PARALLEL MULTISCALE GAUSS-NEWTON-KRYLOV METHODS FOR INVERSE WAVE PROPAGATION*

VOLKAN AKCELIK†, GEORGE BIROS‡, AND OMAR GHATTAS§

Abstract. One of the outstanding challenges of computational science and engineering is large-scale nonlinear parameter estimation of systems governed by partial differential equations. These are known as *inverse problems*, in contradistinction to the *forward problems* that usually characterize large-scale simulation. Inverse problems are significantly more difficult to solve than forward problems, due to ill-posedness, large dense ill-conditioned operators, multiple minima, space-time coupling, and the need to solve the forward problem repeatedly. We present a parallel algorithm for inverse problems governed by time-dependent PDEs, and scalability results for an inverse wave propagation problem of determining the material field of an acoustic medium. The difficulties mentioned above are addressed through a combination of total variation regularization, preconditioned matrix-free Gauss-Newton-Krylov iteration, algorithmic checkpointing, and multiscale continuation. We are able to solve a synthetic inverse wave propagation problem though a pelvic bone geometry involving 2.1 million inversion parameters in 3 hours on 256 processors of the Terascale Computing System at the Pittsburgh Supercomputing Center.

1. Computational challenges in large-scale inverse problems. One of the outstanding challenges of computational science and engineering is large-scale parameter estimation of systems governed by partial differential equations (PDEs). These are known as *inverse problems*, and are distinguished from the *forward problems* that usually characterize large-scale simulations. In the forward problem, the PDE parameters—initial conditions, boundary and domain sources, material coefficients, and the domain geometry—are known, and the state of the system is determined by solving the PDEs. The inverse problem reverses the process: some components of the state are observed, while some subset of PDE parameters is unknown. An objective function that minimizes the differences between the observed states and those predicted by the PDEs is formulated, and solution of a nonlinear optimization problem yields components of the unknown parameters.

The inverse problem is often significantly more difficult to solve than the forward problem, because:

- forward solution is just a subproblem of the inverse problem;
- the inverse problem is often ill-posed despite the well-posedness of the forward problem;
- the inverse operator couples the entire time history of the system's response, as opposed to the usual case with forward evolution operators; and
- the inverse problem can have numerous local solutions.

Not surprisingly, most of the work in large-scale simulation has been directed at the forward problem. But sustained advances over the past 20 years have produced a body of efficient,

*This work supported by the National Science Foundation's Knowledge and Distributed Intelligence (KDI) and Information Technology Research (ITR) programs (through grants CMS-9980063 and ACI-0121667), the Department of Energy's Scientific Discovery through Advanced Computation (SciDAC) program through the TOPS Center, and the Computer Science Research Institute at Sandia National Laboratories. Computing resources on the Cray T3E and Compaq TCS systems are provided by the Pittsburgh Supercomputing Center under awards ASC-010025P and ASC-010036P.

†Mechanics, Algorithms, and Computing Laboratory, Department of Civil & Environmental Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania, 15213, USA (volkan@cs.cmu.edu).

‡Courant Institute, for Mathematical Sciences, New York University, New York, NY, 10012, USA (biros@cs.nyu.edu)

§Mechanics, Algorithms, and Computing Laboratory, Departments of Biomedical Engineering and Civil & Environmental Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania, 15213, USA (oghattas@cs.cmu.edu).

0-7695-1524-X/02 $17.00 (c) 2002 IEEE

parallel, and scalable algorithms for many classes of forward PDE simulations, inviting pursuit of the more challenging problem of inversion.

However, designing scalable parallel numerical algorithms for the solution of nonlinear PDE-based inverse problems poses a significant challenge. This is particularly true for large-scale inverse problems, in which the number of inversion parameters is mesh-dependent (for example when the parameters represent unknown initial conditions, distributed sources, or heterogeneous material properties). In this case, scalability—both to large problem sizes and large numbers of processors—is crucial. Newton's method—the archetypical scalable nonlinear solver—entails a number of difficulties for such problems:

- The Hessian matrices required by Newton's method are (formally) dense and of the order of the number of inversion parameters, and their construction involves numerous solutions of the forward problem. Thus, straightforward implementations for grid-based parameterizations require prohibitive memory, work, and communication.
- The denseness, large size, and difficulty of construction of the linearized inverse operators rule out direct solvers. For the same reasons, matrix-based iterative solvers are not practical. On the other hand, matrix-free implementations can be made to exploit problem structure, so that each matrix-vector product requires solution of just a few forward-like PDE problems. However, this freedom from having to form a matrix creates difficulties in constructing a preconditioner, which is essential when more sophisticated regularizers are used to produce a well-posed inverse problem.
- Objective functions for nonlinear inverse problems are often nonconvex; Newton's method will likely diverge without careful attention to globalization. Worse, for many nonlinear inverse problems, the objective functions to be minimized may have many local minima. In particular, for the inverse wave propagation problems targeted in this paper, they are highly oscillatory, with basins of attraction that shrink with the wavelength of the propagating waves. General-purpose global optimization methods are incapable of scaling to the large numbers of inversion parameters induced by grid-based parameterizations. On the other hand, local methods like Newton's become trapped at local minima unless the objective function can be convexified.

Given the above difficulties faced by Newton-like methods, it is not surprising that many large-scale PDE inverse methods are based on gradients alone and avoid Hessian matrices. The simplest such techniques are steepest descent-type methods, in which the current solution is improved by moving in the direction of the negative gradient of the objective function. Better directions can be constructed, still based on gradient information, via nonlinear conjugate gradient (CG) methods. Even better are limited memory quasi-Newton methods (such as L-BFGS), which build a limited amount of curvature information based on changes in the gradient, and reduce to nonlinear CG in certain limiting cases. All of these gradient-based methods avoid the difficulties brought on by large dense Newton Hessians mentioned above. Moreover, their parallel efficiency can be high, since the gradient computation involves PDE-like solves, for which there exists a large body of parallel algorithms and libraries on which to draw. However, unlike Newton methods, they do not scale well with increasing problem size, suffering a reduction from quadratic to linear asymptotic convergence, and perform poorly for ill-conditioned problems. Nevertheless, for moderately-sized problems (in both parameter and state spaces), they offer a reasonable compromise between performance and ease of implementation, and this explains their popularity.

However, there is a desire to solve ever-larger inverse problems in such areas as global climate change, weather forecasting, ocean modeling, air and groundwater pollutant trans-

port, earthquake modeling, astrophysics, hazardous substance attacks, non-destructive materials testing, medical imaging, electromagnetic sensing, turbulence modeling, reservoir simulation, and so on. In many of these areas, increasing confidence in the fidelity of PDE models, combined with the greater power of PDE solvers and computing hardware, have resulted in PDE simulations that are sufficiently predictive to invite inversion of highly-resolved and highly-parameterized models. For these large-scale problems, linearly convergent methods such as L-BFGS are usually not viable; instead we need algorithms that are scalable, both with increasing problem size and numbers of processors.

In this paper we present and assess the scalability of a parallel algorithm for time-domain PDE-based inverse problems. The target application is 3D inverse wave propagation, in which the goal is to determine coefficients of the material parameter field of a heterogeneous medium, given a source and waveform observations at receiver locations on its boundary. Inverse problems of this type arise in seismic exploration, earthquake modeling, ultrasound imaging, matched field processing, ocean acoustics, treaty verification, obstacle detection, nondestructive evaluation, and structural health monitoring. We present results for inversion of the velocity field of a scalar linear wave equation, which is representative of this class of inverse wave propagation problems and contains the major difficulties of rank-deficiency, ill-conditioning, large-scale, space-time coupling, and numerous local minima with basins of attraction that shrink with increasing frequency of propagating waves.

Because many of the inverse wave propagation problems of interest are "100-wavelength problems," i.e. there is a desire to reconstruct feature sizes that are several hundred times finer than characteristic dimensions, is it essential that inversion algorithms be able to scale to thousand-grid-point resolution in each spatial dimension, which for heterogeneous 3D problems results in billion-parameter inverse problems. For example, one of our target problems is the estimation of the elastic mechanical properties (Lamé moduli and density) of the greater Los Angeles Basin, given seismograms of past earthquakes in the region, an elastic wave propagation forward model, and models of the seismic sources. Terascale (and beyond) computing is needed for such large-scale problems, and thus scalability to large processor counts is essential.

To treat multiple local minima, we employ multiscale continuation, in which a sequence of initially convex, but increasingly oscillatory, approximations to the objective function are minimized over a sequence of increasingly finer discretizations of state and parameter spaces, which in our experience keeps the sequence of minimizers within the basin of attraction of the global minimum. Total variation regularization is used to address ill-posedness and preserve sharp interfaces in the inversion parameter space. To overcome the difficulties of large, dense, expensive-to-construct, indefinite Newton Hessians, we combine Gauss-Newton linearization with matrix-free inner Krylov iterations. These require good preconditioners in the case of the total variation regularization operator, which presents considerable difficulties due to its strong heterogeneity and anisotropy. Our preconditioner is based on the L-BFGS algorithm to approximate curvature information using changes in directional derivatives, initialized by second order stationary iterations applied to the total variation operator.

All of these computational components require linear work and parallelize well in a domain-based manner (i.e. fine granularity parallelism induced by a partition of the mesh). The remaining requirement for assuring overall scalability is mesh-independence of (or more realistically weak dependence on) the outer nonlinear and inner linear iterations. Newton methods often deliver mesh-independence for nonlinear operators for sufficiently-fine grids, and appropriately preconditioned inverse operators can produce spectra that are favorable for Krylov methods. However the total variation operator becomes increasingly ill-conditioned with decreasing mesh size, which often results in an increase of Newton iterations with mesh

size. In this paper we study the performance of our algorithm with respect to these scalability issues in the context of the inverse wave propagation problem.

We begin in Section 2 by formulating the inverse problem as a PDE-constrained optimization problem. In Section 3, we present the Newton-Krylov method, in infinite-dimensional strong form for simplicity of presentation. In Section 4, we give numerical results on the Terascale Computing System (TCS) at the Pittsburgh Supercomputing Center (PSC), and draw conclusions.

2. Inverse heterogeneous wave propagation. In this section, we formulate the inverse wave propagation problem as a problem in PDE-constrained optimization, and derive the first order necessary conditions that characterize its solution. The optimization approach to a general form of this inverse problem was first elaborated by Tarantola [14]. Consider an acoustic medium with domain Ω and boundary Γ. The medium is excited with a known acoustic energy source $f(x,t)$ (for simplicity we assume a single source event), and the pressure $u^*(x,t)$ is observed at N_r receivers, corresponding to points x_j in the domain. Our objective is to infer from these measurements the squared acoustic velocity distribution $q(x)$, which is a property of the medium. From now on we refer to q as the material model and $u(x,t)$ as the state variable. We can formulate the inverse problem as a nonlinear least squares parameter estimation problem with PDE constraints. We seek to minimize, over the period $t=0$ to T, an L_2 norm difference between the observed state and that predicted by the PDE model of acoustic wave propagation, at the N_r receiver locations. The PDE-constrained optimization problem is:

$$
\begin{aligned}
\min_{u,q} \frac{1}{2}\sum_{j=1}^{N_r} \int_0^T \int_\Omega \left[u^* - u\right]^2 \delta(x-x_j)\, d\Omega\, dt + \beta\mathcal{F}(q) \\
\text{subject to} \quad \ddot{u} - \nabla\cdot q\nabla u = f \quad \text{in} \quad \Omega\times(0,T] \\
q\nabla u\cdot n = 0 \quad \text{on} \quad \Gamma\times(0,T] \\
u = \dot{u} = 0 \quad \text{in} \quad \Omega\times\{t=0\}
\end{aligned}
\tag{2.1}
$$

The first term in the objective function is the misfit between observed and predicted states, the second term $\beta\mathcal{F}(q)$ symbolizes the regularization functional with regularization parameter β, and the constraints are the initial–boundary value problem for acoustic wave propagation./footnoteThe form in which the acoustic equation is given assumes a constant bulk modulus and variable density; our method is equally applicable to a variable bulk modulus and constant density, or the case when both are variable. Here we have chosen zero initial conditions and Neumann boundary conditions, but the algorithms we discuss in this paper extend to more general conditions (as well as more general wave propagation PDEs).

In the absence of regularization, the problem as formulated above is ill-posed: the solution may not be unique nor depend continuously on the given data [12]. Discretization of the inverse problem leads to inverse operators that are ill-conditioned and rank deficient. The null space of the inversion operator contains high frequency components of the material model q, which cannot be determined from the band-limited response u. Moreover, certain regions of the medium may be "hidden" as a result of source/receiver density and location, further enlarging the null space of the inverse operator. One remedy is to regularize with the L_2 norm of the gradient of the material model,

$$
\mathcal{F}(q) = \frac{\beta}{2}\int_\Omega \nabla q\cdot\nabla q\, d\Omega
$$

Such *Tikhonov* regularization eliminates the null space by penalizing oscillatory components of q; the more oscillatory, the greater the damping of those components. The effect becomes

more pronounced with larger values of β, which must be chosen large enough to damp out frequency components of q that are larger than those justified by the observations u^*, but not too large that true information is lost. See [15, Ch.7] for a discussion of techniques for choosing the value of β.

Tikhonov regularization will smooth discontinuities in the material model q, and is thus inappropriate for inverse problems with sharp interfaces, for example those between bone and soft tissue, sedimentary basin and rock, concrete and rebar, ocean water and sediment, or across cracks. For such problems, in (2.1) we choose the total variation (TV) regularization functional

$$\mathcal{F}^{TV}(q) = \beta \int_\Omega |\nabla q|_\varepsilon \, d\Omega,$$

where

$$|\nabla q|_\varepsilon = (\nabla q \cdot \nabla q + \varepsilon)^{\frac{1}{2}}.$$

$\mathcal{F}^{TV}$ is the bounded variation seminorm of the material model, modified by the parameter ε. The addition of ε makes the regularization term differentiable. TV regularization effectively preserves (line in 2D, surface in 3D) jump discontinuities in q, since the TV functional, unlike Tikhonov, is bounded in this case; at the same time, it damps out spurious oscillations in smooth regions [1]. The tradeoff is that the discrete TV operator is very ill-conditioned; it is effectively a nonlinear Poisson operator with widely-varying and highly anisotropic coefficient. For both Tikhonov and TV, we assume $\nabla q \cdot n = 0$ on Γ, i.e. the material model does not change in the normal direction on the boundary.

However, regularization does not cure all the problems. In addition to the null space problems described above, the objective function can be highly oscillatory in the space of material model q. Thus, there can be many local minima to contend with. Moreover, this oscillatory behavior becomes more pronounced with increasing frequency of propagating waves: the diameter of the basin of attraction of the global minimum varies as their wavelength, and local optimization methods become increasingly doomed to failure [13, 7].

To overcome this problem, we use multilevel grid and frequency continuation to generate a sequence of solutions that remain in the basin of attraction of the global minimum; that is, we solve for increasingly higher frequency components of the material model, on a sequence of increasingly finer grids with increasingly higher frequency sources [6]. Figure 2.1 shows an example of the TV-regularized solution of a synthetic inverse wave propagation problem, and the ability of multiscale continuation to keep Newton's method in the basin of attraction of the global minima. Note the sharp resolution of reconstructed interfaces on the finest grid (second-to-last frame). Figure 2.2 illustrates the effect of multiscale continuation and TV regularization. Without multiscale continuation, the optimization algorithm becomes trapped in a local minimum (Frame B) far from the global optimium. Frame C demonstrates the smoothing of interfaces due to the use of Tikhonov H^1 seminorm regularization.

The equations that determine the solution to the optimization problem (2.1) can be derived by requiring stationarity of a *Lagrangian functional*. These so-called optimality conditions take the form of a coupled nonlinear three-field system of integro-partial-differential equations in the state variable u, the *adjoint* variable λ, and the material model q:

$$\begin{aligned} \ddot{u} - \nabla \cdot q \nabla u &= f \quad \text{in} \quad \Omega \times (0,T], \\ q \nabla u \cdot n &= 0 \quad \text{on} \quad \Gamma \times (0,T], \\ u = \dot{u} &= 0 \quad \text{for} \quad \Omega \times \{t = 0\}, \end{aligned} \tag{2.2}$$

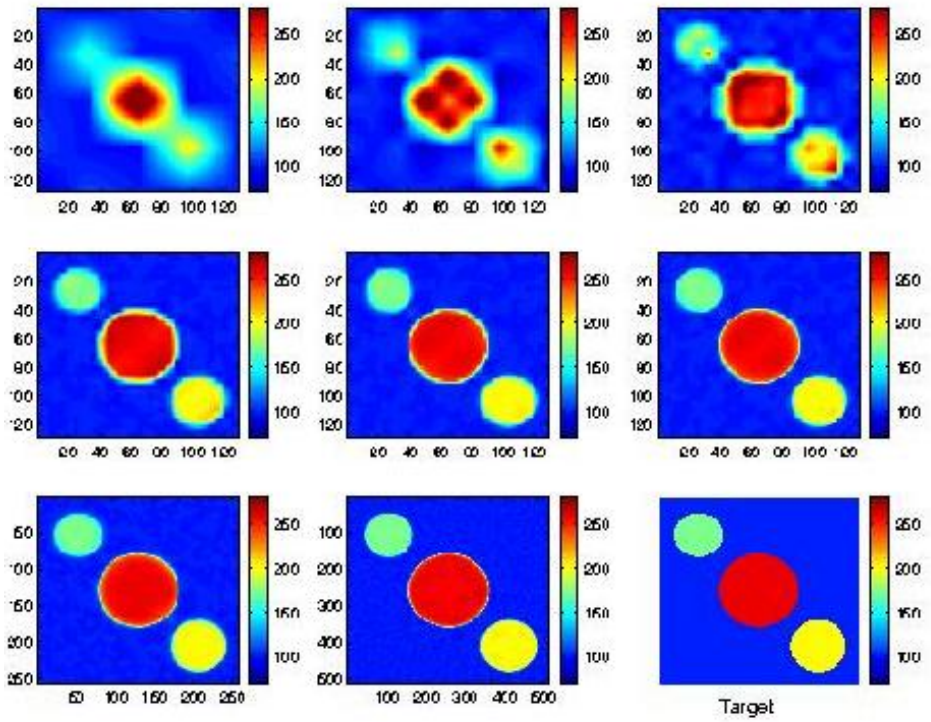


FIG. 2.1. *Inversion of a piecewise-homogeneous material model corresponding to three circular scatterers using TV regularization. Receiver locations on top, sources on bottom, receiver data is synthesized based on forward solution of target model shown in last (naturally-ordered) frame. Initial guess is a homogeneous medium with a single parameter; increasingly finer inversion models are computed on increasingly finer grids, with increasingly higher frequency of excitation, culminating with a* 512×512 *material and wave propagation grid in the next to last frame. This is almost indistinguishable from the target. Computations on 64 processors of PSC Cray T3E.*

$$
\begin{aligned}
\ddot{\lambda} - \nabla \cdot q \nabla \lambda - \sum_{i=1}^{N_r} [u^* - u]\, \delta(x - x_j) = 0 \quad &\text{in} \quad \Omega \times (0,T], \\
q \nabla \lambda \cdot n = 0 \quad &\text{on} \quad \Gamma \times (0,T], \\
\lambda = \dot{\lambda} = 0 \quad &\text{for} \quad \Omega \times \{t = T\},
\end{aligned}
\tag{2.3}
$$

$$
\begin{aligned}
\int_0^T \nabla u \cdot \nabla \lambda \, dt - \beta \nabla \cdot (|\nabla q|_\varepsilon^{-1} \nabla q) = 0 \quad &\text{in} \quad \Omega, \\
\nabla q \cdot n = 0 \quad &\text{on} \quad d\Gamma.
\end{aligned}
\tag{2.4}
$$

The state equation (2.2) is an initial value problem. Since the wave operator is self-adjoint in time and space, the adjoint equation (2.3) has the same form as the state equation, i.e. it is also an acoustic wave equation. However, there is a crucial difference: the adjoint wave equation is a *final value problem*, as seen in the final, as opposed to initial, conditions on λ in (2.3), and it has a different source, which depends on the state variable u. Finally, the material equation (2.4) is integro-partial-differential and time-independent. This last equation is shown for the case of TV regularization; for H^1 Tikhonov regularization, the term involving β would read $-\beta \nabla \cdot \nabla q$.

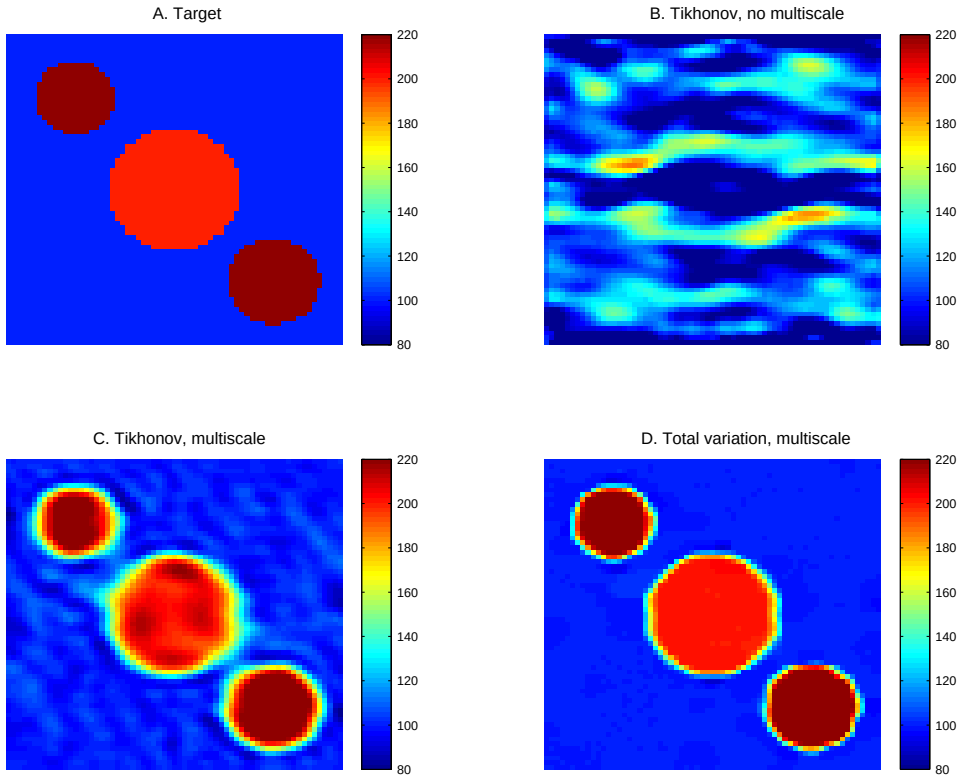


FIG. 2.2. *Effect of multiscale continuation and total variation regularization. Problem is same as in Figure 2.1, except* 64×64 *(finest) grid is used. Target field is in Frame A. Solution of inverse problem on a single grid, without benefit of multiscale continuation (in grid size and source frequency), converges to a local minimum very far from the global, as shown in Frame B. Multiscaling, but with Tikhonov regularization, locates the global optimum, but smears the interfaces, as shown in Frame C. Finally, Frame D depicts inversion with both multiscaling and TV regularization (further continuation produces the sharp reconstructed interfaces of Figure 2.1)*

.

The system (2.2)–(2.4) is an extremely formidable system to solve. When appropriately discretized, the dimension of each of u and λ is equal to the number of grid points N_g multiplied by time steps N_t, q is of dimension N_g, and thus the total system is of dimension $2N_gN_t + N_g$. This is can be very large for problems of interest—for example, in the largest problem solved in this paper, the system contains 3.4×10^9 unknowns. The time dimension cannot be "hidden" with the usual time-stepping procedures, since the system (2.2)–(2.4) contains coupled initial and final value problems; that is, both directions in time are explicit in this system. Therefore, in the next section we will discuss a method that solves in the reduced space of the material model, by eliminating the state and adjoint variables and equations from the system above.

3. A parallel reduced Gauss-Newton-Krylov method. Because the optimality system (2.2)–(2.4) is four-dimensional (three-space and time), we pursue a reduced space approach in which we consider the state u and adjoint λ as implicit functions of the material model q, and solve the material equation (2.4) for q, which is "just" in the 3D space of material parameters. This is represented symbolically by the so-called *reduced gradient*

$$\mathcal{G}(u, \lambda, q) = 0, \tag{3.1}$$

in which $u = u(q)$ and $\lambda = \lambda(u(q), q)$. This equation is nonlinear in q. Therefore, whenever we need to evaluate the reduced gradient $\mathcal{G}(q)$, we solve the state equation (2.2) for u given q, and then the adjoint equation (2.3) for λ given q and u. One difficulty is the time-reduction operator in (2.4) requires the state time history u, which is computed forward in time, and the adjoint time history λ, which is computed backward in time, neither of which we can afford to store (indeed that was one of the motivations for the reduction above). To overcome this problem, we use checkpointing ideas that originated in the automatic differentiation literature [9].

We solve the reduced gradient system (3.1) using Newton's method, which entails solving the Newton system

$$\mathcal{H}(q)\bar{q} = -\mathcal{G}(q), \tag{3.2}$$

for the Newton direction $\bar{q}$. The update is then $q^+ = q + \alpha\bar{q}$, where q is the current material model, q^+ is the updated model, and α permits some step along the Newton direction. The *reduced Hessian* operator when discretized gives rise to a dense matrix that is of dimension of the material parameterization. As mentioned in the introduction, we cannot afford to form, store, and factor this matrix. For example, in the example solved below, forming the matrix would require over an hour at a sustained petaflops/s; storing it would require 36 terabytes of memory, and factoring it would require several hours at a sustained petaflops/s. And the problem solved below is still *a factor of 500 smaller* than one of our ultimate target inverse wave propagation problems: inverting earthquake observational data for the mechanical properties of the Greater Los Angeles Basin.

Therefore, we solve the Newton system (3.2) with the linear conjugate gradient method. At each CG iteration, we require the Hessian-vector product $\mathcal{H}\hat{q}$, where $\hat{q}$ is the CG direction for updating $\bar{q}$. This product is given by an appropriate discretization of the Fréchet derivative of $\mathcal{H}$ in the direction of $\hat{q}$, i.e.

$$\int_0^T (\nabla\bar{\lambda}\cdot\nabla u + \nabla\lambda\cdot\nabla\bar{u})\; dt - \beta\nabla\cdot[\mathbf{K}(q)\nabla\hat{q}], \tag{3.3}$$

where given $(u, q, \hat{q})$, the incremental state variable $\bar{u}$ is found by solving forward-in-time the wave propagation problem

$$\begin{aligned} \ddot{\bar{u}} - \nabla\cdot(q\nabla\bar{u}) &= \nabla\cdot(\hat{q}\nabla u) \quad \text{in} \quad \Omega\times(0,T], \\ q\nabla\bar{u}\cdot n &= 0 \quad \text{on} \quad \Gamma\times(0,T], \\ \bar{u} = \dot{\bar{u}} &= 0 \quad \text{for} \quad \Omega\times\{t=0\}, \end{aligned} \tag{3.4}$$

and given $(\bar{u}, \lambda, \hat{q}, q)$, the incremental adjoint variable $\bar{\lambda}$ is the solution of the backward-in-time adjoint wave propagation problem

$$\begin{aligned} \ddot{\bar{\lambda}} - \nabla\cdot(q\nabla\bar{\lambda}) &= \nabla\cdot\hat{q}\nabla\lambda - \sum_{i=1}^{N_r}\delta(x-x_j)\bar{u}(x_j) \quad \text{in} \quad \Omega\times(0,T], \\ q\nabla\bar{\lambda}\cdot n &= 0 \quad \text{on} \quad \Gamma\times(0,T], \\ \bar{\lambda} = \dot{\bar{\lambda}} &= 0 \quad \text{for} \quad \Omega\times\{t=T\}. \end{aligned} \tag{3.5}$$

The time integral in (3.3) comes from the least squares misfit term in the objective; the second term comes from the TV functional, and includes an anisotropic coefficient tensor $\mathbf{K}(q)$ given by

$$\mathbf{K}(q) = \frac{1}{|\nabla q|_\varepsilon}\left(\mathbf{I} - \frac{\nabla q\otimes\nabla q}{|\nabla q|_\varepsilon^2}\right) \tag{3.6}$$

where

$$|\nabla q|_\varepsilon = (\nabla q \cdot \nabla q + \varepsilon)^{\frac{1}{2}}$$

The tensor is written as the product of two terms. The first is the heterogeneous coefficient $|gradq|_varepsilon^{-1}$, which when discretized varies from $\mathcal{O}(\frac{1}{h})$ near an interface to $\mathcal{O}(\frac{1}{\varepsilon^{0.5}})$ in homogeneous regions. The second term in the product is a tensor with eigenvectors normal and tangential to the interface, and eigenvalues equal to 1 and $1 - |\nabla q|^2 |\nabla q|_\varepsilon^{-2}$. For $\varepsilon = 0$, the second eigenvalue is 0, and the tensor is the projection of the laplacian on the interface. Thus, the TV operator smoothes the material field along the interface, but not normal to it. The addition of a small ε eliminates this defect in the tensor and provides a small amount of smoothing normal to the interface; this makes the nonlinear Poisson operator in (2.4) nonsingular but highly anisotropic. On the other hand, in homogeneous regions, the second eigenvalue is nearly 1, and thus the tensor is nearly isotropic. This combination of high heterogeneity (between smooth regions and interfaces), and high anisotropy (at material interfaces) suggests that the TV term's contribution to $\mathcal{H}$ will be very ill-conditioned for problems with strong material gradients. Finally, just as in computing the reduced gradient $\mathcal{G}(q)$, adjoint checkpointing ideas are used to manage the fact that $(u, \bar{u})$ are computed forward in time, and $(\lambda, \bar{\lambda})$ backward in time. These techniques permit one to trade off available memory for additional work.

Unfortunately, the reduced Hessian $\mathcal{H}$ is guaranteed to be positive definite only when sufficiently close to a minimum. Therefore, far from the solution, we use the *Gauss-Newton* approximation, which drops terms involving λ in (3.3) and in the adjoint wave propagation problem (3.5) and produces a positive definite reduced Hessian [11]. Even with this approximation, convergence can be near-quadratic if the PDEs are sufficiently good models of the data, and the data has sufficiently little noise [11], in which case we can expect a good fit and a small norm of λ, as can be seen from (2.3). In fact, the results presented in this paper are based on using the Gauss-Newton approximation exclusively. In this case, we can represent the reduced Hessian operator as

$$\mathcal{H}^{GN} = \mathcal{J}(q)^* \mathcal{J}(q) + \beta \mathcal{K}(q),$$

where $\mathcal{J}^*\mathcal{J}$ is the Gauss-Newton approximation of the least squares misfit term, and $\mathcal{K}$ is the TV regularization operator. Because of positive definiteness of the reduced Hessian, the Gauss-Newton method can be made globally convergent with a suitable line search strategy. We use a simple Armijo-based backtracking line search [11]. Finally, we terminate the inner CG iterations inexactly according to Eisenstat-Walker criteria [8].

Since forming the action of $\mathcal{H}$ or $\mathcal{H}^{GN}$ on $\hat{q}$ requires the solution of forward and adjoint wave propagation problems, i.e. the "inversion" of PDE operators, it is clear that the reduced Hessian will be dense. The matrix-free Gauss-Newton-Krylov method described above permits us to avoid forming and storing $\mathcal{H}$. However, this strategy will work only if relatively few iterations can be taken. This places a premium on preconditioning. The choice of a preconditioner is compounded by the complicated spectrum of the reduced Hessian. Its smooth eigenfunctions are associated with large eigenvalues and are dictated by the least squares misfit reduced Hessian $\mathcal{J}^*\mathcal{J}$, which is a compact operator, with eigenvalues that cluster at zero. The highly oscillatory eigenfunctions are dictated by the TV regularization operator $\mathcal{K}$. An effective preconditioner should address error components associated with both ends of the spectrum, as well as the middle where neither operator dominates and the eigenfunctions are influenced by both operators. Moreover, it is crucial that the preconditioner be cheap to apply; in this case, we insist that it not involve solutions of the wave equation on the fine grid, since these are very expensive.

Our preconditioning strategy addresses this structure of $\mathcal{H}^{GN}$ as follows. First, we employ an L-BFGS preconditioner that builds up a subspace of basis vectors for approximating the curvature of the objective function in the model space. Here we use a variant in which the curvature is approximated by Hessian-vector products collected during the CG iterations [10]. Since CG requires a constant preconditioner, this preconditioner is updated just once every Newton iteration. Even though L-BFGS is not effective as a solver for the inverse problem (as the results in the next section support), it may be effective as a preconditioner by capturing the smooth eigenfunctions of $\mathcal{J}^*\mathcal{J}$, since those are error components that the CG iteration tends to eliminate first (CG is a rougher for compact operators), and can be represented by a small dimension subspace.

L-BFGS requires an initialization of the inverse of the Hessian; for this we take several iterations of a Frankel method on the operator $\alpha\mathcal{I} + \beta\mathcal{K}$ whenever the inverse Hessian initializer is required. The Frankel method [2] is a two-step stationary iterative method that, like CG, requires matrix-vector products with the coefficient matrix at each iteration, and has convergence constant related to the square root of the condition number. Unlike CG, no inner products are required (which is attractive for large numbers of processors), and since it is stationary can be used within CG iterations. The drawback is that it require estimates of the extremal eigenvalues, but these can be obtained relatively inexpensively with several Lanczos iterations. Since $\mathcal{J}^*\mathcal{J}$ involves forward and adjoint wave propagation solutions, it is much too expensive to be used in the preconditioner, and we replace it with $\alpha\mathcal{I}$, which is meant to capture the relatively flat interior spectrum of $\mathcal{H}^{GN}$. We are working on replacing this with a coarse grid low rank approximation of $\mathcal{J}^*\mathcal{J}$, since compact operators can be well approximated on coarse subspaces, and "inverting" this approximation via the Sherman-Morrison-Woodbury formula with a multilevel solver for $\mathcal{K}$. Details will be provided in the final paper. Nevertheless, we have found the combination of L-BFGS preconditioning with a few iterations of Frankel's method to be very effective in the context of Navier-Stokes optimal flow control [4, 5].

In summary, the method has the following nested components:

- multiscale continuation on a sequence of finer grids with successively higher source frequencies, to capture increasingly finer components of the inversion model and keep the iterates in the basin of attraction of the global minimum;
- Gauss-Newton iteration for solving the nonlinear optimization problem at each grid and frequency level, with the option of maintaining a proper Newton reduced Hessian close to the solution for large misfit problems;
- Inexactly-terminated conjugate gradient solution of the Gauss-Newton linear system, the majority of which is dominated by a forward and an adjoint wave propagation solution, plus a sparse local rank three tensor product and time reduction involving gradients of the state and adjoint fields; and
- A limited memory BFGS preconditioner intended to approximate the smooth eigenfunctions of the reduced Hessian, initialized with several iterations of Frankel's two-step stationary method to approximate the inverse of an approximation of the remaining part of reduced Hessian.

4. Scalability results. We have implemented the inversion method described in the previous section on top of the PETSc library [3]. Spatial approximation is by Galerkin finite elements, in particular piecewise trilinear basis functions for the state u, adjoint λ, and material model q fields. For the class of wave propagation problems we are interested in, the Courant-limited time step size is on the order of that dictated by accuracy considerations, and therefore we choose to discretize in time via explicit central differences. Thus, the number of time steps is of the order of cube root of the number of grid points. Since we require time ac-

curate resolution of wave propagation phenomena, the 4D "problem dimension" scales with the $\frac{4}{3}$ power of the number of grid points.

The overall work is dominated by the cost of the CG iteration, which, because the preconditioner is time-independent, is dominated by the Hessian-vector product. With the Gauss-Newton approximation, the CG matvec requires the same work as the reduced gradient computation (3.1): a forward wave propagation, an adjoint wave propagation, possible checkpointing recomputations based on available memory, and the reduction of the state and adjoint spatio-temporal fields onto the material model space via terms of the form $\int \nabla u \cdot \nabla \lambda \; dt$. These components are all "PDE-solver-like," and can be effectively parallelized in a fine-grained domain-based way, using many of the building blocks of sparse PDE-based parallel computation: sparse grid-based matrix-vector products, vector sums, scalings, and inner products, etc. Indeed, we use routines from the PETSc parallel numerical library [3] for these operations.

We first report results of fixed-size scaling on the Cray T3E at the PSC (Compaq TCS timings will be presented at SC2002). We expect the overhead due to communication, synchronization, and sequential bottlenecks to be very low, since one of the key features of our method is that it recasts the majority of the work in solving the inverse problem in terms of explicitly-solved wave propagation problems, both forward and backward in time, and local tensor reduction operations to form the reduced gradient and Hessian-vector product. Because the communication patterns for these components are nearest-neighbor, and because there are no barriers to excellent load balance, we expect the code to scale well. There are also some inner products and global reduction operations, associated with each iteration of CG and with the application of the preconditioner, that require global communication. In a standard Krylov-based forward PDE solver, such inner products can start to dominate as processor counts reach into the thousands. Here, however, it is the "PDE solver" that is on the inside, and the (inversion-related) Krylov iterations that are on the outside. Communication costs associated with inner products are thus neglible. Table 4.1 demonstrates the good parallel efficieny obtainable for an eightfold increase in number of processors on a Cray T3E, for a fixed problem size of 262,144 material parameters, and the same number of state and adjoint unknowns per time step. The table show a mild decrease in parallel efficiency with

TABLE 4.1
Fixed-size scalability on a Cray T3E-900 for a 65^3 grid problem corresponding to a two-layered medium

processors	**grid pts/processor**	**time (s)**	**time/gridpts/proc (s)**	**efficiency**
16	16,384	6756	0.41	1.00
32	8192	3549	0.43	0.95
64	4096	1933	0.47	0.87
128	2048	1011	0.49	0.84

increasing problem size. Note the very coarse granularity (a few thousand grid points per processor) for the last few rows of the table. For many forward problems, one would prefer finer grid densities, to enable solution of larger problems. However, for inverse problems, we are necessarily compute–bound, since a sequence of many forward problems need to be solved, and one needs as much parallelism as possible. We are therefore interested in this range of granularity.

We should point out that this particular inverse problem presents a very severe test of scalability. For scalar wave propagation PDEs, discretized with low order finite elements on structured spatial grids (the grid stencils are very compact) and explicit central differences in time, there is very little workload for each processor in each time step. So while

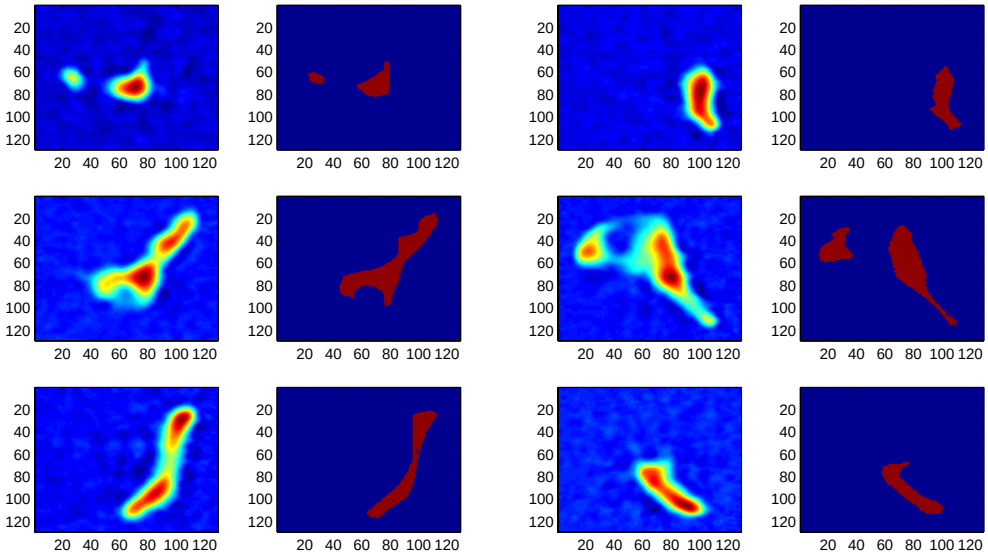
we can express the inverse method in terms of (a moderate number of) forward-like PDE problems, and while this means we follow the usual "volume computation/surface communication" paradigm, it turns out for this particular inverse problem (involving acoustic wave propagation), the computation to communication ratio is about as low as it can get for a PDE problem (and this will be true whether we solve the forward or inverse problem). A nonlinear forward problem, vector unknowns per grid point, higher order spatial discretization, unstructured meshes, or an implicit time stepping scheme—all of these would increase the computation/communication ratio and produce better efficiencies.

By increasing the number of processors with a fixed grid size, we have studied the effect of communication and load balancing on parallel efficiency in isolation of algorithmic performance. We next turn our attention to algorithmic scalability. We characterize the increase in work of our algorithm as problem size increases (mesh size decreases) by the number of inner (linear) CG and outer (nonlinear) Gauss-Newton iterations. The work per CG iteration involves explicit forward and adjoint wave propagation solutions, and their cost scales with the $\frac{4}{3}$ power of the number of grid points; a CG iteration also requires the computation of the integral in (3.3), which is linear in the number of grid points. Ideally, the number of linear and nonlinear iterations would be independent of the problem size; below, we examine this scalability for both Tikhonov and TV regularization.

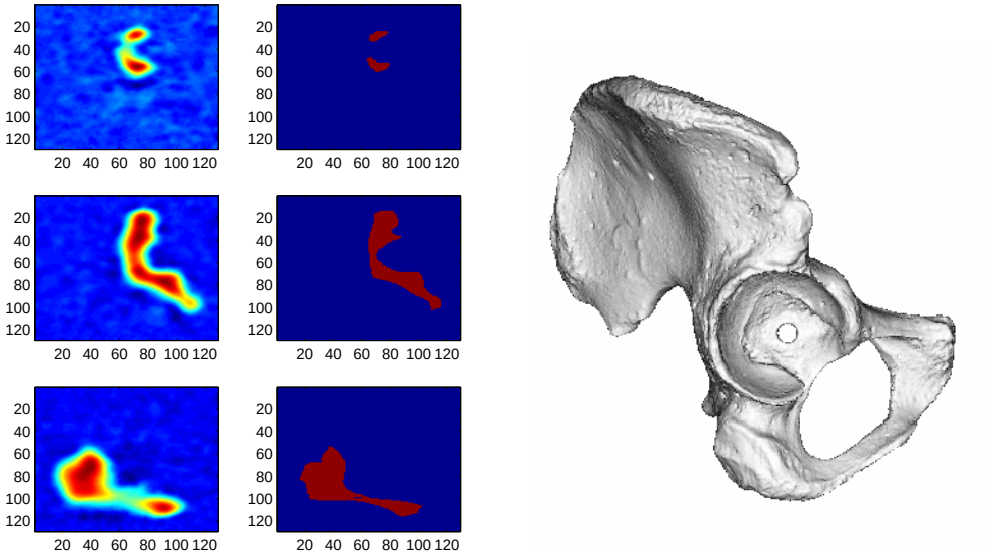
Figure 4.1 shows the result of inversion of surface pressure data for a synthetic velocity model that describes the geometry of a hemipelvic bone and surrounding volume. Figure 4.1d shows a target surface reconstruction, which was used to generate a piecewise-homogeneous (bone and surrounding medium) 3D material model. Cross sections through this 3D model are shown for three orthogonal planes, along with the results of the inversion for the finest grid, which is 129^3. Solution of the inverse problem at this grid level required 3 hours wall clock time on 128 TCS processors for Tikhonov regularization, and the same amount of time, but on 256 processors, for TV regularization (the images in the figure correspond to Tikhonov regularization). Note that unlike ultrasound imaging, here we are inverting for mechanical properties of a medium, not just the geometry of its interface. Note that for this problem, with 129 grid points in each direction, we are resolving at best 10 wavelengths in each direction, which is insufficient to extract fine features of the model. We expect that finer grids will resolve these features in greater detail.

Table 4.2 shows the growth in TV-regularized iterations for the limited memory BFGS (L-BFGS), unpreconditioned Gauss-Newton-Krylov (GNK), and preconditioned Gauss-Newton-Krylov (PGNK) methods as a function of material model resolution. L-BFGS was not able to converge for the 4913 and $35,937$ parameter problems in any reasonable amount of time, and larger problems weren't attempted with the method. This is not surprising; limited memory methods have difficulties with highly ill-conditioned problems. The Gauss-Newton methods showed mesh-independence of nonlinear iterations, until the finest grid. Finer grids produce greater heterogeneity and anisotropy of the "material" tensor of the TV regularization term, as can be seen by (3.6). This results in increasing ill-conditioning in the reduced Hessian, and appears to cause the increase in nonlinear iterations (see [16] for a discussion and examples of this behavior in the context of total variation denoising of images). On the other hand, the inner conjugate gradient iterations appear to be relatively constant per nonlinear iteration for PGNK. To verify that the inner iteration would keep scaling, we ran one nonlinear iteration on a 257^3 grid (nearly 17 million inversion parameters) on 2048 TCS processors. This required about 3 hours and 27 CG iterations, which is comparable to the smaller grids. This suggests that the preconditioner is very effective.

To verify that the TV term is causing the deterioration in nonlinear iterations, we resolved the above problem with Tikhonov regularization for a range of grid sizes from 33^3 to 129^3.



(a) (b)



(c) (d)

FIG. 4.1. *Performance of inversion algorithm on finest grid (129^3). Domain size $L = 100$; target velocity is 20 for bone, 10 for surrounding medium. Total duration is 20. 15×15 source/receiver arrays on four sides of the volume are used; source is a Ricker wavelet with central frequencies of 0.3 (65^3 wave propagation grid) and 0.6 (129^3 grid). Figure shows cross-sections of inverted and target models. (a) at x=0.25L, 0.5L, 0.75L; (b) at y=0.25L, 0.5L, 0.75L; (c) at z=0.25L, 0.5L, 0.75L; and (d) surface model of the target.*

TABLE 4.2

Algorithmic scaling with TV regularization for limited memory BFGS (L-BFGS), unpreconditioned Gauss-Newton-Krylov (GNK), and Gauss-Newton-Krylov-BFGS (PGNK) methods as a function of material model resolution. For L-BFGS, the number of iterations is reported, and for both L-BFGS solver and preconditioning, $m = 200$ L-BFGS vectors are stored. For GNK and PGNK, the total number of CG iterations is reported, along with the number of Newton iterations in parentheses. On all material grids up to 65^3, the forward and adjoint wave propagation problems are posed on 65^3 grid $\times$ 400 time steps, and inversion is done on 64 TCS processors; for the 129^3 material grid, the wave equations are on 129^3 grids $\times$ 800 time steps, on 256 processors. In all cases, work per iteration reported is dominated by a reduced gradient (L-BFGS) or reduced-gradient-like (GNK, PGNK) calculation, so the reported iterations can be compared across the different methods. The regularization and TV smoothing parameters are $\beta = 0.004$ and $\varepsilon = 0.01$. Convergence criterion is 10^{-5} relative norm of the reduced gradient. "" indicates lack of convergence; † indicates number of iterations extrapolated from converging value after 6 hours of runtime.*

grid size	material parameters	L-BFGS its	GNK its	PGNK its
65^3	8	16	17 (5)	10 (5)
65^3	27	36	57 (6)	20 (6)
65^3	125	144	131 (7)	33 (6)
65^3	729	156	128 (5)	85 (4)
65^3	4913	*	144 (4)	161 (4)
65^3	35,937	*	177 (4)	159 (6)
65^3	274,625	—	350 (7)	197 (6)
129^3	2,146,689	—	$1470^\dagger$ (22)	409 (16)

Indeed, the number of nonlinear iterations did not deteriorate with the 129^3 grid, and the linear iterations remained constant as well.

These results are encouraging, and demonstrate that the multiscale L-BFGS preconditioned Gauss-Newton-Krylov method seems to be scaling reasonably well for such a highly nonlinear and ill-conditioned problem as inverse wave propagation. The increase in nonlinear iterations with the TV-regularized Gauss-Newton method at the finest grid suggests a switch to a more robust (but just linearly convergent) Picard iteration [16] far from the optimum, coupled with (Gauss)-Newton once near the region of quadratic convergence. A coarse grid projection to approximate the compact part of the reduced Hessian, combined with a multilevel solver for the TV component, should help address the mild increase in CG iterations. Overall, we are able to solve a highly nonlinear inverse wave propagation problem with over 2 million unknown inversion parameters in just three hours on 256 TCS processors.

REFERENCES

[1] R. Acar and C. R. Vogel, *Analysis of bounded variation penalty methods for ill-posed problems*, Inverse Problems, Vol. 10, p. 1217–1229, 1994.

[2] O. Axelsson, *Iterative Solution Methods*, Cambridge Press, 1994.

[3] S. Balay, K. Buschelman, W.D. Gropp, D. Kaushik, L. Curfman McInnes, B.F. Smith, *PETSc Home Page*, `http://www.mcs.anl.gov/petsc`, 2001.

[4] G. Biros and O. Ghattas, *Parallel Newton-Krylov-Schur Methods for PDE Constrained Optimization. Part I: The Krylov-Schur Solver*, Technical Report, Laboratory for Mechanics, Algorithm, and Computing, Carnegie Mellon University, 2000.

[5] G. Biros and O. Ghattas, *Parallel Newton-Krylov-Schur Methods for PDE Constrained Optimization. Part II: The Lagrange Newton Solver*, Technical Report, Laboratory for Mechanics, Algorithm, and Computing, Carnegie Mellon University, 2000.

[6] C. Bunks, F. M. Saleck, S. Zaleski, and G. Chavent, *Multiscale Seismic Waveform Inversion*, Geophysics, Vol. 50, p. 1457-1473, 1995.

[7] G. Chavent and F. Clement, *Waveform Inversion Through MBTT Formulation.* Technical Report 1839, INRIA, 1993.

[8] S.C. Eisenstat and H.F. Walker, *Globally Convergent Inexact Newton Methods*, SIAM Journal on Optimization, Vol. 4, No. 2, p. 393–422, 1994.

[9] A. Griewank, *Achieving logarithmic growth of temporal and spatial complexity in reverse automatic differentiation*, Optimization Methods and Software, Vol.1, p. 35-54, 1992.
[10] J.L. Morales and J. Nocedal, *Automatic Preconditioning by Limited Memory Quasi-Newton Updating*, SIAM Journal on Optimization, Vol. 10, p. 1079-1096, 2000.
[11] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer, 1999.
[12] W. W. Symes, *Velocity Inversion: A Case Study in Infinite-Dimensional Optimization*, Mathematical Programming, Vol. 48 , p. 71-102, 1990.
[13] W. W. Symes and J. J. Carazzone, *Velocity Inversion by Differential Semblance Optimization*, Geophysics, Vol. 56, No. 5, p. 654-663, 1991.
[14] A. Tarantola, *Inversion of Seismic Reflection Data in the Acoustic Approximation*, Geophysics, Vol. 49, No. 8, p. 1259–1266, 1984.
[15] C. R. Vogel, *Computational Methods for Inverse Problems*, SIAM, Frontiers in Applied Mathematics, 2002.
[16] C.R. Vogel and M.E. Oman, *Iterative Methods for Total Variation Denoising*, SIAM Journal on Scientific Computing, Vol. 17, p. 227-238, 1996.