Extending the search space of time-domain adjoint-state FWI with randomized implicit time shifts

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

Abstract

Adjoint-state full-waveform inversion aims to obtain subsurface properties such as velocity, density or anisotropy parameters, from surface recorded data. As with any (non-stochastic) gradient based optimization procedure, the solution of this inversion procedure is to a large extend determined by the quality of the starting model. If this starting model is too far from the true model, these derivative-based optimizations will likely end up in local minima and erroneous inversion results. In certain cases, extension of the search space, e.g. by making the wavefields or focused matched sources additional unknowns, has removed some of these non-uniqueness issues but these rely on time-harmonic formulations. Here, we follow a different approach by combining an implicit extension of the velocity model, time compression techniques and recent results on stochastic sampling in non-smooth/non-convex optimization.

Introduction

To reduce the dependence of full-waveform inversion (FWI, (Virieux and Operto, 2009)) on the quality of starting models, we apply recent work by Burke et al. (2005) on non-convex/non-smooth optimization to adjoint-state FWI in the time-domain. By using this technique, we intend to tackle an important limitation of classical adjoint-state methods that ask for a starting that too accurate to be practical. Our approach is similar in spirit to other methods aimed to overcome cycle skip problems but instead of extending the search space—as for instance in Wavefield Reconstruction Inversion (WRI, van Leeuwen et al. (2014), van Leeuwen and Herrmann (2015)) where both the velocity model and wavefields are unknowns—our method extends the search space implicitly while relying on robust gradient sampling algorithm for nonsmooth, nonconvex optimization problems (Burke et al., 2005). By working on sets of nearby velocity models instead of working with a single velocity model as in conventional adjoint-state, our approach scales to high frequencies in 3D at low computational and memory overhead. To set the stage, we start by introducing the gradient sampling algorithm followed by our proposal how to incorporate this technique into FWI without incurring massive computational costs. We conclude by demonstrating the advantages of this method compared to regular FWI on a synthetic but realistic 2D model.

Gradient sampling for FWI

While smooth, i.e. the objective of FWI is differentiable with respect to the velocity model, FWI is non-convex and suffers as a consequence from local minima and sensitivity to starting models. A recent result by Curtis and Que (2013) extends Burke et al. (2005)'s original work on gradient sampling for non-smooth optimization problems to non-convex problems with encouraging results. Encouraged by these findings we propose to adapt this method to overcome issues related to the non-convexity of FWI. As a result, we expect the proposed method to make FWI less reliant on accurate starting models. As we show below, the cornerstone of this algorithm is to work with local neighborhoods (read sets of perturbed velocity model around the current model estimate) instead with a single velocity model. By working with these neighborhoods, we are able to reap global information on the objective from local gradient information at small cost.

Gradient sampling for non-smooth and non-convex optimization

We minimize at iteration k the objective function $\Phi(\mathbf{x})$ with respect to the unknown variable $\mathbf{x} \in R^N$ by using local gradient $\nabla\Phi(\mathbf{x})$ information by

- sampling $N+1$ vectors $\mathbf{x}_{ki}$ in a ball $B_{\varepsilon_k}(\mathbf{x}_k)$ defined as all $\mathbf{x}_{ki}$ such that $\|\mathbf{x}_k - \mathbf{x}_{ki}\|_2^2 < \varepsilon_k$, where ε_k is the maximum distance between the current estimate and a sampled vector;
- calculating gradients for each sample—i.e., $\mathbf{g}_{ki} = \nabla\Phi(\mathbf{x}_{ki})$;
- computing descent directions as a weighted sum over all sampled gradients—i.e.,

$$\mathbf{g}_k \simeq \sum_{i=0}^{p} \omega_i \mathbf{g}_{ki} \text{ , such that } \sum_{i=0}^{p} \omega_i = 1,\ \omega_i > 0\ \forall i \qquad (1)$$

- updating the model according to $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha \mathbf{H}^{-1}\mathbf{g}_k$, where α is a step length obtained from a line search and $\mathbf{H}^{-1}$ is an approximation of the inverse Hessian.

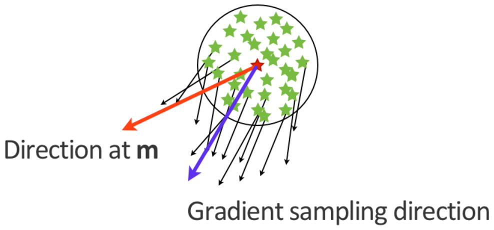

Figure 1: Gradient Sampling step.

We illustrate the calculation of a single iteration of this algorithm in Figure 1. While the above procedure demonstrably is capable of finding minima of non-convex minimization problems using local gradient information only, its computational costs are prohibitive. For instance, for a model of size N (in number of grid points) we would need to calculate $N+1$ gradients, a cost way out of reach for even the smallest FWI problem. The other major drawback is that the obtention of the weights $\omega_i$requires the solution of a quadratic subproblem over all the gradients $\mathbf{g}_{ki}$, which requires too much storage. We circumvolve these issues in two ways. First, we approximate the solution of the quadratic subproblems with predetermined random positive weights. This choice will not always give accurate gradient sampling weights but will always guaranty that the obtained update direction will perform as well as FWI in the worst case and as well as gradient sampling in the best case. Random weights that satisfy the constraints (positive and sum to one) of the subproblem improve then the chances to be in the best case scenario as the descent direction will be a non-optimal but feasible solution of the subproblem. Secondly, we exploit the structure of time-stepping solutions of the wave equation to build an implicit approximation of sampling of velocity models in the ball $B_{\varepsilon_k}(\mathbf{x}_k)$

Implicit time shift and on-the-fly Gradient sampling

As we just stated, gradient sampling relies on the computation of gradients at (too) many sampling points in the neighborhood of the current model. Unfortunately, this requirement is not feasible in a realistic setting because gradients of the FWI objective $\Phi_s(\mathbf{m})$ for an acoustic medium (Virieux and Operto, 2009) involves

$$\nabla\Phi_s(\mathbf{m}) = -\sum_{\mathbf{t}\in I}\left[\operatorname{diag}(\mathbf{u}[\mathbf{t}])(\mathbf{D}^\top\mathbf{v}[\mathbf{t}])\right], \tag{2}$$

where $\mathbf{m}$ is the square slowness and $\mathbf{q}_s$ the source. In this expression, the superscript $\top$ denotes the transpose and the matrix $\mathbf{D}$ contains the upwind discretization of the second-order time derivative $\frac{\delta^2}{dt^2}$. The sum in this expression for the correlation between the forward $\mathbf{u}$ and adjoint $\mathbf{v}$ wavefields, calculated from

$$A(\mathbf{m})\mathbf{u} = \mathbf{q}_s, \quad A^\top(\mathbf{m})\mathbf{v} = \mathbf{P}_r^\top\delta\mathbf{d}_s, \tag{3}$$

runs over the time index set I and involves inverting the discrete acoustic wave equation operator $A(\mathbf{m})$, the receiver projection operator $\mathbf{P}_r$ and the data residual $\delta\mathbf{d}_s$. As we can see, these gradient calculations require the solution of at least two PDEs so generating N gradient would require at least $2N$ PDE solves. However, as we will show below we can reduce these costs to only two PDE solves by randomly perturbing the velocity at each time step. Our main argument is based on the fact that the effects of slightly perturbed velocity models contained within the vicinity of the current velocity model can be approximated by time shifts. This means that we can obtain approximations of of the forward and reverse-time wavefields corresponding to perturbed velocity models, in the neighborhood of our current model estimate, by simply time shifting these wavefields. So, we argue that for a slightly perturbed velocity model $\tilde{\mathbf{m}}$ nearby $\mathbf{m}$, we can write

$$\mathbf{u}(\tilde{\mathbf{m}})[\mathbf{t}] \approx \mathbf{u}(\mathbf{m})[\mathbf{t}+\tau], \quad \mathbf{v}(\tilde{\mathbf{m}})[\mathbf{t}] \approx \mathbf{v}(\mathbf{m})[\mathbf{t}-\tau]. \tag{4}$$

Note that the shift in time τ is reversed for the adjoint wavefield as the propagation time axis is reversed (propagated backward in time). With this approximation, we can now write a simplified expression for the corresponding gradients with respect to perturbed models $\tilde{\mathbf{m}}$:

$$\nabla\Phi_s(\tilde{\mathbf{m}}) = -\sum_{\mathbf{t}\in I}\left[\operatorname{diag}(\mathbf{u}[\mathbf{t}+\tau])(\mathbf{D}^\top\mathbf{v}[\mathbf{t}-\tau])\right]. \tag{5}$$

We now have a way to compute multiple gradients for a range of velocity models in the neighborhood $B_{\varepsilon_k}(\mathbf{x}_k)$ at roughly the same computational cost incurred when calculating the usual FWI gradient. Moreover, by limiting the maximum time shift to $\tau_{max} = \frac{1}{f_0}$, where f_0 is the peak frequency of the source wavelet, we effectively define neighborhoods $B_{\varepsilon_k}(\mathbf{x}_k)$ at each iteration k. With this choice, we guaranty wavefields not to be shifted by more than half a wavelength meaning we only consider shifts contained within a single wavefront. While the above described method allows us to apply time shifts as proxies for evaluating gradients in perturbed velocity models, it requires storage of the forward and reverse-time (adjoint) wavefield — an unfeasible proposition. To overcome this memory issue, we propose to use recently developed technique that allows us to work with drastically (10×) randomly time subsampled wavefields (Louboutin and Herrmann, 2015). We mitigate artifacts induced by this time subsampling by generating different jitter sampling sequences for each shot such that maximum time gaps are controlled and coherent artifacts are prevented from building up. To avoid storage and explicit calculations of gradients for each perturbed model, we approximate the descent directions implicitly by weighted sums evaluated at jittered time samples:

$$\overline{\nabla\Phi_s(\mathbf{m})} := -\sum_{t\in\bar{I}}\left[\mathrm{diag}(\bar{\mathbf{u}}[\mathbf{t}])(\overline{\mathbf{D}^\top\mathbf{v}}[\mathbf{t}])\right], \quad \bar{\mathbf{u}} = \sum_{\mathbf{t}=t_i}^{t_{i+1}} \alpha_t\mathbf{u}[\mathbf{t}], \quad \overline{\mathbf{D}^\top\mathbf{v}} = \sum_{\mathbf{t}=t_i}^{t_{i+1}}\alpha_t\mathbf{D}^\top\mathbf{v}[\mathbf{t}]. \tag{6}$$

where $\bar{I} = [t_1, t_2, ..., t_n]$ are the jittered time sampled from $[1, 2, 3, ..., n_t]$. We arrived as these expressions, where $\alpha_t > 0$'s are random weights chosen such that $\sum\alpha_t = 1$, by expanding and factorizing the sum as a function of common time-shifts. We avoid the need of extra storage because we calculate these averaged wavefields on the fly and store them during forward propagation. The averaged adjoint wavefield is computed during back propagation. Details of our implicit gradient sampling scheme are included in Algorithm 1, where H^{-1} is the inverse of quasi-Newton Hessian.

Algorithm 1 |11. **End** Caption: Time-Compressed Gradient Sampling FWI algorithm (TC-GSFWI)

```
Input: Measured data $\mathbf{d}$ and starting model $\mathbf{m}_0$
Result: Estimate of the square slowness $\mathbf{m}_{k+1}$ via Gradient Sampling adjoint state
1. Set the subsampling ratio for the wavefield $\tau_{max}$
2. For k=1:niter
    3. For s=1:nsrc
        4. Draw an new set of time indices $\bar{I}$ for the wavefields for each shot
        5. Compute the average forward wavefield $\bar{\mathbf{u}}$ via Equation 3 & [#ubar] (on the fly sum)
        6. Compute the gradient via Equation 6 (on the fly sum of the adjoint)
        7. Stack with the previous gradients $\mathbf{g} = \mathbf{g} + \overline{\nabla\Phi_s(\mathbf{m}_k)}$
    8. End
    9. Get the step length $\alpha$ via linesearch
    10. Update the model $\mathbf{m}_{k+1} = \mathbf{m}_k - \alpha\mathbf{H}^{-1}\mathbf{g}$
```

Example

We demonstrate our method on a synthetic 2D model and include results from conventional FWI for comparison. To illustrate the advantage of our Time-Compressed Gradient Sampling FWI algorithm (TC-GSFWI), we deliberately choose a poor starting model, yielding a conventional FWI result that is cycle skipped (See Figure), derived from the synthetic BG's Compass model. Aside from being complex, i.e., the model contains fine-scale heterogeneity constrained by well data, this model has a challenging velocity kick back at 800 m depth that is not present in the starting model. In our experiment, we use 50 sources located at 100m intervals and at 25m depth, 201 receivers at 25m intervals at 100m depth (at the ocean bottom) and a Ricker wavelet with a 15Hz peak frequency as a source. The total recording time is 4s and the shot records are sampled at 4ms. As described above, the subsampling ratio for our TC-GSFWI algorithm is set according to the peak frequency of the wavelet, yielding jittered subsampled wavefields at an average temporal sampling interval of only 100ms. Because the randomized jittering samples on average at 25 % of Nyquist, we achieve a reduction in memory cost of 95% compared to memory requirements of conventional FWI. The inversion itself is carried out with a projected quasi-Newton method The inverted velocity models are shown in Figure 2. We observe that the initial velocity model is indeed cycle skipped for FWI, as the result shows a global low velocity instead of the layered shape structure of the true model. The inverted velocity with GSFWI, on the other hand, includes the main features of the true velocity model. This result is consistent with the fact that gradient sampling uses global information to minimize non-convex objectives—i.e., objectives with local minima due to cycle skipping. While our method for this example is clearly less sensitive to the quality of the starting model, the jitter sampling and implicit velocity perturbations lead to inversion artifacts. These artifacts can be removed by either composing certain (smoothness) constraints on the model or by reusing this inversion as a starting model for conventional FWI with these constraints. This inversion result is exciting because it reduces memory usage and was obtained at approximately the same computational cost as conventional FWI.

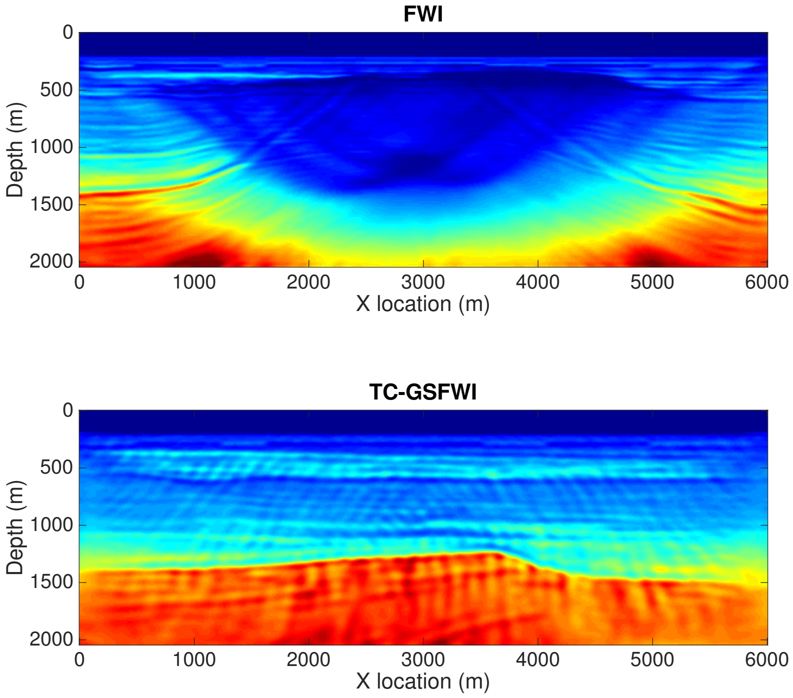

Figure 2: Inverted velocities

Conclusions

By combining randomized time subsampling in conjunction with the observation that random time shits applied to the forward and reverse-time wavefields can be considered as proxies for wavefields generated in randomly perturbed velocity models, we arrive at a computationally feasible formulation of gradient sampling for full-waveform inversion. Compared to ordinary gradient based algorithms, gradient sampling reaps global information making it suitable to minimize non-convex optimization problems. While we do not have a formal proof of our method, there are indication that our method is less sensitive to cycle skips. As a byproduct of the randomized sampling, we are also able to significantly reduce the storage requirements of gradient calculations making our method suitable for 3D models.

Acknowledgements

This research was carried out as part of the SINBAD II project with the support of the member organizations of the SINBAD Consortium.

References

Burke, J.V., Lewis, A.S. and Overton, M.L. [2005] A Robust Gradient Sampling Algorithm for Nonsmooth, Nonconvex Optimization. *SIAM Journal on Optimization*, **15**(3), 751–779.

Curtis, F.E. and Que, X. [2013] An Adaptive Gradient Sampling Algorithm for Nonsmooth Optimization. *Optimization Methods and Software*, **28**(6), 1302–1324.

van Leeuwen, T. and Herrmann, F.J. [2015] A penalty method for PDE-constrained optimization in inverse problems. *Inverse Problems*, **32**(1), 015007. (Inverse Problems).

van Leeuwen, T., Herrmann, F.J. and Peters, B. [2014] A new take on FWI: wavefield reconstruction inversion. *EAGE Annual Conference Proceedings*.

Louboutin, M. and Herrmann, F.J. [2015] Time compressively sampled full-waveform inversion with stochastic optimization. (submitted to the SEG conference).

Virieux, J. and Operto, S. [2009] An overview of full-waveform inversion in exploration geophysics. *GEOPHYSICS*, **74**(5), WCC1–WCC26.