correlation and applying the imaging conditions. To limit the imprint of this random subsamplings, we draw independent subsamplings for each source and after each model update. As a consequence, the subsampling-related artifacts average out as the inversion procedure progresses.

Our outline is as follows. First, we present our formulation of acoustic FWI followed by the proposed subsampling method. Next, we discuss it performance on a complex synthetic.

ACOUSTIC FULL-WAVEFORM INVERSION (FWI)

We start by defining usual adjoint-state time-domain full-waveform inversion (FWI) for an acoustic media (Virieux and Operto, 2009). The continuous acoustic wave equation for the pressure wavefield u is define by the partial differential equation (PDE)

$$m\frac{\partial^2 u}{\partial t^2} - \nabla^2 u = q \tag{1}$$

where m is the spatial distribution of the square slowness, ∇^2 is the Laplacian and q is the source. In the following, we will only consider acoustic media only with discrete measurements, collected in the vector $\mathbf{d}$, of the discrete pressure field $\mathbf{u}$. Our unknown parameter is the discretized square slowness $\mathbf{m}$ and we will denote by $\mathbf{u}[t] = u(t\delta t)$ the discrete wavefield at discrete time t with n_t the total number of time steps.

For a single source $\mathbf{q}_s$, time-domain inversion for the adjoint-state method derives from the following discrete PDE-constrained optimization problem:

$$\begin{aligned} &\underset{\mathbf{m},\mathbf{u}}{\text{minimize}}\ \frac{1}{2}\left\|\mathbf{P}_r\mathbf{u}-\mathbf{d}\right\|_2^2 \\ &\text{subject to } A(\mathbf{m})\mathbf{u} = \mathbf{q}_s, \end{aligned} \tag{2}$$

where $\mathbf{P}_r$ is the discrete projection operator restricting the synthetic wavefield to the receiver locations, $A(\mathbf{m})$ is the discretized wave equation matrix, $\mathbf{u}$ is the synthetic pressure wavefield, $\mathbf{q}$ is the discrete source and $\mathbf{d}$ is the measured data. The parameters to be estimated are the discrete square slownessses collected in the vector $\mathbf{m} = \{v_i^{-2}\}_{i=1\cdots N}$ with N the size of the discretization of the model and v the velocity. The adjoint-sate method solves the above problem by eliminating the PDE constraint and can be formulated as

$$\underset{\mathbf{m}}{\text{minimize}}\,\Phi_s(\mathbf{m}) = \frac{1}{2}\left\|\mathbf{P}_r A^{-1}(\mathbf{m})\mathbf{q}_s - \mathbf{d}\right\|_2^2. \tag{3}$$

The gradient of the above objective is given by the action of the adjoint of the Jacobian on the data residual $\delta\mathbf{d} = (\mathbf{P}_r\mathbf{u}-\mathbf{d})$ and reads

$$\nabla\Phi_s(\mathbf{m}) = -\sum_{t=1}^{n_t}\left\{\left((\mathbf{D}\mathbf{u})[t]\right)^T \mathrm{diag}(\mathbf{v}[t])\right\} = \mathbf{J}^T\delta\mathbf{d} \tag{4}$$

where $\mathbf{u}$ is the forward wavefield computed forwards in time via

$$A(\mathbf{m})\mathbf{u} = \mathbf{q}_s, \tag{5}$$

and $\mathbf{v}$ is the adjoint wavefield computed backwards in time via

$$A^*(\mathbf{m})\mathbf{v} = \mathbf{P}_r^*\delta\mathbf{d}. \tag{6}$$

In these expressions the superscript ^{T} denotes the transpose and the matrix $\mathbf{D}$ represents the discrete second-order time derivative. We define the discrete second-order time derivative as

$$\frac{\partial^2 \mathbf{u}}{\partial t^2}[t] = \frac{\mathbf{u}[t-2] - 2\mathbf{u}[t-1] + \mathbf{u}[t]}{\delta t^2} \tag{7}$$

so that the matrix $\mathbf{D}$ becomes a lower-triangular matrix and its transpose an upper-triangular matrix. Given this convention, we compute the second-time derivative of the forward wavefield with the left lower-triangular derivative matrix and conversely we apply the right upper-triangular derivative matrix to the time-reversed adjoint wavefield. With this convention we match the computational order so we have the necessary wavefields at the the different times available. We put the transpose on D so we only need to store one time-step of the forward wavefield.

Assuming $I = [1,2,3,....,n_t]$ and with the time derivative defined as in Equation 7, we can write an equivalent of the gradient in Equation 4 as

$$\nabla \Phi_s(\mathbf{m}) = -\sum_{t \in I} \Big[\text{diag}(\mathbf{u}[t])(\mathbf{D}^T \mathbf{v}[t]) \Big] = \mathbf{J}^T \delta \mathbf{d}. \tag{8}$$

The advantage of the alternative expression for the gradient, with the application of the derivative switched, is that we avoid storage of wavefield at additional time steps. While this alternative formulation is beneficial, gradient computations need storage of wavefields for all sources, which is computationally prohibitive.

APPROXIMATE FWI VIA STOCHASTIC GRADIENTS

The main disadvantage of time-domain adjoint state methods is the need of access to both the forward and adjoint wavefields. While certain techniques, mentioned in the introduction, overcome this need we follow a different strategy by allowing (random) errors in the gradient while still computing the objective accurately. Instead of using wavelet or other domain lossy compression techniques, we propose to randomly subsample the time histories of the forward and adjoint wavefields. In this way, we circumvent the costly recomputations part of optimal checkpointing, which requires $\mathcal{O}(n_t \log n_t)$ additional computations. Before giving details on our sampling method, let us first define the subsampling rate as

$$r = \frac{n_{\text{corr}}}{n_{\text{Nyquist}}} \tag{9}$$

where n_{corr} is the number of terms in the correlation calculations that are part of the gradient. Naively, these correlations are carried out over all time steps of the simulations but this not necessary since the simulation time steps are much smaller then the Nyquist sample interval. Therefore, we define our subsampling rate with respect to the n_{Nyquist}, which is the number of remaining terms after sampling at Nyquist.

Since we allow for errors, preferably random, in the gradient calculations we are interested in the cases where $r \ll 1$ that require much less storage. To that end, we redefine the gradient in Equation 8 by its subsampled counterpart given by

$$\tilde{\nabla} \Phi_s(\mathbf{m}) = -\sum_{t \in \tilde{I}} \Big[\text{diag}(\mathbf{u}[t])(\mathbf{D}^T \mathbf{v}[t]) \Big] = \tilde{\mathbf{J}}^T \delta \mathbf{d}. \tag{10}$$

In this approximation (denoted by the $\tilde{\text{I}}$), we sum over the subset of times $\tilde{I} \subset [1 \cdots n_t]$ with $\#(\tilde{I}) \ll n_{\text{Nyquist}}$. To reduce the buildup of subsampling-related artifacts, we draw for each gradient step a new independent random time history for each source. As we will show below, this generated fewer artifacts compared to deterministic periodic subsampling since the artifacts from different gradient updates tend to cancel. Instead of choosing the samples uniformly random, we employ a jitter sampling technique (Hennenfent and Herrmann, 2008) to control the maximum time gaps for a given subsampling rate r.

Random subsampling of the different terms in the correlation correlations is not the only way to subsample. As we have learned from the field of Compressive Sensing, we can also randomly mix and subsample the time histories. Since we do not have the complete time history available, this mixing can not be done with dense matrices but is feasible with sampling matrices $\mathbf{M}$ for which $\mathbf{M}^T\mathbf{M}$ is block diagonal. In this case, the gradient for a single shot and at a single position in the subsurface can be written as

$$\tilde{\nabla} \Phi_s(\mathbf{m})[\mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i] = -\mathbf{u}[\mathbf{x_i}, \mathbf{y_i}, \mathbf{z_i}]^T \mathbf{M}^T \mathbf{M} (\mathbf{D}^T \mathbf{v}[\mathbf{x_i}, \mathbf{y_i}, \mathbf{z_i}]), \tag{11}$$

where $\mathbf{M}$ is the time-mixing and subsampling matrix.

Algorithm 1 Time domain FWI via approximate gradients

Data: Measured data $\mathbf{d}$
Result: approximate solution of FWI $\mathbf{m}$ via approximate adjoint state
Choose a subsampling ratio for the wavefield
Set initial solution $\mathbf{m}_0$ to a smooth background
For k=1:niter
  For s=1:nsrc
  Draw an new set of time indexes $\tilde{I}$ for the wavefield
  Compute the forward wavefield $\mathbf{u}$ via Equation 1
  Compute the gradient via Equation 10
  stack with the previous gradients $\mathbf{g} = \mathbf{g} + \tilde{\nabla} \Phi_s(\mathbf{m^k})$
  End
  Get the step length α via Line Search
  Update the model $\mathbf{m}^{k+1} = \mathbf{m}^k - \alpha \mathbf{g}$
End
Solve FWI via approximates gradients

LINE SEARCH

In Algorithm 1, we use the weak Wolf line search in order to find the correct step length for our updates. However, remember that we are not working with the true gradient because of the subsampling. For that reason, we check whether the Wolfe line-search conditions (Skajaa, 2010) are met. We check whether the objective is decreasing and whether the curvature condition is met. If the curvature is not met but the objective is decreasing, we consider the step length to be correct. This guarantees we are still a decent direction.

EXAMPLES

To demonstrate the performance of our stochastic optimization approach to FWI, we first consider the performance of stochastic-gradient descent. We introdruce stochasticity by random subsampling the forward and adjoint wavefield which

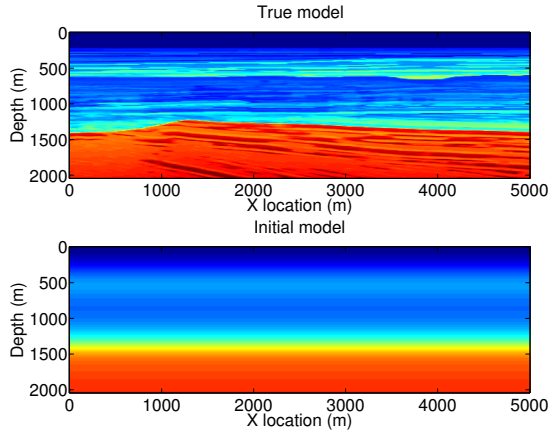

Figure 1: Initial and true model.

causes random errors in the gradients but not in the objective. For a synthetic, we invert for the square slowness from pressure data recorded at the surface. We have chosen a model small enough to insure we can store the complete time history at Nyquist sampling rate. We use 50 sources at 100m intervals and 201 receivers at 25m intervals and Ricker wavelet with a 15Hz peak frequency as a source. To time step given by the CFL stability condition is $\delta t = .5$ms yielding 4801 timesteps for the 2.4 s of recording. Given the central frequency of the Ricker wavelet the Nyquist sample interval is 4ms yielding $n_{\text{Nyquist}} = 601$. To limit the memory inprint we choose a subsampling ratio of $r = .05$ for our approximate gradient calculations according to Algorithm 1 . This reduces the number of time samples to be stored to 30—34.

To make a fair comparison, we fixed the number of gradient steps to 50 given a a relatively good 1D starting model (see Figure {1) obtained by smoothing. We compare our subsampling methods, namely jitter sample and mixed-subsampled, to gradients calculated for correlations carried out at Nyquist. The results are summarized in Figure 2 . As we can see from this Figure, the results obtained with approximate FWI sampled at $r = .05$ show generally the same features as the inversion result sampled at Nyquist. The artifacts appear random and relatively mundane. In Figure 3 we compare the relative model errors with respect to the true model and these plots show that periodic subsampling at $r = 0.05$ performs the worst, that Nyquist and jitter subsampled perform relatively the same throughout the iterations while the result with mixing and subsampling behaves better. This is not totally surprising because we have to remember that we mix and subsample the wavefield sampled at the time-stepping interval of the simulations, which is much smaller.

While the above example is certainly promising, it does not use second-order information. It is well-known that stochastic optimization techniques for second order method are challenging and a topic of open research. Having said that using vanilla out-of-the-box l-BFGS (Skajaa, 2010) performs remarkably well as we can observe from Figure 4, where we juxtapose results at Nyquist and at $r = 0.05$ for the jittered subsampling. It is clear

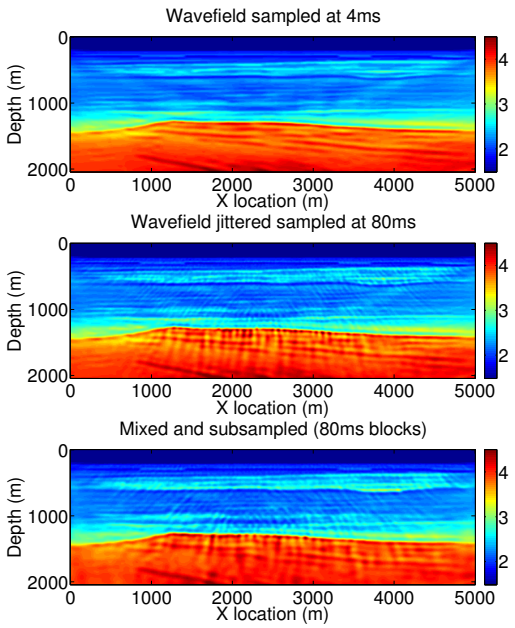

Figure 2: Inverted velocity for Gaussian subsampling.

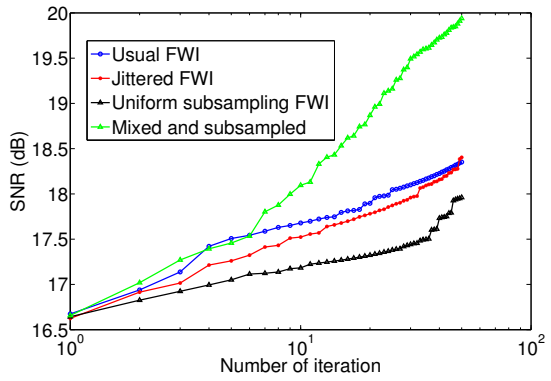

Figure 3: Recovered velocity SNR.

from this result that the subsampling-related artifacts nearly disappeared even for the jittered sampled case. We think this remarkable because we were able to obtain this excellent result with only 5% of the time history of the forward and adjoint wavefields sampled at Nyquist. Compared the sampling rate of the simulation this is a 160-fold reduction.

CONCLUSIONS AND DISCUSSION

We presented a new method to solve the time-domain adjoint-state full-waveform inversion problem without relying on complete storage or recomputation of the forward and reverse-time adjoint (receiver) wavefields. We accomplished this by combining insights from stochastic optimization and randomized subsampling where gradients are allowed to contain random sampling-related errors. Compared to random subsampling

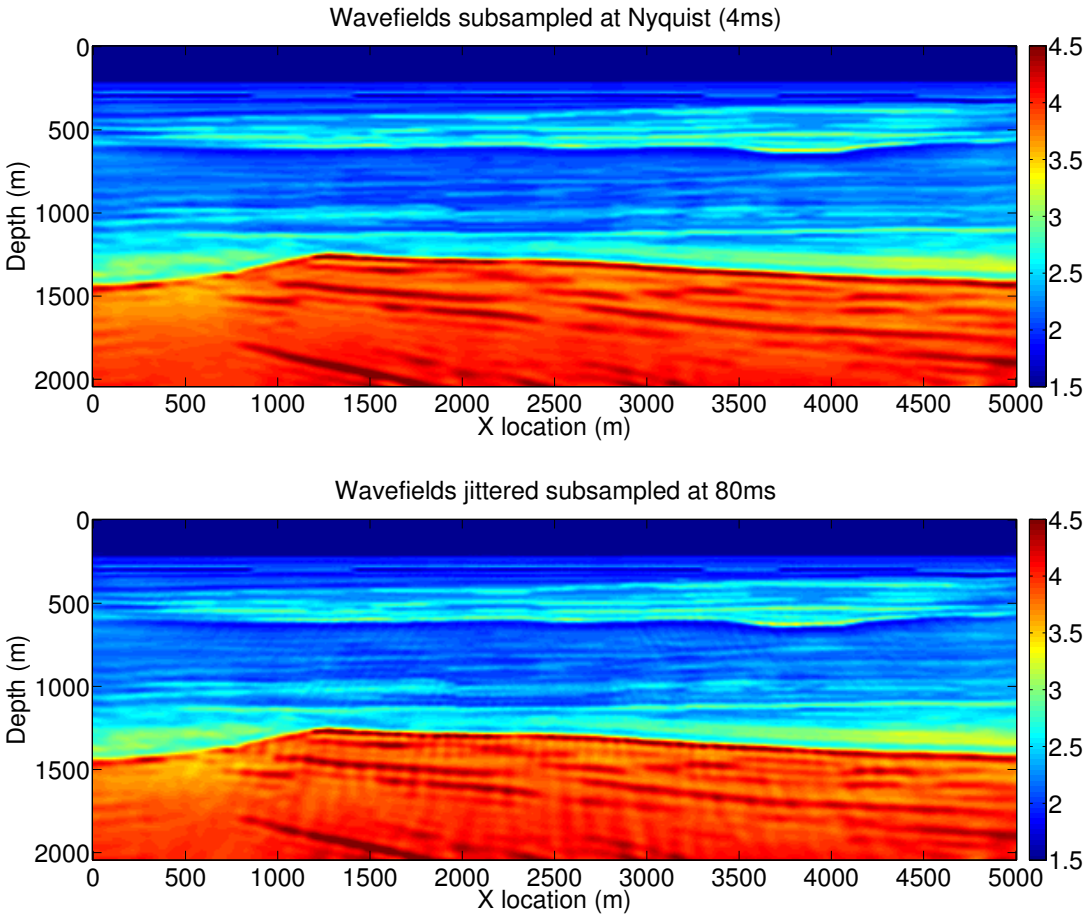

Figure 4: Inverted velocity with l-BFGS.

along the sources, an approach now widely used to reduces the number of wave-equation simulations, our method only approximates the gradient and not the objective by randomly subsampling the forward and adjoint wavefields prior to correlations and applying the imaging condition. We find that despite large subsampling ratios, 20-fold compared to Nyquist and 160-fold compared to the time-stepping of the simulations, excellent inversion results can be obtained by having independent samplings for each source, and possibly each gridpoint in the model, that are redrawn after each model update. The results for stochastic gradient decent clearly benefited from random sampling and we even found that the results for random mixing and subsampling were better than sampling at Nyquist. This means that the mixing picks up information at the fine time scales, which may explain the improvements in the nonlinear inversion. The results for quasi-Newton (with l-BFGS) are also excellent and show very little artifacts in case of jittered subsampling. We expect that these results will improve when we mix and subsample. As a results, we obtained an alternative method to reduce the memory and computational demands that arise from the need to have simultaneous access to the forward and reverse-time wavefield in time-domain full-waveform inversion based on time stepping.

ACKNOWLEDGEMENTS

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, BGP, CGG, Chevron, ConocoPhillips, Petrobras, PGS, WesternGeco, and Woodside.

REFERENCES

Clapp, 2009, Reverse time migration with random boundaries.

Griewank, A., and A. Walther, 2000, Algorithm 799: Revolve: An implementation of checkpoint for the reverse or adjoint mode of computational differentiation: ACM Transactions on Mathematical Software, **26**, 19–45. (Also appeared as Technical University of Dresden, Technical Report IOKOMO-04-1997.).

Hennenfent, G., and F. J. Herrmann, 2008, Simply denoise: wavefield reconstruction via jittered undersampling: Geophysics, **73**, V19–V28.

Plessi, 2006, A review of the adjoint-state method for computing the gradient of a functional with geophysical applications: Geophysical Journal International, **167**, 495–503.

Skajaa, A., 2010, Limited memory bfgs for nonsmooth optimization: masters, Courant Institute of Mathematical ScienceNew York University.

Symes, 2007, Reverse time migration with optimal checkpointing: GEOPHYSICS, **72**, SM213–SM221.

van Leeuwen, T., and F. J. Herrmann, 2014, 3D frequency-domain seismic inversion with controlled sloppiness: SIAM Journal on Scientific Computing, **36**, S192–S217. ((SISC)).

Virieux, J., and S. Operto, 2009, An overview of full-waveform inversion in exploration geophysics: GEOPHYSICS, **74**, WCC1–WCC26.