FAST TRACK PAPER

Full three-dimensional tomography: a comparison between the scattering-integral and adjoint-wavefield methods

Po Chen, Thomas H. Jordan and Li Zhao

1 Lamont-Doherty Earth Observatory, Columbia University, USA. E-mail: pochen@ldeo.columbia.edu
2 Department of Earth Sciences, University of Southern California, USA
3 Institute of Earth Sciences, Academia Sinica, Taipei 115, Taiwan

Accepted 2007 February 26. Received 2007 February 26; in original form 2006 July 17

SUMMARY

This paper analyses the computational issues of full 3-D tomography, in which the starting model as well as the model perturbation is 3-D and the sensitivity (Fréchet) kernels are calculated using the full physics of 3-D wave propagation. We compare two formulations of the structural inverse problem: the adjoint-wavefield (AW) method, which back-propagates the data from the receivers to image structure, and the scattering-integral (SI) method, which sets up the inverse problem by calculating and storing the Fréchet kernels for each data functional. The two inverse methods are closely related, but which one is more efficient depends on the overall problem geometry, particularly on the ratio of sources to receivers, as well as trade-offs in computational resources, such as the relative costs of compute cycles to data storage. We find that the SI method is computationally more efficient than the AW method in regional waveform tomography using large sets of natural sources, although it requires more storage.

Key words: adjoint wavefield, 3-D Earth structure, receiver Green tensor, scattering integral, tomography, waveform inversion.

1 INTRODUCTION

In seismic tomography, the data in a vector \( \mathbf{d} \) are represented as functionals of the earth model \( \mathbf{m} \), which specifies the wave velocities, mass density, and other seismic properties as functions of position \( x \) throughout the model volume \( V \). To set up the inverse problem, the data functionals are calculated for a starting model \( \mathbf{m}_s \), and the forward problem is linearized:

\[
\delta \mathbf{d} = A \delta \mathbf{m} = \int dV(x) \mathbf{K}_d(\mathbf{m}_s, x) \cdot \delta \mathbf{m}(x) \approx \delta \mathbf{d}(\mathbf{m}) - \delta \mathbf{d}(\mathbf{m}_s). \tag{1}
\]

The data sensitivity kernels in \( \mathbf{K}_d \) are the functional (Fréchet) derivatives of the data with respect to the model parameters around the starting model. The tomographic inverse problem is to estimate the best representation of the real Earth from observations of \( \mathbf{d} \). The approximations applied in tomographic methods thus involve the dimensionalities of the starting model \( \mathbf{m}_s \), the model perturbation \( \delta \mathbf{m} = \mathbf{m} - \mathbf{m}_s \), and the Fréchet kernels \( \mathbf{K}_d \).

Early tomographic studies built three-dimensional (3-D) model perturbations on top of 1-D starting models and, in constructing the Fréchet kernels, typically applied geometrical ray theory or 1-D approximations, such as path-averaging and local phase-velocity methods (e.g. Woodhouse & Dziewonski 1984; Li & Tanimoto 1993; Li & Romanowicz 1995; Ritsema & van Heijst 2000). Over the last decade, seismologists have improved the accuracy and spatial resolution of tomographic methods by calculating the Fréchet kernels for finite-bandwidth signals using asymptotic ray theory (Marquering et al. 1999; Dahlen et al. 2000) and normal-mode summation (Katzman et al. 1999; Zhao & Jordan 1998). The latter provides an essentially exact algorithm for computing 3-D Fréchet kernels for a general class of band-limited data functionals (Zhao et al. 2000), although it still relies on a 1-D starting model.

Advances in parallel computing technology and numerical methods (Olsen 1994; Graves 1996; Acikelik et al. 2003; Olsen et al. 2003; Komatitsch et al. 2004) have made large-scale 3-D numerical simulations of seismic wavefields much more affordable, and they open up the possibility of ‘full 3-D tomography’, in which \( \delta \mathbf{m} \) as well as \( \mathbf{K}_d \) is 3-D and \( \mathbf{K}_d \) is calculated using the full physics of 3-D wave propagation. Full 3-D tomography can account for the non-linearity of the structural inverse problem through iteration, thus providing the most efficient means for assimilating seismic observations into dynamic ground-motion models. This type of tomography was formulated as an adjoint problem in space–time domain in the 1980s (Tarantola 1984, 1988) and in space–frequency domain in the 1990s (Pratt 1990; Pratt et al. 1998). It has been applied in the seismic exploration industry for more than a decade (Biondi 2006); however, its adaptation to large-scale structural studies that use earthquake sources remains a work in progress (Tromp et al. 2005). We have...
recently demonstrated the ability to compute the Fréchet kernels for generalized seismological data functionals at regional distances using a 3-D starting model (Zhao et al. 2005), and we have applied this capability to improving existing 3-D models of crustal structure in the Los Angeles region by full 3-D tomography (Chen et al. 2007).

In this paper, we compare two formulations of the structural inverse problem: the adjoint-wavefield (AW) method, which back-propagates the data from the receivers to image structure (Ackelik et al. 2002; Tromp et al. 2005; Liu & Tromp 2006), and the scattering-integral (SI) method, which sets up the inverse problem by calculating and storing the Fréchet kernels for each data functional (Zhao et al. 2005; Chen et al. 2007). The two inverse methods are closely related (Tromp et al. 2005), but which one is more efficient depends on the overall problem geometry, particularly on the ratio of sources to receivers, as well as trade-offs in computational resources, such as the relative costs of compute cycles to data storage.

2 DATA FUNCTIONALS

We presume the seismogram observed on the \( i \)-th component of the \( r \)-th receiver from the \( s \)-th seismic source can be approximated by an instrument-filtered wavefield \( u'_s(t,x,i) \) calculated from an unknown (target) earth model \( m \). The quality of the approximation will depend on the ambient seismic noise as well as inadequacies in the way \( m \) represents the actual Earth (‘signal-generated noise’). For each seismogram, we consider a finite set of data functionals, indexed by \( n \), that measure the misfit between \( u'_s(t,x,i) \) and the wavefield \( \tilde{u}'_s(t,x,i) \) synthesized from the starting model \( \hat{m} \):

\[
d_t \delta d_{in} = D_t u'_s(t,x,i) - \tilde{u}'_s(t,x,i). \tag{2}
\]

The measurement process generally involves non-linear operations on both the observed and synthetic seismograms. We assume it is constructed to satisfy \( D_t [\tilde{u}'_s(t,x,i), \tilde{u}'_s(t,x,i)] = 0 \), so that the term \( \delta d(\hat{m}) \) in eq. (1) is zero. Because these differential data functionals are seismogram-specific, their Fréchet derivatives with respect to \( \tilde{u}'_s(t,x,i) \) can be expressed in terms of space-independent kernels,

\[
\delta d_{in} = \int dt \ J_{in}^t(t) \delta u'_s(t,x,i) . \tag{3}
\]

We consider a starting model \( \hat{m} \) that comprises estimates of the mass density \( \hat{\rho}(x) \) and elastic moduli \( \hat{c}_{ijkl}(x) \) and perturbations that are real and time-independent,

\[
\delta m(x) = \begin{bmatrix} \delta \hat{\rho}(x) \\ \delta \hat{c}_{ijkl}(x) \end{bmatrix} . \tag{4}
\]

This model perturbation can be related to seismogram perturbation using the first-order Born approximation (Dahlen & Tromp 1998; Zhao et al. 2000):

\[
\delta u'_s(t,x,i) = -\int dV(x) \int dt \sum_j \left[ G_{ij}(x,u_t-t) \partial_t^p u'_j(t,x) \delta \hat{\rho}(x) + \sum_{jklm} \partial_{ij} G_{ij}(x,u_t-t) \partial_t^p u'_j(t,x) \delta \hat{c}_{ijkl}(x) \right] . \tag{5}
\]

Here, \( \partial_t^p \) is the partial derivative with respect to \( \chi_t \) and \( \partial_t \) is the partial derivative with respect to \( t \). Substituting (5) into (3) yields the Fréchet derivative of \( d_{in}^t \) with respect to \( \hat{m} \). The resulting kernel is exact (Zhao et al. 2000), but the quality with which \( d_{in}^t \) can be approximated by a linear perturbation \( \delta d_{in}^t \) from \( \hat{m} \) depends on the type of data functional. We consider four examples.

1. Perhaps the most obvious data functional for waveform inversion is the difference between the target and starting seismograms at time \( t_n \) (Nolet 1987; Tarantola 1988; Ackelik et al. 2002):

\[
d_{in}^n = u'_s(t,x_i,t_n) - \tilde{u}'_s(t,x_i,t_n). \tag{6}
\]

In this case, the seismogram perturbation kernel \( J_{in}^t(t) \) is just the Dirac delta function \( \delta(t - t_n) \) , and the model perturbation kernel \( \nabla_m \delta u_{in}^n \) is given directly by the Born eq. (5). However, the Born approximation \( u'_s(t,x_i,t_n) \approx \tilde{u}'_s(t,x_i,t_n) + \delta u'_s(t,x_i,t_n) \) is limited in utility in tomography, because it effectively involves the linearization of the Fourier shift operator, \( e^{iTt} \) ; that is, for an observed waveform with centroid frequency \( \omega_0 \), it is valid only when \( \omega_0 \Delta T_{in} < 1 \), where \( \Delta T_{in} \) is the traveltime shift relative to the synthetic.

2. The traveltime shift \( d_{in}^n = \Delta T_{in}^n \) of an isolated waveform in a time window \( t_n, t'_n \) can be estimated by maximizing the cross-correlation between the observed and synthetic waveforms (e.g. Woodward & Masters 1991). For band-limited signals, the cross-correlation traveltime anomaly is usually a good approximation to the phase delay at the dominant frequency (Gee & Jordan 1992). The perturbation kernel is proportional to the waveform derivative (Luo & Shuster 1991; Zhao & Jordan 1998),

\[
J_{in}^t(t) = \frac{-1}{\Delta T_{in}^n} \int_t^{t'} dt \int_t^{t'} dt' [\delta t u'_s(t,x,i) \delta u'_s(t',x,i)] . \tag{7}
\]

where \( H(t) \) is the Heaviside function. Approximating the Born perturbation in terms of asymptotic ray theory yields the ‘banana-doughnut’ kernels for \( \Delta T_{in}^n \) described by Marquering et al. (1999) and Dahlen et al. (2000). Fréchet kernels for \( \Delta T_{in}^n \) that fully account for wavefield interactions in a 1-D starting model can be computed by expressing \( \delta T_{in}^n \) as a coupled-mode summation (Zhao et al. 2000).

3. The amplitude anomaly \( \Delta U_{in}^n \) of an isolated waveform in a time window \( t_n, t'_n \) can also be estimated by waveform cross-correlation (Ritsema et al. 2002; Tromp et al. 2005). The perturbation kernel for the relative amplitude anomaly, \( d_{in}^n = \Delta (\ln U_{in}^n) \approx \Delta U_{in}^n / U_{in}^n \) , is proportional to the waveform (Dahlen & Baig 2002),

\[
J_{in}^t(t) = \int_t^{t'} dt' \left[ u'_s(t,x_i,t) / U_{in}^n(t) \right] H(t - t_n) - H(t - t'_n) . \tag{8}
\]

4. The cross-correlation traveltime shift and relative amplitude anomaly can be generalized to frequency-dependent quantities that completely describe the discrepancies between the target and synthetic waveforms. In the frequency domain, the synthetic can be mapped into the target by an exponential (Ryttov) operator, \( e^{i \omega (\Delta \tau_{in}(\omega) + \Delta \tau_{out}(\omega))} \). The generalized seismological data functionals (GSDFs) are observational approximations to the phase-delay times \( \Delta \tau_{in}(\omega) \) and amplitude-reduction times \( \Delta \tau_{out}(\omega) \) , obtained by time-windowing and narrow-band filtering of the waveform cross-correlagram (Gee & Jordan 1992). The former is the frequency-dependent traveltime shift,

\[
d_{in}^n = \Delta T_{in}^n (\omega_n) , \tag{9}
\]

and the latter is proportional to the logarithmic amplitude anomaly,

\[
d_{in}^n = -1 / \omega_n \Delta (\ln U_{in}^n (\omega_n)) . \tag{10}
\]

The minus sign appears because a positive value of \( \Delta \tau_{in} \) corresponds to an amplitude decrease. The integration kernel \( J_{in}^t(t) \) for either type of GSDF measurement can be written as the convolution of the
synthetic waveform with a five-parameter Gaussian wavelet that has an effective centroid frequency $\omega_0$ and bandwidth $\sigma_0$ dependent on the windowing and filtering parameters (Chen 2005).

GSDFs are well suited for the tomographic inverse problem. In particular, their linearization depends on the Rytov approximation, which is valid for large phase-shifts as long as the phase perturbation per wavelength is small (Chernov 1960; Snieder & Lomax 1996). This is far less restrictive than the Born approximation, which requires small phase-shifts. A practical aspect of the GSDF method is that the phase-delay measurements need to be properly unwrapped, which can usually be achieved through bootstrapping from low to high frequencies.

### 3 MISFIT FUNCTIONAL AND ITS FRÉCHET DERIVATIVE

The tomographic methods discussed here are formulated as optimizations of the quadratic misfit functional,

$$\chi^2(m, \tilde{m}) = \frac{1}{2} \sum_{s1} \sum_{s2} \sum_{i} \sum_{s} |d_i^m(m, \tilde{m})|^2.$$  \hfill (11)

We explicitly sum over the $N_s$ sources and $N_r$ receivers that define the geometry of the inverse problem; $i$ and $n$ range across instrumental components and seismogram-specific measurements used in the data vector $d$. The data-space inner product can be weighted by the data covariance operator $C_d$, which we omit for notational simplicity.

The perturbation of the misfit functional with respect to $\tilde{m}$ can be written as

$$\delta \chi^2(m, \tilde{m}) = \sum_{s1} \sum_{s2} \sum_{i} d_i^m \delta d_i^{\tilde{m}} = \sum_{s1} \sum_{s2} \sum_{i} d_i^m \int dt J_i^{\mu}(t) \delta u_i(x, t)$$

$$= \int dV(x) K_{x, i}^{\mu}(x) \delta \rho(x)$$

$$+ \sum_{j, k, l, m} \int dV(x) K_{x, jklm}^{\tilde{c}, \delta \rho}(x) \delta c_{jklm}(x),$$ \hfill (12)

where the Fréchet kernels are

$$K_{x, i}^{\mu}(x) = -\sum_{s1} \int dt \int d\tau d_i^{\mu} J_i^{\mu}(t)$$

$$\times \sum_j G_{ji}(x, t - \tau; \tilde{x}) \delta \rho(x, \tau),$$ \hfill (13)

$$K_{x, jklm}^{\tilde{c}, \delta \rho}(x) = -\sum_{s1} \int dt \int d\tau d_i^{\mu} J_i^{\mu}(t) \delta c_{jklm}(x, \tau),$$ \hfill (14)

In these expressions we have applied the seismic reciprocity principle (Aki & Richards 1980),

$$G_{ji}(x, t - \tau; \tilde{x}) = G_{ji}(x, t - \tau; \tilde{x}).$$ \hfill (15)

The kernels of the misfit functional (13) and (14) are just the data-weighted sum of the Fréchet kernels in eq. (1); that is, $\delta \chi^2 = d^T A \delta m$.

### 4 ADJOINT-WAVEFIELD METHOD I: CONJUGATE-GRADIENT IMPLEMENTATION

One implementation of the AW method minimizes the misfit functional using the kernels $K_{x, i}^{\mu}(x)$ and $K_{x, jklm}^{\tilde{c}, \delta \rho}(x)$, which give the descent directions in a gradient-based optimization algorithm such as the conjugate-gradient method (Tromp et al. 2005). We call this type of implementation ‘AW–CG’. If we reverse time by making the substitution $t \rightarrow -t$, eqs (13) and (14) can be expressed in terms of the source wavefield and an adjoint wavefield:

$$K_{x, i}^{\tilde{\mu}}(x) = -\sum_s \int dt \sum_j \delta^2 u_i^\tau(x, t) \left[u_i^\tau\right]^\dagger(x, -\tau),$$ \hfill (16)

$$K_{x, jklm}^{\tilde{c}, \delta \rho}(x) = -\sum_s \int dt \delta \rho_i^\tau(x, t) \delta c_{jklm}(x, -\tau).$$ \hfill (17)

The adjoint wavefield is defined as

$$\left[u_i^\tau\right]^\dagger(x, t) = \int dt \sum_{i, \mu} G_{ji}(x, t - \tau; x_s) J_{i, \mu}^{\mu}(-t) d_i^\mu.$$ \hfill (18)

In practice, the adjoint wavefield is not constructed using the transposed Green’s tensor operator, but rather by solving the adjoint wave equation with the adjoint source field (Tarantola 1988; Ackelik et al. 2003; Tromp et al. 2005),

$$\left[f_i^\tau\right]^\dagger(x, t) = \sum_{i, \mu} J_{i, \mu}^{\mu}(-t) d_i^\mu \delta(x - x_s).$$ \hfill (19)

The solution is then integrated in time with the derivatives of the forward wavefield to construct $K_{x, i}^{\tilde{\mu}}$ and $K_{x, jklm}^{\tilde{c}, \delta \rho}$. Because the adjoint source field sums over the receiver index $r$, only one adjoint calculation is needed for each seismic source $s$, no matter how many receivers are used. The computational cost for one adjoint calculation is about the same as calculating the forward wavefield.

Because the forward wavefield $u$ and the adjoint field $[u^\tau]^\dagger$ are both source-specific, $2N_s$ wavefield simulations are needed for one calculation of $K_{x, i}^{\tilde{\mu}}$ and $K_{x, jklm}^{\tilde{c}, \delta \rho}$ ($N_s$ forward simulations and $N_r$ adjoint simulations) as well as $N_r$ operations to compute the time-integration of the forward and the adjoint wavefields, where $N_r$ is the number of spatial grid points. If the conjugate-gradient algorithm is used to minimize the misfit functional (Press et al. 1992), each iteration requires one calculation of $K_{x, i}^{\tilde{\mu}}$ and $K_{x, jklm}^{\tilde{c}, \delta \rho}$ to determine the conjugate direction in model space, followed by a line search to determine the optimal step to take in this conjugate direction. A cubic interpolation for the line search (Tromp et al. 2005) requires an additional $N_r$ forward calculations and $N_r$ adjoint calculations; hence, the total number of simulations per iteration is $4N_s$, and the total number of time integration operations is $2N_rN_s$ (Table 1). A quadratic interpolation scheme is often adequate, however (Tape et al. 2006), which reduces the total number of simulations per iteration to $3N_s$ and the total number of time integration per iteration to $N_rN_s$.

During the construction of $K_{x, i}^{\tilde{\mu}}$ and $K_{x, jklm}^{\tilde{c}, \delta \rho}$, we need simultaneous access to both the forward wavefield $u$ and the adjoint field $[u^\tau]^\dagger$ in order to perform the time integration operation. If we choose to store $u$ and read it back during the construction of $K_{x, i}^{\tilde{\mu}}$ and $K_{x, jklm}^{\tilde{c}, \delta \rho}$, the storage requirement for the adjoint method is the space–time volume of one wavefield, which is proportional to $N_rN_sN_T$, where $N_T$ is the total number of time steps for the wavefield simulation. If we choose to reconstruct $u$ on the fly, we need to store only the last time frame of $u$ as a function of space, which has volume proportional to $N_r$, although additional CPU time is then required to recover the forward wavefield by solving the wave equation backward in time, taking the final state of the forward wavefield as its initial condition (Gauthier et al. 1986; Tromp et al. 2005). The CPU time for doing one backward wave propagation is about the same as doing one forward simulation.
The computational cost for reading/writing data from/to storage facilities (I/O cost) depends on the number of time integration operations and whether the forward wavefield \( \mathbf{u} \) is stored or reconstructed on the fly. If a cubic interpolation scheme is used in the line search and \( \mathbf{u} \) is reconstructed on the fly, the I/O cost is proportional to \( 2N_tN_r \). The computational cost for a single conjugate-gradient step of this type of AW method is summarized in Table 1 under the column ‘AW—CG’.

### Table 1. Comparison of computational costs for a single optimization step of the scattering-integral (SI) and adjoint-wavefield (AW) methods.

<table>
<thead>
<tr>
<th>Cost</th>
<th>SI method</th>
<th>AW—CG(^a)</th>
<th>AW—GN(^b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Storage requirement</td>
<td>( 3N_tN_rN_T )</td>
<td>( N_T )</td>
<td>( N_T )</td>
</tr>
<tr>
<td>Number of simulations</td>
<td>( 3N_t + N_r )</td>
<td>( 6N_t )</td>
<td>( 4N_{CG}N_u + 2N_s )</td>
</tr>
<tr>
<td>Number of time integrations</td>
<td>( 2N_tN_rN_s )</td>
<td>( 2N_rN_s )</td>
<td>( N_r(2N_{CG}N_s + N_T) )</td>
</tr>
<tr>
<td>I/O cost</td>
<td>( N_TN_rN_s )</td>
<td>( 2N_rN_T )</td>
<td>( (2N_{CG}N_s + N_T)N_T )</td>
</tr>
<tr>
<td>Requires solving a linear system?</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Optimization algorithm</td>
<td>Gauss–Newton</td>
<td>Conjugate-Gradient</td>
<td>Gauss–Newton</td>
</tr>
<tr>
<td>Number of iterations needed to match one Gauss–Newton step</td>
<td>1</td>
<td>6–7</td>
<td>1</td>
</tr>
</tbody>
</table>

\(^a\)The AW—CG method listed here is based on cubic line search as implemented in Tromp et al. (2005).

\(^b\)The AW—GN method listed here is based on the implementation by Akcelik et al. (2002). The computational cost listed here is for a single Gauss–Newton optimization step, which involves \( N_{CG} \) conjugate-gradient iterations for solving the Gauss–Newton normal equation.

The transposed Green tensor \( G_{ji}(x, t - \tau; x_r) \) contains the wavefields generated by three orthogonal impulsive point forces acting at the receiver location \( x_r \), so we call it the receiver Green tensor (RGT). The data-specific Fréchet kernels can be computed by convolving the forward wavefield for source \( s \) with the RGT for receiver \( r \) and integrating the results against \( J^T_{ji} (t) \). Eqs (20) and (21) are equivalent to the scattering integrals (16)–(19) in Zhao et al. (2005).

We can construct the RGT once for each receiver, store it (at cost \( \sim N_tN_rN_T \)), and use it for every source in the dataset, which can be efficient when the number of seismic sources is larger than the number of receivers. A complete calculation of the data-specific kernels requires \( 3N_t \) simulations (considering 3 orthogonal unit forces at each receiver) and one forward simulation for each source—a total of \( 3N_t + N_r \), plus a storage cost proportional to \( 3N_tN_rN_T \). The number of time integration operations is determined by the convolution between the forward wavefields and the RGTs and is proportional to \( 2N_t \), where \( N_t \) is the number of time steps within the support of \( J^T_{ji} (t) \), the number of spatial grid points \( N_x \), and the total number of seismograms \( N_u \) used in the inversion. For three-component receivers, \( N_u \leq 3N_tN_r \). Owing to noise and instrumental reasons for eliminating seismograms, \( N_u \) is often much less than this bound. The computational cost for the SI method is summarized in Table 1 under the column ‘SI’.

In the SI method, the quadratic misfit is minimized by a Gauss–Newton algorithm. The misfit functional is expanded to second order in the model perturbation,

\[
\chi^2(m, \mathbf{m}) \approx \chi^2(\mathbf{m}, \mathbf{m}) + a(\mathbf{m})(m - \mathbf{m})^T + \frac{1}{2}(m - \mathbf{m})^TH(\mathbf{m})(m - \mathbf{m}),
\]

and its gradient with respect to the target model is set to zero,

\[
\nabla_m \chi^2(m, \mathbf{m}) = a(\mathbf{m}) + H(\mathbf{m})(m - \mathbf{m}) = 0.
\]

In these expressions, \( a = \nabla_m \chi^2 = -A(\mathbf{m})^T \mathbf{d} \) is the data-weighted Fréchet kernels, and \( H \) is the Hessian (Press et al. 1992),

\[
H = \nabla_m \nabla_m \chi^2 = A^T A + (\nabla_m A)^T \mathbf{d}.
\]

The second term in the Hessian is routinely neglected (e.g. Nolet, 1987; Sambridge, 1990), which leads to the Gauss–Newton normal equation,

\[
A^T \mathbf{d} = \mathbf{A}(m - \mathbf{m}).
\]

Because of the positive definiteness of the reduced Hessian \( A^T A \), the solution of eq. (25) will always give a descent direction. Eq. (25) will provide a good approximation to (23) if either the data vector \( \mathbf{d} \) or \( \nabla_m A \) is small. The latter will be the case when the data functionals are only weakly non-linear with respect to model parameters (Tarantola 2005). In this respect, the Rytov approximation used in constructing the GSDFs of (9) and (10) is usually much superior to the Born approximation required for the waveform difference functional (6).

If \( A^T A \) approximates the exact Hessian, the Gauss–Newton algorithm converges to a minimum of the misfit functional almost quadratically, a big advantage over gradient-based optimization algorithms such as steepest descent and conjugate gradient, which usually have linear or superlinear convergence rate (Press et al. 1992). In practice, the normal equation need not be formed, because its solution can be computed by solving the linear system \( A\tilde{\mathbf{m}} = d \) via a relaxation method, such as LSQR (Paige & Saunders 1982). This system can be scaled by a data covariance matrix \( C_d \) and augmented to include a model regularization term in the misfit functional (e.g. Snieder & Trampert 1999). A Bayesian solution can be obtained.
by referencing the model perturbation to some \( \mathbf{m}_{\text{prior}} \) and scaling the regularization using an \textit{a priori} model covariance operator \( \mathbf{C}_m \) (Tarantola 2005).

For each iteration of the Gauss–Newton’s method, we require \( 3N_r + N_s \) simulations, \( N_s \) convolution operations to construct kernels for individual data functionals, and additional CPU time (though not much) for solving the augmented linear system. The I/O cost for the SI method is proportional to \( N_r N_f N_T \) (Table 1).

6 ADJOINT-WAVEFIELD METHOD II: GAUSS–NEWTON IMPLEMENTATION

As pointed out in Tromp et al. (2005), the Fréchet kernel for the misfit functional can be expressed as a weighted summation of the Fréchet kernels for individual data functionals,

\[
K^m_{\gamma}(x) = \sum_{srin} d^m_{sr} K^m_{d_{sr}}(x)
\]

(26)

In the AW–CG method, the adjoint sources located at all receivers are propagated back simultaneously in one adjoint calculation, so the integration of the forward and the adjoint wavefields gives only a weighted summation over receivers,

\[
\int d^\tau \partial_x u^j_\gamma(x, \tau) [u_j^\gamma]^+(x, -\tau) = \sum_{srin} d^m_{sr} K^{m}_{d_{sr}}(x),
\]

(27)

\[
\int d^\tau \partial_x u^m_\gamma(x, \tau) \partial_x [u_j^\gamma]^+(x, -\tau) = \sum_{srin} d^m_{sr} K^{m_{\gamma,j}}(x).
\]

(28)

Therefore, the AW–CG method provides no access to the kernels for individual data functionals, and the Hessian cannot be constructed; only gradient-based algorithms can be used to minimize the misfit functional.

As pointed out in Akcelik et al. (2002, 2003), the “Hessian vector product” \( \mathbf{p} = \mathbf{Hq} \) can be obtained through one forward and one adjoint simulation and one time integration operation for each source without explicitly forming \( \mathbf{H} \) or \( \mathbf{A} \). For the forward simulation, we solve the wave equation for seismic source \( s \) with the following source field

\[
h^j_\gamma(x, \tau) = \sum_{klm} \delta_k \left[ \partial_t u^m_\gamma(x, \tau) \partial_t q_{j,klm}(x) \right],
\]

(29)

where \( q_{j,klm}(x) \) is \( q \) written in component form. We denote the solution at the receiver location \( x_r \) due to this source field as \( u^j_\gamma(x_r, t) \).

We then solve the adjoint wave equation for seismic source \( s \) with the following adjoint source field

\[
[h^j_\gamma]^+(x, \tau) = - \sum_{r} J^m_{sr} (-\tau) \delta(x - x_r) \int d^\tau J^m_{sr}(\tau) u^j_\gamma(x, \tau),
\]

(30)

and obtain the adjoint solution \( [w_j^\gamma]^+(x, t) \). The Hessian vector product can then be constructed by one integration operation for each seismic source \( s \)

\[
p_{j,klm}(x) = \sum_{s} \int dt \partial_t u^m_\gamma(x, t) \partial_t [w_j^\gamma]^+(x, -t),
\]

(31)

where \( p_{j,klm}(x) \) is \( p \) written in component form. If we solve the Gauss–Newton normal eq. (25) using the conjugate-gradient method, vector \( q \) is the conjugate direction and each conjugate-gradient step requires two Hessian vector products. To construct the right-hand side of eq. (25) requires one adjoint calculation and \( N_f \) time-integration operations for each source. If solving eq. (25) requires \( N_{CG} \) conjugate-gradient steps, then each Gauss–Newton iteration will require \( 4N_{CG}N_s + 2N_s \) simulations and \( N_f (2N_{CG}N_s + N_f) \) time-integration operations. The computational cost for this type of AW method is summarized in Table 1 under the column ‘AW–GN’.

7 APPLICATION OF THE SI METHOD TO THE LOS ANGELES REGION

We have applied the SI method to a full 3-D tomography using phase-delay measurements of crustal \( P \) and 3 phases from 67 small earthquakes recorded by 48 stations of the Southern California Seismic Network. We inverted 7364 frequency-dependent phase delays in the frequency band from 0.2 to 1.0 Hz, which we measured relative to synthetics calculated from version 3.0 of the SCEC 3-D Community Velocity Model (CVM3.0) (Kohler et al. 2003). The wavefields and the RGTs were computed on a mesh with 200-m grid spacing (36 million nodes) using a fourth-order, staggered-grid, finite-difference code developed by K. Olsen.

The computational aspects of the inversion cycle are summarized in Table 2. All computations and I/O operations were performed in parallel using the Message Passing Interface (Pacheco 1997). Each wave-propagation simulation took about 90 min on a Linux cluster of 128 Pentium-4 CPUs. Constructing each data kernel required about 4 min to read the RGTs from local disk and 1 min to do the convolutions. To represent \( 8 \mathbf{m} \) and compute the kernels, we decimated the simulation grid to a 2-km node spacing. The linearized perturbation equations were scaled by the data-covariance matrix \( \mathbf{C}_d \) and augmented to incorporate the data nuisance parameters (source and station anomalies) and the inverse model-covariance operator \( \mathbf{C}_m^{-1} \), needed for a Bayesian solution. The resulting linear system, which measured 77,695 columns by 84,896 rows, was inverted by a parallelized version of LSQR in less than 1 hr—a negligible computing expense. The variance reduction, calculated by resynthesizing the seismograms using the perturbed model and remeasuring the GSDFs, was almost 80% for phase-delay times. For the amplitude-reduction times, which were measured but not inverted, the perturbed model achieved a variance reduction of about 40%, consistent with the improvements visible in the waveform fits. A complete description of the inversion results is given in Chen et al. (2006).

On the 128-processor cluster, about 20 day of CPU time and 24 terabytes (TB) of storage space were required for one complete iteration of the SI method. Based on the scalings in Table 1, we estimate that the AW–CG method would require about 100 day of CPU time for six conjugate-gradient steps, which is probably the minimum number needed to achieve the data variance reduction obtained in a single Gauss–Newton step (Table 1; Tape et al. 2006). Numerical experiments conducted in Akcelik (2002) show that for the AW–GN method \( N_{CG} \) usually ranges from 10 to 30. Therefore, more CPU time would be needed for the AW–GN method to match one SI step. The AW storage cost would be very small (~0.04 TB).

Table 2. Computational parameters for the Los Angeles Basin tomography using the SI method.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of stations ( N_r )</td>
<td>48</td>
</tr>
<tr>
<td>Number of earthquakes ( N_s )</td>
<td>67</td>
</tr>
<tr>
<td>Number of seismograms ( N_m )</td>
<td>2000</td>
</tr>
<tr>
<td>Number of FD simulations ( 3N_r + N_s )</td>
<td>211</td>
</tr>
<tr>
<td>Simulation grid spacing, time interval ( 200 ; \text{m}, 0.01 ; \text{s} )</td>
<td></td>
</tr>
<tr>
<td>Simulation grid points ( N_f ), time steps ( N_T )</td>
<td>36 140 440, 6000</td>
</tr>
<tr>
<td>Number of CPUs</td>
<td>128</td>
</tr>
<tr>
<td>Total CPU time per iteration</td>
<td>62 000 CPU-hours</td>
</tr>
<tr>
<td>Total disk space ( 3N_r N_f N_T )</td>
<td>24 TB</td>
</tr>
</tbody>
</table>
When disk storage is cheap and the number of earthquakes is larger than the number of receivers, as in our case, the SI method is more economical than the AW methods. As the number of sources used in the inversion increases, the efficiency of the SI method increases. Moreover, the storage requirements and I/O overhead of the SI method can be reduced, perhaps significantly, by deploying more efficient data compression algorithms (e.g. Bassioni 1985).

Storing the RGTs has other advantages. By applying the reciprocity principle, we can efficiently use them to calculate synthetic seismograms and obtain the source parameters of new earthquakes; we need only extract a small, source-centred volume from each RGT to compute three-component seismograms and the partial derivatives required for source estimation. We have implemented automated procedures that invert the waveform data from small to moderate-sized earthquakes in Southern California for centroid moment tensors (Zhao et al. 2006) and finite moment tensors (Chen et al. 2005).

For example, the source parameters used in the full 3-D tomography described here were refined using the RGTs computed for CVM3.0 prior to the structural inversion.

ACKNOWLEDGMENTS

This research was supported by the Southern California Earthquake Center; SCEC is funded by NSF Cooperative Agreement EAR-0106924 and USGS Cooperative Agreement 02HQAG0008. We thank J. Tromp and J. Bielak for helpful discussions of AW methods. High-throughput computations were supported by the SCEC Community Modeling Environment under NSF grant EAR-0122464 and the USC Center for High Performance Computing and Communications; we thank P. Maechling and other members of CME collaboration and G. Staples from USC/HPC for help in implementing the simulation and inversion algorithms. The SCEC contribution number for this paper is 1011.

REFERENCES


Akcelik, V. et al., 2003. High resolution forward and inverse earthquake modeling on terascale computers, SC2003, Phoenix, AZ.


