Seismic waveform inversion with Gauss-Newton-Krylov method

Yogi A. Erlangga and Felix J. Herrmann, UBC-Seismic Laboratory for Imaging and Modeling*

SUMMARY

This abstract discusses an implicit implementation of the Gauss-Newton method, used for the frequency-domain full-waveform inversion, where the inverse of the Hessian for the update is never formed explicitly. Instead, the inverse of the Hessian is computed approximately by a conjugate gradient (CG) method, which only requires the action of the Hessian on the CG search direction. This procedure avoids an excessive computer storage, usually needed for storing the Hessian, at the expense of extra computational work in CG. An effective preconditioner for the Hessian is important to improve the convergence of CG, and hence to reduce the overall computational work.

INTRODUCTION

Full waveform inversion–either in the time domain or the frequency domain–is one of the important methods for obtaining information on the earth's subsurface structure. Given data at the receivers, an image is obtained by minimizing the misfit between the data and the predicted wavefields, which are computed from a mathematical model, usually partial differential equations (PDEs), that governs the wave propagation in the earth (Tarantola, 1984). This process is an instance of PDE-constrained optimization, which requires methods to solve the underlying PDEs associated with the state and the adjoint variables, and methods to systematically update the model. In time-domain waveform inversions, the state variables are obtained by solving an initial-boundary value problem, to propagate the initial wavelet forward in time. The adjoint variables, however, are the solution of a final-boundary value problem, which propagate residuals back in time. The latter introduces difficulties because it in principle requires all time history, which is impossible to store. One remedy is to use checkpointing (Griewank, 1992; Symes, 2007).

Frequency-domain waveform inversions, on the other hand, do not have "history" problems. Furthermore, while in their formalism, all frequencies in the band should be included, inversion can be performed by using only a subset of frequencies (Sirgue and Pratt, 2004; Mulder and Plessix, 2004; Herrmann et al., 2008). Typically, the inversion is performed by solving optimization problems of increasing frequencies (Pratt, 1999; Plessix, 2006). This offers additional advantages of saving computational work at low frequencies–because these problems require less grid points–and of enlarging the basin of attraction. One disadvantage is that the associated state and adjoint equations are now of implicit forms, which are also known to be more difficult to solve. Only for a not-too-large 2-D problem that the inverse of the operators of these equations can be formed explicitly.

One important issue in the waveform inversion is the convergence rate of the inversion method. A descent method, which only requires the gradient of the functional to be minimized, converges only linearly. A faster convergence can be attained by using a Newton-like method (Nocedal and Wright (1999) and examples in seismic applications in Pratt et al. (1998)). This method can exhibit a (or near) quadratic convergence, provided that the initial model is close enough to the minimizer, and that the inverse of the Hessian (the second derivative of the functional) is available. For large problems–as in seismic waveform inversions–the latter is in general dense, and it is impossible to store.

In this paper, we discuss an implicit inversion of the Hessian in the Gauss-Newton framework used for the frequency-domain waveform inversion. This inversion is based on an iterative method (conjugate gradient (CG)), which requires only an action of the Hessian on the CG search direction that updates the CG approximation to the second variation of the model. The resultant method is an instance of Gauss-Newton-Krylov methods. We show that a useful inversion result can be obtained by using this procedure with a small number of updates, and with less CPU time than standard gradient methods for the same quality of results. Combined with an iterative method for solving the state and adjoint equations (Erlangga et al., 2006; Erlangga and Nabben, 2007; Lin et al., 2008; Erlangga and Herrmann, 2008), a matrix-free seismic waveform inversion algorithm can be designed. For the time-domain counterpart, see Akcelik et al. (2002).

FULL-WAVEFORM INVERSION

The frequency-domain waveform inversion can be formulated as follows. Given the data $u_d(x_r,\omega) \in \mathbb{C}$, with ω the angular frequency and $x_r \in \Omega_R \subset \Omega$ the receiver position, find $q = 1/c^2(x) \in Q \subset \mathbb{R}$, $x \in \Omega$ ($c(x)$ the velocity model) such that

$$\min_q \int_\Omega \frac{1}{2}\sum_{f=1}^{N_f}\sum_{s=1}^{N_s}\sum_{r=1}^{N_r} \overline{(u_d - u_{f,s})}(u_d - u_{f,s})\delta(x-x_r)d\Omega, \quad (1)$$

subject to

$$\mathscr{A}_f u_{f,s} := -\nabla\cdot\nabla u_{f,s} - \omega_f^2 q u_{f,s} = b_s, \quad \forall s,f, \quad (2)$$

$u \in U \subset \mathbb{C}$, that are equipped with proper boundary conditions. ($\bar{\cdot}$ indicates the complex conjugation.) Here, N_s, N_r, and N_f are respectively the number of shots, receivers, and frequencies.

The usual way to solve the above minimization problem is via the Lagrange functional (Nocedal and Wright (1999), Vogel (2002)):

$$\begin{aligned} &\mathscr{L}(u,v,q) \\ = \quad & \int_\Omega \frac{1}{2}\sum_{f=1}^{N_f}\sum_{s=1}^{N_s}\sum_{r=1}^{N_r} \overline{(u_d - u_{f,s})}(u_d - u_{f,s})\delta(x-x_r)d\Omega \\ - \quad & \int_\Omega v_{f,s}^T\left(-\nabla\cdot\nabla u_{f,s} - \omega_f^2 q u_{f,s} - b_s\right)d\Omega \end{aligned}$$

$$
\begin{aligned}
&= \int_\Omega \frac{1}{2}\sum_{f=1}^{N_f}\sum_{s=1}^{N_s}\sum_{r=1}^{N_r}\overline{(u_d-u_{f,s})}(u_d-u_{f,s})\delta(x-x_r)d\Omega \\
&- \sum_{f=1}^{N_f}\sum_{s=1}^{N_s} v_{f,s}^T\left[\int_{\partial\Omega=\Gamma}\nabla u_{f,s}d\Gamma\right. \\
&- \left.\int_\Omega\left(\nabla v_{f,s}^T\cdot\nabla u_{f,s}-\omega_f^2 v_{f,s}^T q u_{f,s}-v_{f,s}^T b_s\right)\right]d\Omega, \quad (3)
\end{aligned}
$$

after making use of the divergence theorem, with $v \in V \subset \mathbb{C}$ the Lagrange multiplier. A descent method computes the minimizer $\{u^*, v^*, q^*\}$ via an update $\{u,v,q\} \leftarrow \{u,v,q\} + \gamma g$, with g the search direction, which is the negative of the gradient, and γ the step length. This minimization however has to be done over the large space $U \times V \times Q$. The computational burden can be reduced by considering the *reduced* gradient

$$
\delta_q\mathscr{L}(u,v,q)[\hat{q}] = \sum_{f=1}^{N_f}\sum_{s=1}^{N_s}\int_\Omega \omega_f^2 v_{f,s}^H \hat{q} u_{f,s} d\Omega =: -g, \quad (4)
$$

the first variation of the Lagrange functional with respect to the model q, if the conditions that $u_{f,s}, v_{f,s}$ solve respectively (2) and the adjoint equations

$$
\mathscr{A}_f^T v_{f,s} = \mathscr{A}_f v_{f,s} = \sum_{r=1}^{N_r}\overline{(u_d-u_{f,s})}\delta(x-x_r), \quad \forall s,f, \quad (5)
$$

for an estimate q are imposed. In this case, the optimization is now carried out only over the model space Q with updates $q \leftarrow q - \gamma\delta_q\mathscr{L}(u,v,q)$. This procedure however results in a linear convergence towards the minimizer. This convergence is suboptimal and may prove prohibitive in practice.

A faster convergence can be attained by considering Newton updates $\{u,v,q\} \leftarrow \{u,v,q\} + \{\tilde{u},\tilde{v},\tilde{q}\}$, where $\{\tilde{u},\tilde{v},\tilde{q}\}$ solves the system

$$
\delta^2\mathscr{L}(u,v,q)[\hat{u},\hat{v},\hat{q};\tilde{u},\tilde{v},\tilde{q}] = -\delta\mathscr{L}(u,v,q)[\hat{u},\hat{v},\hat{q}], \quad (6)
$$

with $\delta^2\mathscr{L}$ the second variation of the Lagrangian. First, the system (6) requires the solution from the space $U \times V \times Q$. The size of the problem can be reduced by imposing the condition that the current approximation u and v solve the state and the adjoint equations (2) and (5), respectively. Consequently, we have $\delta_v\mathscr{L} = \delta_u\mathscr{L} = 0$. Also remark that unless the current approximation is close enough to the minimizer, the system in Eq. (6) is not guaranteed to be positive definite. We obtain a symmetric positive definite system by dropping the terms with v from Eq. (6). This results in the Gauss-Newton update. With these two conditions, the equation (6) can be reduced to

$$
\mathscr{H}^{GN}(u,v,q)[\hat{q}] = -\delta_q\mathscr{L} = g, \quad (7)
$$

with $\mathscr{H}^{GN}$ the Schur complement of the $(\tilde{q},\tilde{q})$ block in $\delta^2\mathscr{L}$, also called the *reduced* Hessian. For instance, with $N_f = 1$, and $N_s = 2$ (one frequency and two shots), the *full* Gauss-Newton system reads

$$
\begin{bmatrix}
\tilde{\mathscr{W}}_1 & 0 & \tilde{\mathscr{A}}^T & 0 & 0 \\
0 & \tilde{\mathscr{W}}_2 & 0 & \tilde{\mathscr{A}}^T & 0 \\
\tilde{\mathscr{A}} & 0 & 0 & 0 & \tilde{\mathscr{P}}_1 \\
0 & \tilde{\mathscr{A}} & 0 & 0 & \tilde{\mathscr{P}}_2 \\
0 & 0 & \tilde{\mathscr{Q}}_1 & \tilde{\mathscr{Q}}_2 & 0
\end{bmatrix}
\begin{pmatrix} \tilde{u}_1 \\ \tilde{u}_2 \\ \tilde{v}_1 \\ \tilde{v}_2 \\ \tilde{q} \end{pmatrix}
=
\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ -\delta_q\mathscr{L} \end{pmatrix}, \quad (8)
$$

where $\tilde{\mathscr{P}}_s = \tilde{\mathscr{Q}}_s^T = \int_\Omega \omega^2 \tilde{v}_s^T \hat{q} u_s d\Omega$, which gives

$$
\mathscr{H}^{GN}(u,v,q) = \sum_{s=1}^{2}\tilde{\mathscr{Q}}_s^T\tilde{\mathscr{A}}^{-T}\tilde{\mathscr{W}}_s\tilde{\mathscr{A}}^{-1}\tilde{\mathscr{Q}}_s. \quad (9)
$$

(The definitions of each operator are not derived here and only given in the Appendix.) Clearly, the operator (9) is symmetric positive definite (spd). Furthermore, if $\{u,v,q\}$ and $\{\tilde{u},\tilde{v},\tilde{q}\}$ are known, the action of the $\mathscr{H}^{GN}$ on $\tilde{q}$, discussed in the next section, can be easily determined.

NUMERICAL PROCEDURE

We first present the Gauss-Newton method in Algorithm 1 below, starting with the initial model q^0.

Algorithm 1. Gauss-Newton

Choose q^0;
Do $i = 0, 1, \ldots$ until convergence
  With q^i, solve the state equation for $u^i_{f,s}$, $\forall f, s$
  With $u^i_{f,s}$ and q^i, solve the adjoint equation for $v^i_{f,s}$
  With $u^i_{f,s}$, $v^i_{f,s}$ and q^i, compute g^i;
  Solve $\mathscr{H}^{GN,i}[\tilde{q}^i] = g^i$ for $\tilde{q}^i$;
  $q^{i+1} := q^i + \gamma\tilde{q}^i$;
  If $\|g^i\| \leq$ tol then convergent;
end

An important numerical procedure in the implicit Gauss-Newton method is the iterative method for spd matrices to approximately solve Eq. (7). An iterative method has an advantage that it only requires an action of the to-be-inverted operator on a vector. Since $\mathscr{H}^{GN}$ is spd, conjugate gradient (CG, Axelsson, 1994) is the method of choice. In CG applied to Eq. (7), the new approximation to $\tilde{q}$ is done via an update: $\tilde{q} \leftarrow \tilde{q} + \alpha w$, where $w := \mathscr{H}^{GN}[p]$ is compute via Eq. (9) with p the CG search direction. To do this, we need a forward solve associated with the operator $\tilde{\mathscr{A}}^{-1}\tilde{\mathscr{Q}}_s$ and an adjoint solve associated with $\tilde{\mathscr{Q}}_s^T\tilde{\mathscr{A}}^{-T}$ to compute the second variations, $\tilde{u}$ and $\tilde{v}$, based on u, v, and q at the i-th Gauss-Newton update and $\tilde{q} = p$ at the j-th CG update. This procedure is summarized below.

Algorithm 2. Computing the action of the Hessian $\mathscr{H}^{GN}$

At the i-th Gauss-Newton step, with $u^i_{f,s}$, q^i and $v^i_{f,s} = 0$, $\forall i$, and at the j-th CG step:
  With $\tilde{q}^i_j$, solve the state equation for $\tilde{u}^i_{f,s,j}$, $\forall f, s$;
  With $\tilde{u}^i_j$, solve the incremental adjoint equation for $\tilde{v}^i_{f,s,j}$;
  Compute Equation (9).

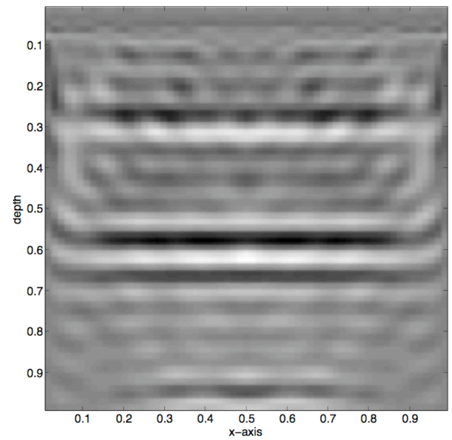
EXAMPLE

We consider an "academic" example based on a 1-D (hard) velocity model with two reflectors to show the concept. In this model, the domain is scaled into a unit square domain, and the frequency and velocity are then scaled into the reduced frequency $k = \omega L/c$, with L a characteristic length. The model space Q is discretized into 64×64 grid points. Synthetic data for this model are obtained from 64 shot positions, recorded at 64 receiver positions. The initial model for the inversion is a constant velocity background obtained by removing the two reflectors from the hard model. Figure 1 compares the inversion results based on the gradient method and the implicit Gauss-Newton method. In both cases, we did not use an optimal step length, which can be determined, for example, via a backtracking procedure. Instead we fixed the step length throughout the iterations. In the Gauss-Newton method, the Hessian is inverted by 8 unpreconditioned CG iterations. Figure 1(a) shows the inversion result after the two-norm of the gradient is reduced below 10^{-6} by only 16 Gauss-Newton updates. Figure 1(b) is the inversion result with 500 gradient updates ($\|g\|_2 = 2.7 \times 10^{-5}$). About 9 and 87 seconds of CPU-time per update are needed by the gradient and the Gauss-Newton method, respectively. Overall, the Gauss-Newton method requires 1400 seconds to converge, less than 4500 seconds needed by 500 gradient steps. With only 16 gradient updates, the inversion result is far from accurate (Figure 1(c)) and the second reflector is not yet recovered.

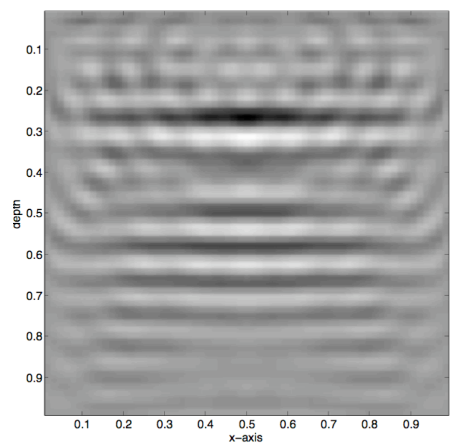
DISCUSSION

Inversion results shown here are not surprising: Newton-based directions provide a very good search direction for an update and – under certain conditions – guarantee a fast convergence. The main aspect of this paper, however, is the iterative procedure (CG) to invert the Hessian, which translates to performing the action of the Hessian on the CG search direction, and which avoids an explicit computation and excessive storage of the inverse of the Hessian. Clearly the rate of convergence will depend on the accuracy of this implicit inversion of the Hessian. We have two options: either to allow more CG iterations or to use a preconditioner for the Hessian. The first remedy may not be computationally efficient because at every CG iteration, $2N_sN_f$ state and adjoint equation solves have to be performed. This procedure also allows a matrix-free implementation, where no matrix (either Hessian or the state and adjoint equations) has to be stored. Furthermore, the wavefields u and v need not be stored and can be computed on the fly. We note that the state and adjoint operators are commonly inverted explicitly by an LU-type factorization, which has to be done in each Gauss-Newton and CG iteration because of the change of q. With an iterative method as proposed in Erlangga et al. (2006) and Erlangga and Nabben (2007), expensive computations of this factor and extra memory to store it can also be avoided. Of importance is that in the frequency-domain waveform inversion the Hessian is embarrassingly parallel in frequency and shot; cf. Eq. (9). It is therefore possible to compute each term in the Hessian individually in a parallel machine and then gather them.

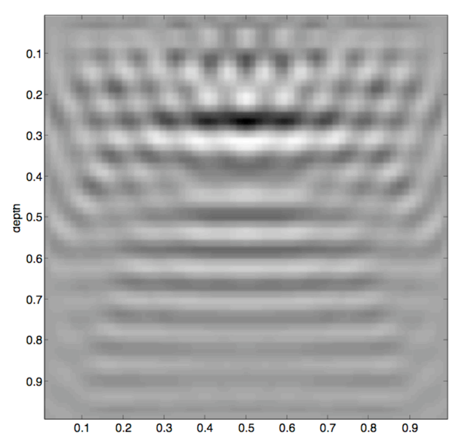
Better inversion results can be expected by including a penalty term (e.g., Tikhonov or Total Variation (TV)) in the minimization functional. TV is especially suitable for models with edges or discontinuities, because of the boundedness of the TV functional and its ability to damp oscillations in the smooth region. Algorithm-wise, we note that, with a penalty term, the $(\tilde{q}, \tilde{q})$ block is no longer a zero block, and the reduced Hessian (7) will have an additive term associated with this penalty functional, whose implementation is straightforward. What is more important, however, is the fact that seismic inversion tries to focus the energy at the zero offset. A spread of energy towards nonzero offset creates defocusing, which appears in the image matrix as nonzero off diagonals (the diagonal is the energy focused at the zero offset, the image). This energy spreading is



(a)



(b)



(c)

Figure 1: Full waveform inversion: **(a)** Implicit Gauss-Newton method at convergence, **(b)** gradient method after 500 iterations ($\|g\|_2 = 2.7 \times 10^{-5}$), **(c)** gradient method after 16 iterations (the same number of iterations required by the Gauss-Newton method to converge).

typically produced by a correlation-based approach, which is equivalent to a gradient method without a penalty term. The Hessian in the Gauss-Newton method is one way to penalize the energy at the nonzero offset and to focus it at the zero offset, which can be improved further by TV. In this direction, an alternative to TV is the use of an ℓ_1-norm-based minimization, which promotes sparsity by minimizing the energy at the nonzero offset, corresponding to the off-diagonal terms in the image matrix. This minimization, however, has to be carried out over a matrix space associated with the model, and requires an extension of the standard, vector-space, ℓ_1 solvers. Herrmann (2009) (this conference) shows that with a mixed (1,2)-norm solver, which penalizes the energy at the far offset and exploits joint sparsity at the near offset, this approach can be a viable alternative to TV. This approach is also conducive to undersampled data (in this case, image), which can bring a significant reduction in the size of the problem.

ACKNOWLEDGMENT

This work was partly supported by NSERC Grant and CRD Grant DNOISE.

APPENDIX

Definitions of operators in equation (9).

$$\tilde{\mathscr{W}}_{f,s} = \int_\Omega \sum_{r=1}^{N_r} \overline{\tilde{u}}_s \hat{u}_s \delta(x - x_r) d\Omega. \tag{10}$$

$$\tilde{\mathscr{A}}_f = \int_\Omega \hat{v}_s \left(-\nabla \cdot \nabla \tilde{u}_s - \omega^2 q \tilde{u}_s \right) d\Omega. \tag{11}$$

$$\tilde{P}_{f,s} = -\int_\Omega \omega_f^2 \hat{v}_{f,s} \tilde{q} u_{f,s} d\Omega. \tag{12}$$

$$\tilde{Q}_{f,s} = -\int_\Omega \omega_f^2 \tilde{v}_{f,s} q \hat{u}_{f,s} d\Omega. \tag{13}$$

REFERENCES

Akcelik, V., G. Biros, and O. Ghattas, 2002, Parallel multiscale Gauss-Newton-Krylov methods for inverse wave propagation: Presented at the Proceedings of the IEEE/ACM SC2002 Conference.

Axelsson, O., 1994, Iterative solution methods: Cambridge Press.

Erlangga, Y. A., and F. J. Herrmann, 2008, An iterative multilevel method for computing wavefields in frequency-domain seismic inversion: SEG Technical Program Expanded Abstract, 1957–1960.

Erlangga, Y. A., and R. Nabben, 2007, On multilevel projection Krylov method for the preconditioned Helmholtz system: Electronic Transaction of Numerical Analysis. (to appear).

Erlangga, Y. A., C. W. Oosterlee, and C. Vuik, 2006, A novel multigrid-based preconditioner for the heterogeneous Helmholtz equation: SIAM Journal on Scientific Computing, **27**, 1471–1492.

Griewank, A., 1992, Achieving logaritmic growth of temporal and spatial complexity in reverse automatic differentiation: Optimization Method and Software, **1**, 35–54.

Herrmann, F. J., 2009, Compressive imaging via group sparsity: Presented at the SEG Technical Program Extended Abstract, this conference.

Herrmann, F. J., Y. A. Erlangga, and T. T. Y. Lin, 2008, Compressive simultaneous full-wave form simulation: Geophysics. (to appear).

Lin, T. T. Y., E. Lebed, Y. A. Erlangga, and F. J. Herrmann, 2008, Interpolating solutions of the Helmholtz equation with compressed sensing: SEG Technical Program Expanded Abstract, 2122–2126.

Mulder, W. A., and R. E. Plessix, 2004, How to choose a subset of frequencies in frequency-domain finite-difference migration: Geophysical Journal International, **158**, 801–812.

Nocedal, J., and S. J. Wright, 1999, Numerical optimization: Springer.

Plessix, R. E., 2006, A review of the adjoint-state method for computing the gradient of a functional with geophysical applications: Geophysical Journal International, **167**, 495–503.

Pratt, R. G., 1999, Seismic waveform inversion in the frequency domain, Part 1: Theory and verification in a physical model: Geophysics, **64**, P888–P901.

Pratt, R. G., C. Shin, and G. J. Hicks, 1998, Gauss-Newton and full Newton methods in frequency-space seismic waveform inversion: Geophysical Journal International, **133**, 341–362.

Sirgue, L., and R. G. Pratt, 2004, Efficient waveform inversion and imaging: A strategy for selecting temporal frequencies: Geophysics, **69**, P231–P241.

Symes, W. W., 2007, Reverse time migration with optimal checkpointing: Geophysics, **72**, SM213–SM221.

Tarantola, A., 1984, Inversion of seismic reflection data in the acoustic approximation: Geophysics, **49**, 1259–1266.

Vogel, C. R., 2002, Computational methods for inverse problems: SIAM.