Decentralized full-waveform inversion

Ajinkya Kadu[1] and Rajiv Kumar[2]
[1]Utrecht University, The Netherlands,
[2]University of British Columbia, Vancouver

Abstract

With the advent of efficient seismic data acquisition, we are having a surplus of seismic data, which is improving the imaging of the earth using full-waveform inversion. However, such inversion suffers from many issues, including (i) substantial network waiting time due to repeated communications of function and gradient values in the distributed environment, and (ii) requirement of the sophisticated optimizer to solve an optimization problem involving non-smooth regularizers. To circumvent these issues, we propose a decentralized full-waveform inversion, a scheme where connected agents in a network optimize their objectives locally while being in consensus. The proposed formulation can be solved using the ADMM method efficiently. We demonstrate using the standard marmousi model that such scheme can decouple the regularization from data fitting and reduce the network waiting time.

Introduction

In recent times, full-waveform inversion (FWI) has gained popularity in seismic imaging due to the improvements in details it brings to an image of earth. It is a classical procedure to find the parameters of earth, for example, velocity, density, etc, by matching the simulated wavefields to the observed field data collected at receivers (Virieux and Operto, 2009). We can cast the FWI as a discrete optimization problem described below:

$$\underset{\mathbf{m}}{\text{minimize}} \quad \frac{1}{2}\sum_{i=1}^{n_f} \|\mathbf{b}_i - \mathbf{F}_i(\mathbf{m},\mathbf{Q}_i)\|^2 + G(\mathbf{m}) \tag{1}$$

where $\mathbf{b}_i$ represents the vectorized monochromatic shot records, $\mathbf{F}_i(\mathbf{m},\mathbf{Q}_i) = \mathbf{P}\mathbf{H}_i^{-1}(\mathbf{m})\mathbf{Q}_i$ represents the simulated monochromatic data for all source experiments, $\mathbf{Q}_i$ is $n_s \times n_s$ matrix with n_s corresponds to the number of source experiments, and $G(\mathbf{m}) : \mathbb{R}^n \rightarrow \mathbb{R}$ is a regularization function, which incorporates the prior information about the model. Here, n_f represents the total number of frequencies, $\mathbf{H}_i$ is the discretization of the two-way Helmholtz operator $(\omega_i^2\mathbf{m} + \nabla^2)$ for temporal frequency ω_i and for a gridded squared slowness $\mathbf{m}$ while operator $\mathbf{P}$ restricts the data to the receiver positions. $\|\cdot\|$ represents ℓ_2 norm throughout the paper. For realistic seismic acquisition, n_s varies from 10^3 to 10^6 and ω is in the range of $4-40$ Hz with a sampling interval of 0.1 Hz. Given a large amount of data, the common strategy of solving the equation (1) is to use the multi-scale approach where we first divide the full frequency bandwidth into small segments. We then perform the inversion starting from low-frequency batches and use the inverted model as an initial guess for higher frequency batches (Pratt, 1999). For good performance, we need to ensure that our starting model at the lowest frequency batch is not cycle skipped. Even though inverting small batches circumvent the high cost of FWI, which involves extremely large multi-experiment data volumes, it requires repeated communication of gradients and loss function at every iteration to compute the model updates. Moreover, one needs to store all the data corresponding to small batches at the same place so that it can be easily available during inversion.

For large-scale seismic acquisition, both communication and storage become prohibitively expensive, thus increasing the turnaround time of FWI problem. To circumvent this, we propose a decentralized FWI, which is parallelizable over a connected network of agents, i.e., nodes or machines sitting at different remote locations. The proposed optimization method, known as the Alternating Direction Method of Multipliers (ADMM), uses the connected agents and solves equation (1) via collaboratively minimizing the sum of the local objective function over a common global variable, which is updated once in a while during the inversion. The abstract is organized as follows. First, we discuss the decentralized full waveform inversion, where we talk about splitting the inversion problem using local and global variables. Next, we utilize the ADMM framework to solve the decentralized waveform inversion and show that we can mitigate the communication cost while dealing with large-scale seismic data. Finally, we demonstrate the performance of our method on the Marmousi model.

Decentralized full-waveform inversion

Given the fact that FWI problem is *separable* over frequencies, we can rewrite the equation (1) with local variables $\mathbf{m}_i \in \mathbb{R}^n$ and a global variable $\mathbf{z} \in \mathbb{R}^n$ as:

$$\underset{\mathbf{m_i},\,\mathbf{z}}{\text{minimize}} \quad \frac{1}{2}\sum_{i=1}^{n_f} K_i(\mathbf{m}_i) + G(\mathbf{z}) \quad \text{subject to} \quad \mathbf{m}_i = \mathbf{z} \quad \text{for} \quad i = 1,\ldots,n_f. \tag{2}$$

Here $\mathbf{m}_i$ are the local variables (at the local agents), $\mathbf{z}$ is called as a *global* or *consensus* variable (typically, handled by master agent), and $K_i(\mathbf{m}_i) = \|\mathbf{b}_i - \mathbf{F}_i(\mathbf{m}_i,\mathbf{Q}_i)\|^2$ represents the loss function associated with i^{th} monochromatic frequency. The optimization problem in equation (2) tries to minimize the global decision variable $\mathbf{z}$ by locally minimizing the objectives in consensus. This formulation, known by the name of *Consensus* optimization, has been introduced in the 1980s (Bertsekas and Tsitsiklis, 1989) and an extensive survey (Nedic and Ozdaglar, 2009) has been available for an interested reader. The augmented Lagrangian (in *scaled* form) for such problem is written as:

$$\mathcal{L}_\rho(\mathbf{m}_i,\mathbf{z},\mathbf{u}_i) = \sum_{i=1}^{n_f} K_i(\mathbf{m}_i) + G(\mathbf{z}) + \frac{\rho}{2}\sum_{i=1}^{n_f} \|\mathbf{m}_i - \mathbf{z} + \mathbf{u}_i\|^2, \tag{3}$$

where $\mathbf{u}_i \in \mathbb{R}^n$ is a Lagrange multiplier associated with an equality constraint $\mathbf{m}_i - \mathbf{z} = 0$, and $\rho > 0$ is an augmented Lagrangian parameter. We call $\mathbf{m}_i$ as the primal variable and $\mathbf{u}_i$ as the dual variable. It is easy to see that the augmented Lagrangian function in equation (3) is *separable*. Hence, we can utilize the ADMM method (Boyd et al., 2011) to find the updates:

$$\mathbf{m}_i^{k+1} = \operatorname*{arg\,min}_{\mathbf{m_i}} \left\{ K_i(\mathbf{m}_i) + \frac{\rho}{2} \|\mathbf{m}_i - \left(\mathbf{z}^k - \mathbf{u}_i^k\right)\|^2 \right\}, \qquad \text{for } i = 1, \ldots, n_f \tag{4}$$

$$\mathbf{z}^{k+1} = \operatorname*{arg\,min}_{\mathbf{z}} \left\{ G(\mathbf{z}) + \frac{n_f \rho}{2} \|\mathbf{z} - \left(\bar{\mathbf{m}}^{k+1} + \bar{\mathbf{u}}^k\right)\|^2 \right\}, \tag{5}$$

$$\mathbf{u}_i^{k+1} = \mathbf{u}_i^k + \mathbf{m}_i^{k+1} - \mathbf{z}^{k+1}, \qquad \text{for } i = 1, \ldots, n_f \tag{6}$$

where $\bar{\mathbf{m}} = \frac{1}{n_f} \sum \mathbf{m}_i$ and $\bar{\mathbf{u}} = \frac{1}{n_f} \sum \mathbf{u}_i$ are the averages of primal variable and dual variable respectively. From equation (4), the minimization objective for $\mathbf{m}_i$ contains a data misfit term plus a Tikhonov regularization term. This implies that the objective is *well-conditioned* even if $K_i(\mathbf{m}_i)$ is not (as $\rho > 0$). This minimization (in $\mathbf{m}_i$) can be interepreted as solving a local problem on local nodes. Before we look at the global variable update, we introduce the *proximity* operator, also known as Moreau-Yosida regularization (Moreau, 1965). It is defined as:

$$\operatorname{prox}_{G,\upsilon}(\mathbf{y}) = \operatorname*{arg\,min}_{\mathbf{z}} \left\{ G(\mathbf{z}) + \frac{\upsilon}{2} \|\mathbf{z} - \mathbf{y}\|^2 \right\}.$$

The important property of the proximity operator is its evaluation. If $G(\mathbf{z})$ is simple enough, then the operator can be evaluated analytically (Combettes and Pesquet, 2011). For example, the proximity operator for Manhattan norm, $G(\mathbf{z}) = \lambda \|\mathbf{z}\|_1$ with $\lambda \geq 0$, is given by:

$$\operatorname{prox}_{G,\upsilon}(\mathbf{y}) := S_{\lambda/\upsilon}(\mathbf{y}), \quad \text{where} \quad S_\kappa(a) = (a - \kappa)_+ - (-a - \kappa)_+,$$

where S is known as the *soft thresholding* operator. Hence, the $\mathbf{z}$-update is simply applying an appropriate proximal operator on the average of primal and dual variables. In summary, the ADMM steps to solve decentralized FWI can be seen as the following: *(i)* Solve n_f independent sub-problems in parallel to compute $\mathbf{m}_i$ for $i = 1, 2, \ldots, N$, *(ii)* Collect computed $\mathbf{m}_i$'s in the central unit and update $\mathbf{z}$ by applying proximal operator on average, *(iii)* Broadcast computed $\mathbf{z}$ to n_f parallel units, *(iv)* Update dual variable $\mathbf{u}_i$ at each node using the received $\mathbf{z}$. The flow diagram for the decentralized FWI is shown in Figure 1.

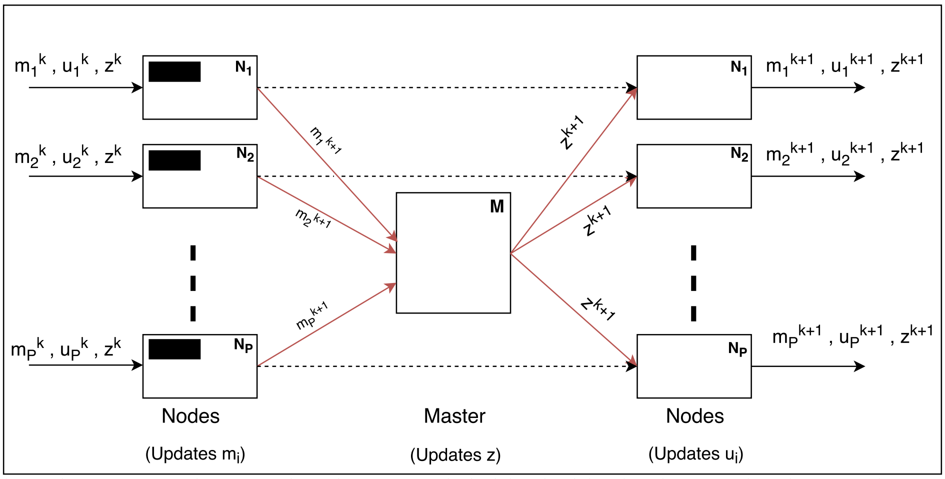


***Figure 1** Flow diagram of decentralized FWI. Solid (black) blocks denote the data at the node i. We use data only to update $\mathbf{m}_i$. Burgundy-colored lines denote the communications between nodes and master.*

Computational advantage

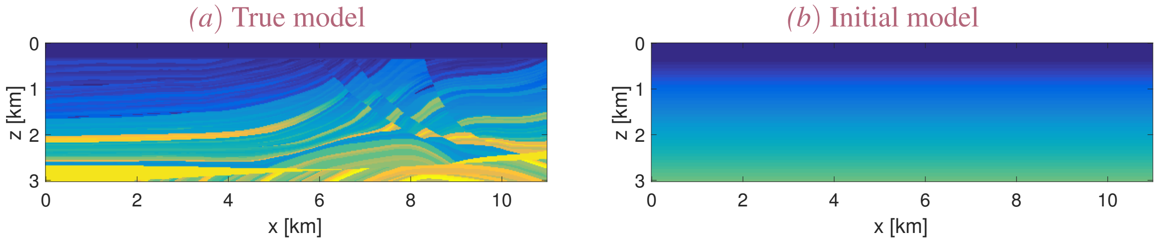
The equation (4) tells us that the primal variable updates $\mathbf{m}_i$ are computed through solving an unconstrained optimization problem. This allows the user to choose a 1st or 2nd-order method from the convex optimization literature without worrying about the constraints. The constraints on the model parameter are explicitly handled by the global variable. The $\mathbf{z}$-update is simply applying the proximity operator which can be done efficiently for many functions within an $\mathcal{O}(n^2)$, where n is the size of variable $\mathbf{z}$. Apart from the advantages of handling the constraints, communication cost in the proposed method only occurs at the step where we need to update $\mathbf{z}$. For practical applications, we need to exchange the local variable couple of times to updates $\mathbf{z}$, thus we can reduce the turnaround time drastically for waveform inversion. Also, the proposed method will be very beneficial for distributing the waveform inversion on cloud type environments where machine waiting time during communication of gradient updates can become prohibitively expensive.

Numerical Results

Table 1 Experiment Details

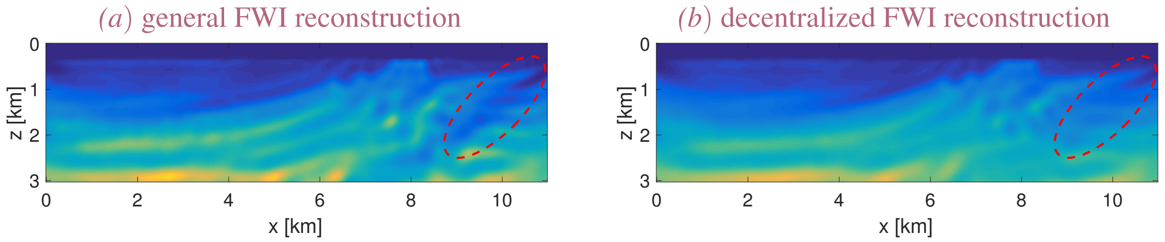
Source-Receiver grid	
number of sources	110
source grid spacing	100 m
source depth	8 m
number of receivers	220
receiver grid spacing	50 m
receiver depth	10 m
Source (Ricker wavelet)	
peak frequency	10 Hz
phase shift	0°
20 frequencies in $\{2,3\}$ Hz	

To demonstrate the feasibility of this approach, we consider a standard Marmousi model (Brougois et al., 1990) discretized on a 20 m × 20 m grid. The true model is shown in Figure 2. We generate the data using acoustic kernel with perfectly matched layer (PML) boundary conditions (Berenger, 1994) on all sides. More experiment details are provided in Table 1. To avoid inverse crime, we add a Gaussian noise to the data with SNR of 10 dB. The model has been constrained by lower (v_{lb} = 1500 m/s) and upper (v_{ub} = 5000 m/s) bound. At the start of the inversion, we consider kinematically-correct smooth model shown in Figure 2*(b)* for both versions of FWI. For general FWI, we consider all the data at once (but distributed on 20 nodes) and compute the single function and gradient vector at each step. Hence in each iteration, a total of 20 communications are performed by a master with nodes to get the information about the function and gradient. We execute a total of 20 L-BFGS iterations using a projected quasi-newton method (Schmidt et al., 2009) to generate the results described in Figure 3*(a)*. For decentralized FWI, we separate the misfit function with respect to frequency by allocating each frequency data to a particular node. Hence, we use 20 local nodes and one master. Each node optimizes its own model using L-BFGS (Nocedal, 1980). We allow a minimum of 5 iterations at each node before collecting the update and sending it to master node to handle the box constraints. Nodes corresponding to lower frequencies may perform more than 5 iterations as their iterations are faster compared to high-frequency. We perform 4 global iterations to match the same computations with general FWI. The reconstruction using this approach has been presented in Figure 3*(b)*.


(a) True model

(b) Initial model

Figure 2 *(a) True Marmousi model (11 km × 3 km) and (b) Initial model.*


(a) general FWI reconstruction

(b) decentralized FWI reconstruction

Figure 3 *reconstructions for (a) general FWI and (b) decentralized FWI.*

To compare the results of both formulations, we define three measures: *(i)* Normalized model misfit (NMM): It is a Euclidean distance between the reconstructed model and true model normalized by the distance between initial and true model. *(ii)* Normalized data misfit (NDM): It is the total data misfit from reconstructed model normalized by the initial data misfit. These measures will always lie in $[0,1]$, and the reconstruction will be considered *relatively* better if the measure is lower. The formulas for these measures are given in equation (7):

$$\text{NMM} := \|\mathbf{m}_{\text{rec}} - \mathbf{m}_{\text{true}}\| \, / \, \|\mathbf{m}_{\text{init}} - \mathbf{m}_{\text{true}}\| \, , \quad \text{NDM} := \|\mathbf{b} - \mathbf{F}(\mathbf{m}_{\text{rec}}, \mathbf{Q})\| \, / \, \|\mathbf{b} - \mathbf{F}(\mathbf{m}_{\text{init}}, \mathbf{Q})\| \, . \qquad (7)$$

And finally we define, *(iii)* Average (network) waiting time: It is an average wait time per node for a successful communication with the master. For example, in the case of general FWI, wavefield propagation at 3Hz takes sufficiently higher time than that of 2 Hz and hence lower frequency nodes need to wait for every iteration before communicating to master. The values of these measures are given in Table 2.

***Table 2** Performance Table*

	general FWI	**decentralized FWI**
Model misfit (NMM)	0.85	**0.78**
Data misfit (NDM)	**0.58**	0.60
Total communications	400	**80**
Average waiting time	50 sec	**4 sec**

From Figure 3, we see that the reconstruction via general FWI suffers from artifacts in the red-dotted region due to fitting the noise in the data. Because of this, the general FWI has a better data misfit compared to the decentralized FWI (refer Table 2). On other hands, a smooth model has been obtained from decentralized FWI due to Tikhonov regularization term in equation (4). The decentralized FWI has managed to reduce the reconstruction error (lower NMM value as seen from Table 2) along with cutting down the network waiting time by a huge margin.

Conclusions

We introduce a decentralized full-waveform inversion to handle large-scale data on the distributed platform. The proposed method separate the decision variable into a local and global component. We invert the local components on the network of agents, where the local update exchange is restricted to the agents. By doing this, we circumvent the communication cost required during the large-scale waveform inversion. To perform decentralized waveform inversion, we propose to use the ADMM method, which consists of three steps: A Tikhonov regularized model parameter estimation followed by a proximal step for handling the regularization on the model parameter and at the end updating the dual variables. The proposed method decouples the regularization from the data fitting and reduces the network waiting time drastically.

Acknowledgments:

A.K. would like to thank Fundamental Research on Matter (FOM) which is part of Netherlands Organisation for Scientific Research (NWO), and Shell Global Solutions International B.V for funding this research. R. K. would like to thank the member organizations of the SINBAD II project and the SINBAD Consortium for supporting this work. Finally, we would like to thank Tristan van Leeuwen for pointing us to the key literature.

References

Berenger, J.P. [1994] A perfectly matched layer for the absorption of electromagnetic waves. *Journal of computational physics*, **114**(2), 185–200.

Bertsekas, D.P. and Tsitsiklis, J.N. [1989] *Parallel and distributed computation: numerical methods*, 23. Prentice hall Englewood Cliffs, NJ.

Boyd, S., Parikh, N., Chu, E., Peleato, B. and Eckstein, J. [2011] Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, **3**(1), 1–122.

Brougois, A., Bourget, M., Lailly, P., Poulet, M., Ricarte, P. and Versteeg, R. [1990] Marmousi, model and data. In: *EAEG Workshop-Practical Aspects of Seismic Data Inversion*.

Combettes, P.L. and Pesquet, J.C. [2011] Proximal splitting methods in signal processing. In: *Fixed-point algorithms for inverse problems in science and engineering*, Springer, 185–212.

Moreau, J.J. [1965] Proximité et dualité dans un espace hilbertien. *Bull. Soc. Math. France*, **93**(2), 273–299.

Nedic, A. and Ozdaglar, A. [2009] Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, **54**(1), 48–61.

Nocedal, J. [1980] Updating quasi-Newton matrices with limited storage. *Mathematics of computation*, **35**(151), 773–782.

Pratt, R.G. [1999] Seismic waveform inversion in the frequency domain, Part 1: Theory and verification in a physical scale model. *Geophysics*, **64**(3), 888–901.

Schmidt, M., Berg, E., Friedlander, M. and Murphy, K. [2009] Optimizing costly functions with simple constraints: A limited-memory projected quasi-newton algorithm. In: *Artificial Intelligence and Statistics*. 456–463.

Virieux, J. and Operto, S. [2009] An overview of full-waveform inversion in exploration geophysics. *Geophysics*, **74**(6), WCC1–WCC26.