Computers & Geosciences 61 (2013) 184–197

Contents lists available at ScienceDirect

Computers & Geosciences

journal homepage: www.elsevier.com/locate/cageo

An optimized parallel LSQR algorithm for seismic tomography

En-Jui Lee [a,*], He Huang [b], John M. Dennis [c], Po Chen [a], Liqiang Wang [b]

[a] Department of Geology and Geophysics, University of Wyoming, Dept. 3006, 1000 E University Ave, Laramie, WY 82071, United States
[b] Department of Computer Science, University of Wyoming, Dept. 3315, 1000 E University Ave, Laramie, WY 82071, United States
[c] Computational Science Section, Scientific Computing Division, National Center for Atmospheric Research, 1850 Table Mesa Drive Boulder, CO 80303, United States

ARTICLE INFO

Article history:
Received 23 May 2013
Received in revised form
8 August 2013
Accepted 30 August 2013
Available online 12 September 2013

Keywords:
LSQR algorithm
Tomographic inversion
MPI
Computational seismology
Inverse problems
Parallel scientific computing

ABSTRACT

The LSQR algorithm developed by Paige and Saunders (1982) is considered one of the most efficient and stable methods for solving large, sparse, and ill-posed linear (or linearized) systems. In seismic tomography, the LSQR method has been widely used in solving linearized inversion problems. As the amount of seismic observations increase and tomographic techniques advance, the size of inversion problems can grow accordingly. Currently, a few parallel LSQR solvers are presented or available for solving large problems on supercomputers, but the scalabilities are generally weak because of the significant communication cost among processors. In this paper, we present the details of our optimizations on the LSQR code for, but not limited to, seismic tomographic inversions. The optimizations we have implemented to our LSQR code include: reordering the damping matrix to reduce its bandwidth for simplifying the communication pattern and reducing the amount of communication during calculations; adopting sparse matrix storage formats for efficiently storing and partitioning matrices; using the MPI I/O functions to parallelize the date reading and result writing processes; providing different data partition strategies for efficiently using computational resources. A large seismic tomographic inversion problem, the full-3D waveform tomography for Southern California, is used to explain the details of our optimizations and examine the performance on Yellowstone supercomputer at the NCAR-Wyoming Supercomputing Center (NWSC). The results showed that the required wall time of our code for the same inversion problem is much less than that of the LSQR solver from the PETSc library (Balay et al., 1997).

1. Introduction

Seismic waves generated by natural or manmade sources and recorded by seismometers carry important information about the physical properties of the subsurface earth structures through which they propagate. Seismic tomography is an imaging technique that assimilates ground-motion observations collected using seismometers to improve structural models of the Earth's interior and it has been one of most effective means for imaging the Earth's interior in the past few decades.

The seismic tomography problem is often formulated as an optimization problem, in which we search for an optimal earth structure model that minimizes an objective function defined in terms of certain misfit measurements that quantify the discrepancies between the observed wave-fields and the corresponding synthetic wave-fields predicted using a reference earth structure model. A typical objective function that is often employed in practice has the quadratic form

$$\chi^2(\mathbf{m}) = \mathbf{d}^T\mathbf{C}_d^{-1}\mathbf{d} + (\mathbf{m}-\tilde{\mathbf{m}})^T\mathbf{C}_m^{-1}(\mathbf{m}-\tilde{\mathbf{m}}), \tag{1}$$

where $\mathbf{d}$ is a vector composed of individual misfit measurements, $\mathbf{m}$ is a vector composed of model parameters, $\tilde{\mathbf{m}}$ is a vector of the reference structure model, $\mathbf{C}_m$ is the *a priori* model covariance matrix and $\mathbf{C}_d$ is the data covariance matrix. This type of objective functions arises in the context of statistical inference based on a Gaussian-Bayesian point of view (e.g., Tarantola, 2005). For an individual misfit measurement d_{in}^{sr}, which is the n-th misfit measurement on the i-th component seismogram generated by source s and recorded at receiver r, the data sensitivity kernel $K_{d_{in}^{sr}}^{\mathbf{m}}(\tilde{\mathbf{m}},\mathbf{x})$ is the functional (Fréchet) derivative of this misfit measurement with respect to the model parameters around the reference model (Backus and Gilbert, 1968), i.e.,

$$\delta d_{in}^{sr} = \int dV(\mathbf{x}) K_{d_{in}^{sr}}^{\mathbf{m}}(\tilde{\mathbf{m}},\mathbf{x})\delta\mathbf{m}(\mathbf{x}). \tag{2}$$

If discretized over space $\mathbf{x}$, the data sensitivity kernel $K_{d_{in}^{sr}}^{\mathbf{m}}(\tilde{\mathbf{m}},\mathbf{x})$ becomes a vector and the spatial integral in Eq. (2) can be expressed as an inner product. The Jacobian matrix $\mathbf{A}_k$ is the

* Corresponding author. Tel.: +1 307 460 1744.
E-mail addresses: elee8@uwyo.edu, rickli92@gmail.com (E.-J. Lee).

0098-3004/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cageo.2013.08.013

matrix with each row given by the discretized data sensitivity kernel for each individual misfit measurement. For nonlinear least-squares problems such as the one defined in Eq. (1), the Gauss–Newton algorithm is often an effective optimization algorithm because the Jacobian matrix, which only involves the first-order Fréchet derivative of every misfit measurement, can provide not only the gradient of the objective function but also an approximation of its Hessian. The exact Hessian of the objective function in Eq. (1) is given by

$$\mathbf{H} = \mathbf{A}_k^T \mathbf{C}_d^{-1} \mathbf{A}_k + \mathbf{C}_m^{-1} + (\nabla_{\mathbf{m}} \mathbf{A}_k)^T \mathbf{C}_d^{-1} \mathbf{d}, \tag{3}$$

which involves the derivative of the Jacobian matrix, therefore second-order derivatives of individual misfits. However, when $\mathbf{d}$ is small and/or the individual misfits are approximately linear with respect to model parameters (i.e., $\nabla_{\mathbf{m}} \mathbf{A}_k$ is small), the last term in Eq. (3) can be neglected. Under such an approximation, if we expand the objective function in Taylor series around the reference model $\tilde{\mathbf{m}}$, truncate the expansion to second-order and then set the derivative of the truncated series with respect to $\mathbf{m}$ to zero, we arrive at the Gauss–Newton normal equation

$$(\mathbf{A}_k^T \mathbf{C}_d^{-1} \mathbf{A}_k + \mathbf{C}_m^{-1})(\mathbf{m} - \tilde{\mathbf{m}}) = \mathbf{A}_k^T \mathbf{C}_d^{-1} \mathbf{d}. \tag{4}$$

In practice, we do not need to explicitly form this equation, because its solution can be computed by solving the linear system

$$\begin{bmatrix} \mathbf{C}_d^{-1/2} \mathbf{A}_k \\ \mathbf{C}_m^{-1/2} \end{bmatrix} (\mathbf{m} - \tilde{\mathbf{m}}) = \begin{bmatrix} \mathbf{C}_d^{-1/2} \mathbf{d} \\ 0 \end{bmatrix} \tag{5}$$

via a relaxation method. The LSQR algorithm of Paige and Saunders (1982) is one of the most popular solvers used in seismic tomography due to its efficiency and stability in solving large, sparse and ill-conditioned linear systems (e.g., Nolet, 1985; Nolet, 1993). Once Eq. (5) is solved, the structure model can be updated and the updated model can become the new reference model for the next iteration. This process can then be iterated until convergence.

In conventional ray-theoretic travel-time tomography, an individual misfit measurement is determined by the difference between the observed travel-time of a specific seismic phase and the corresponding model-predicted travel-time computed using a ray-tracing algorithm in the reference structure model and the data sensitivity kernel is determined by the ray-path connecting the source and the receiver for the selected seismic phase (e.g., Červený, 2005). The "finite-frequency" effect of wave-propagation can be accounted for by combining the Born approximation with the paraxial ray theory and the corresponding data sensitivity kernel exhibits the counterintuitive "banana-doughnut" phenomena, i.e., the sensitivity of the cross-correlation delay-time is non-zero within a tube surrounding the ray path (i.e., the Fresnel zone) but is zero on the ray path (Marquering et al., 1999; Dahlen et al., 2000; Hung et al., 2000; Zhao et al., 2000). Recent advances in parallel computing technology and numerical methods (e.g., Olsen, 1994; Graves, 1996; Bao et al., 1998; Komatitsch and Vilotte, 1998; Komatitsch et al., 2004; Dumbser et al., 2007) have significantly reduced the computational cost for solving acoustic and (visco) elastic seismic wave equations in realistic 3D earth structure models, which has opened up the possibilities for wave-equation-based (i.e., "full-wave") seismic tomography techniques. The adjoint-state method, which was adopted to solve seismic imaging problems in Bamberger et al. (1977, 1982) and later extended to 2D acoustic (Pratt and Worthington, 1990; Pratt et al., 1998) and 3D acoustic and elastic full-wave inversions (e.g., Tarantola, 1984; 1988; Tromp et al., 2004), is numerically efficient for computing the gradient of the objective function, as it only requires one forward and one adjoint wave-propagation simulation per seismic source. For a dataset with N_s seismic sources, the total number of wave-propagation simulations (forward and adjoint) needed for constructing the gradient is $2N_s$. Once the gradient of the objective function is available, gradient-based optimization algorithms such as the steepest-descent and the conjugate-gradient methods can be adopted to minimize the objective function. However the adjoint method is not efficient for constructing the Jacobian matrix, as it will need one forward and one adjoint simulation to compute the data sensitivity kernel for each misfit measurement. For realistic seismic tomography problems involving a large number of misfit measurements, the number of simulations and the computational cost needed to construct the Jacobian matrix using the adjoint method is prohibitive.

The scattering-integral (SI) method (Zhao et al., 2005), which is physically equivalent to, but computationally different from the adjoint method (Chen et al., 2007a), provides a computationally viable approach for constructing the Jacobian matrix. Consider the data sensitivity kernel of the misfit measurement d_{in}^{sr} with respect to the elastic moduli $c_{jklm}(\mathbf{x})$, after applying the reciprocity principle (Aki and Richards, 2002), the data sensitivity kernel can be expressed as (Chen et al., 2007a)

$$K_{d_{in}^{sr}}^{cjklm}(\mathbf{x}) = -\int dt \int d\tau\, J_{in}^{sr}(t) \partial_k G_{ji}(\mathbf{x}, t-\tau; \mathbf{x}_r) \partial_l u_m^s(\mathbf{x}, \tau), \tag{6}$$

where ∂_k represents the partial derivative with respect to x_k, $J_{in}^{sr}(t)$ is the functional derivative of the misfit measurement with respect to the waveform, i.e., $\delta d_{in}^{sr} = \int J_{in}^{sr}(t) \delta u_i^s(\mathbf{x}_r, t)\, dt$, $G_{ji}(\mathbf{x}, t; \mathbf{x}_r)$ is the Green's tensor for a unit impulsive force acting at the receiver location $\mathbf{x}_r$ and is named the "receiver Green's tensor" (RGT), $u_m^s(\mathbf{x}, t)$ is the m-th component forward wave-field generated by the seismic source s. The SI method is based on the observation that the RGTs do not depend on the sources. If we compute and stored them on disk, they can be re-used for constructing the data sensitivity kernels of different seismic sources. The computational cost for carrying out the temporal convolution and integration in Eq. (6) is almost negligible compared to the cost for carrying out a wave-propagation simulation. For a dataset with N_r receivers, the total number of simulations needed to construct the Jacobian matrix using the SI method is $N_s + 3N_r$ for 3D elastic problems. For acoustic and/or 2D problems, the number of simulation is even less. For realistic seismic tomography applications the disk space needed for storing the RGTs can be substantial but still manageable by adopting efficient data compression algorithms.

The SI method has been successfully applied to image the crustal structure of the Los Angeles Basin area in Chen et al. (2007b) and the tomographic inversion is currently being extended to Southern California. The linear system in Eq. (5) is more than 450 times larger in the Southern California inversion than that in the Los Angeles Basin inversion. In Chen et al. (2007b), the LSQR code used for solving Eq. (5) came from the PETSC library (Balay et al., 1997). But when we try to apply the same code to our inversion in Southern California, it does not provide satisfactory performance. In this paper we discuss our optimization of the parallel LSQR algorithm and demonstrate the performance of our code using one Gauss–Newton iteration from our Southern California tomographic inversion. In our Southern California inversion, we have completed 5 Gauss-Newton iterations so far. In each iteration, the updated model from the previous iteration is used as our reference model for the current iteration and the Jacobian matrix is re-computed for the current iteration using the SI method. Our optimized LSQR code is used to solve the Gauss-Newton normal equation in every iteration. The full inversion process based on both the adjoint and the SI method will be documented in a separate publication.