Journal of Physics: Conference Series **386** (2012) 012013 doi:10.1088/1742-6596/386/1/012013

The truncated Newton method for Full Waveform Inversion

L. Métivier[1], R. Brossier[1], J. Virieux[1], S. Operto[2]

[1]ISTerre, Université de Grenoble, BP 53, 38041 Grenoble CEDEX 9, France, [2]Géoazur, La Darse, B.P. 48, 06235 Villefranche sur Mer CEDEX, France

E-mail: ludovic.metivier@ujf-grenoble.fr

Abstract. Full Waveform Inversion (FWI) is a promising seismic imaging method. It aims at computing quantitative estimates of the subsurface parameters (bulk wave velocity, shear wave velocity, rock density) from local measurements of the seismic wavefield. Based on a particular wave propagation engine for wavefield estimation, it consists in minimizing iteratively the distance between the predicted wavefield at the receivers and the recorded data. This amounts to solving a strongly nonlinear large scale inverse problem. This minimization is generally performed using gradient-based methods. We investigate the possibility of applying the truncated Newton (TrN) method to this problem. This is done through the development of general second-order adjoint state formulas that yield an efficient algorithm to compute Hessian-vector products, and the design of an adaptive stopping criterion for the inner conjugate gradient (CG) iterations. Numerical results demonstrate the interest of using the TrN method when multi-scattered waves dominate the recorded data.

1. Introduction

Full Waveform Inversion (FWI) computes quantitative estimates of the subsurface parameters, such as the bulk wave velocity, the shear wave velocity, or the rock density. Applications range from the localisation of natural resources, such as oil and gas, to reservoir and storage monitoring, and seismic risk prevention. The seismic data is collected through the so-called seismic experiment: several sources and receivers are located at the ground surface or in wells, and the wavefield generated by these sources is recorded locally by the receivers. Based on a wave propagation modeling (from the simplest acoustic equation to more sophisticated anisotropic visco-elastic dynamics), the FWI method consists in minimizing a misfit function, which measures a distance between the data predicted by the forward problem and the recorded data. We consider the general wave equation

$$S(p)u = \varphi, \tag{1}$$

where the subsurface parameters are denoted by $p \in \mathbb{R}^m$ (model space), the linear forward problem operator corresponding to the two-way wave equation [1] is denoted by $S(p)$, the source vector is denoted by φ, and the wavefield vector is denoted by u. These notations are general

[1] Note that the operator $S(p)$ depends non-linearly on parameter p

and can be applied either in the time domain or in the frequency domain. The FWI problem is expressed as the nonlinear least-square problem

$$\min_p f(p) = \frac{1}{2}\sum_{s=1}^{N_s} \|R_s u_s(p) - d_s\|^2, \tag{2}$$

where d_s and $u_s(p)$ are respectively the recorded dataset and the solution of the forward problem associated with the source φ_s, N_s is the total number of sources, and R_s is a restriction operator that maps the wavefield u_s to the receiver locations.

The minimization of $f(p)$ is based on the local Newton approach. A sequence p_k is computed from an initial guess p_0 using the update formula

$$p_{k+1} = p_k + \gamma_k \Delta p_k, \tag{3}$$

where the scalar γ_k is computed through a globalization method (linesearch, trust-region) and Δp_k is the solution of

$$\nabla^2 f(p_k)\Delta p_k = -\nabla f(p_k), \tag{4}$$

where the Hessian operator is denoted by $\nabla^2 f(p_k)$. The computation of the gradient $\nabla f(p)$ can be performed efficiently with the adjoint-state method [6]. Because of the large-scale aspect of the FWI problem (even for 2D applications, the problem often involves hundred thousands of unknowns parameters and discrete data), explicit computation and storage of $\nabla^2 f(p)$, *a fortiori* $(\nabla^2 f(p))^{-1}$, is prohibitive. Therefore, standard methods use an approximation Q_k of the inverse Hessian. Different choices can be made. For instance, choosing simply the identity matrix yields the steepest-descent method, which may converge slowly. A more sophisticated and efficient choice is the l-BFGS approximation [1]. This method estimates the effect of $\left(\nabla^2 f(p_k)\right)^{-1}$ on the gradient. At iteration k, this approximation is based on finite differences of the l previous values of the gradient.

However, Pratt et al [7] emphasize the crucial role of the inverse Hessian operator in FWI: its acts as a defocusing filter, improving the resolution of the subsurface parameter estimation, and can help to remove artifacts on the model update related to the presence of large amplitude double-scattered waves. Accounting more accurately for effects of the inverse Hessian operator may thus yield significant improvements. As a consequence, we present in this study an implementation of a matrix-free truncated Newton (TrN) method [5] for FWI. Instead of using an approximation of the inverse Hessian operator, this method partially solves the linear system (4) using a matrix-free conjugate gradient (CG) algorithm. This requires the capability of computing efficiently Hessian-vector products. This can be achieved using the second-order adjoint state formulas we present in the second section. In addition, an efficient adaptive stopping critertion for the CG iterations must be designed, in order to prevent from oversolving the equation (4), which would generate prohibitive computation costs. The Eisenstat stopping criterion [2] fulfills this requirement. Within this framework, numerical examples presented in the third section demonstrate that the computation cost of the l-BFGS method and the TrN method are comparable, and that the presence of large amplitude multi-scattered waves prevents the l-BFGS method from converging while the TrN method provides significantly more reliable results.

2. Method

2.1. Efficient computation of Hessian-vector products

In the following, we consider that $N_s = 1$ and we drop index s. Formulas for $N_s > 1$ are obtained straightforwardly by summation. We define the function $g_v(p)$ such that

$$g_v(p) = (\nabla f(p), v) = \mathcal{R}\left(J^\dagger R^\dagger (Ru(p) - d), v\right), \tag{5}$$

where $u(p)$ is the solution of (1), $J(p) = \partial_p u(p)$, $\mathcal{R}$ denotes the real part operator, and $\dagger$ the adjoint operator. We have $\nabla g_v(p) = \nabla^2 f(p)v$. The Lagrangian function associated with the functional $g_v(p)$ is

$$L_v\left(p,u,\alpha,\lambda,\mu\right) = \mathcal{R}\left(R^\dagger(Ru-d),\alpha\right) + \mathcal{R}\left(S(p)u-\varphi,\mu\right) + \mathcal{R}\left(S(p)\alpha + \sum_{j=1}^{m} v_j \partial_{p_j} S(p)u, \lambda\right), \tag{6}$$

where the first term is related to the function $g_v(p)$, the second term to the wave equation (1) and the third term to the first derivatives of (1). For $\widetilde{u}$ and $\widetilde{\alpha}$ such that

$$S(p)\widetilde{u} = \varphi, \quad S(p)\widetilde{\alpha} = -\sum_{j=1}^{m} v_j \partial_{p_j} S(p)\widetilde{u}, \tag{7}$$

we have

$$\nabla g_v(p) = \partial_p L_v\left(p,\widetilde{u},\widetilde{\alpha},\lambda,\mu\right) + \partial_u L_v\left(p,\widetilde{u},\widetilde{\alpha},\lambda,\mu\right)\partial_p\widetilde{u}(p) + \partial_\alpha L_v\left(p,\widetilde{u},\widetilde{\alpha},\lambda,\mu\right)\partial_p\widetilde{\alpha}(p). \tag{8}$$

We define $\widetilde{\lambda}$ and $\widetilde{\mu}$ such that

$$\partial_u L_v\left(p,\widetilde{u},\widetilde{\alpha},\widetilde{\lambda},\widetilde{\mu}\right) = 0, \quad \partial_\alpha L_v\left(p,\widetilde{u},\widetilde{\alpha},\widetilde{\lambda},\widetilde{\mu}\right) = 0. \tag{9}$$

We have

$$S(p)^\dagger\widetilde{\mu} = -R^\dagger R\widetilde{\alpha} - \sum_{j=1}^{m} v_j\left(\partial_{p_j}S(p)\right)^\dagger\widetilde{\lambda}, \quad S(p)^\dagger\widetilde{\lambda} = -R^\dagger(R\widetilde{u}-d), \tag{10}$$

and

$$(\nabla^2 f(p)v)_i = \mathcal{R}\left(\left(\left(\partial_{p_i}S(p)\right)\widetilde{u},\widetilde{\mu}\right) + \left(\left(\partial_{p_i}S(p)\right)\widetilde{\alpha},\widetilde{\lambda}\right) + \sum_{j=1}^{m} v_j\left(\left(\partial_{p_j}\partial_{p_i}S(p)\right)\widetilde{u},\widetilde{\lambda}\right)\right), i = 1,\ldots,m. \tag{11}$$

Since computation of ∇f through the adjoint state method already requires the computation of $\widetilde{\lambda}$ and $\widetilde{u}$ [6], the computation of one matrix vector product $\nabla^2 f(p)v$ requires to solve one additional forward problem for $\widetilde{\alpha}$ and one additional adjoint problem for $\widetilde{\mu}$, as reported in [3, 4]. For $N_s > 1$, the overall computation cost is multiplied by N_s.

2.2. CG stopping criterion

An efficient implementation of the truncated Newton method requires the definition of an adaptive stopping criterion for the CG criterion. Newton methods are based on the sequential minimization of local quadratic approximations

$$q_k(\Delta p_k) = f(p_k) + (\nabla f(p_k), \Delta p_k) + (\nabla^2 f(p_k)\Delta p_k, \Delta p_k). \tag{12}$$

The accuracy required to solve the system (4) should reflect the accuracy of these local approximations to prevent from oversolving. This is achieved using the Eisenstat stopping criterion [2]:

$$\|H(p_k)\Delta p_k + \nabla f(p_k)\| \leq \eta_k \|\nabla f(p_k)\|, \tag{13}$$

where the forcing term η_k measures the distance between the misfit gradient and its first order development

$$\eta_k = \frac{\|\nabla f(p_k) - \nabla f(p_{k-1}) - H(p_{k-1})\Delta p_{k-1}\|}{\|\nabla f(p_{k-1})\|}. \tag{14}$$

Moreover, far from the solution, the Hessian operator should be indefinite. An additional stopping criterion is thus introduced: as soon as a negative curvature direction is computed during the resolution of the system (4), the CG iterations are stopped.

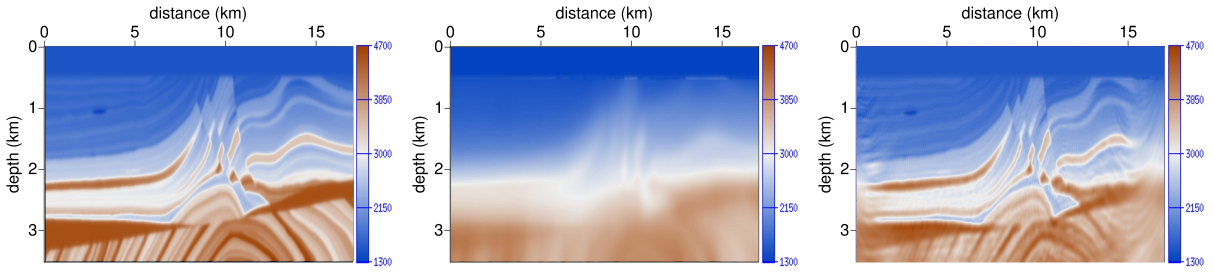

Figure 1. l-BFGS inversion results. Exact model (left), initial model (middle), estimated model (right).

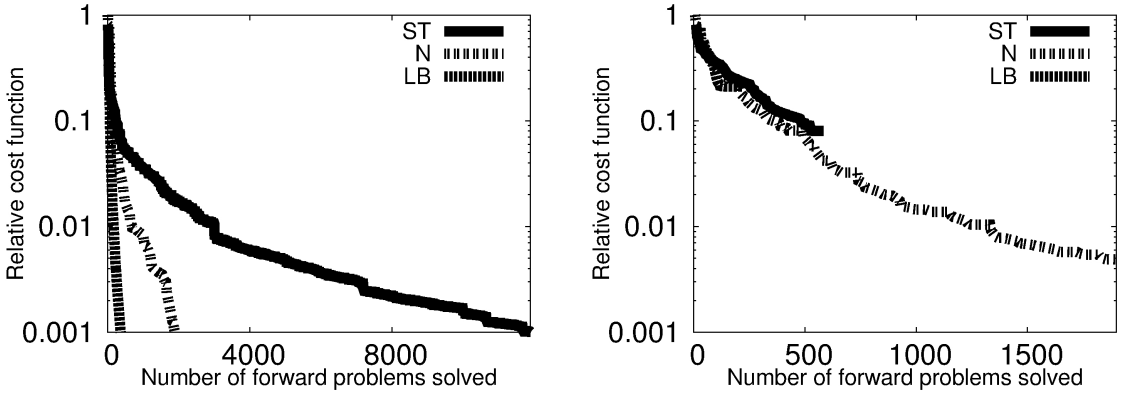

Figure 2. Convergence curves for the Marmousi II test case (left), for the near-surface imaging test case (right). N: truncated Newton method, LB: l-BFGS, ST: steepest-descent

3. Numerical results

3.1. The Marmousi II test case

The numerical results presented in this study are obtained in the 2D acoustic frequency domain FWI context. An estimation of the pressure wave velocity is computed. We first consider the Marmousi II benchmark test case. The Marmousi II model is 16 km wide and 3.5 km deep. The discretization step is set to 25 m. We use 144 sources and 660 receivers located near the surface to generate synthetic data. Four datasets corresponding to the frequencies $3, 5, 8, 12$ Hz are simultaneously inverted. The efficiencies of the steepest-descent, the l-BFGS algorithm and the TrN algorithm are compared. The three methods are implemented with the same linesearch globalization method. The iterations are stopped when the relative cost function $f(p)/f(p_0)$ reaches 10^{-3}. The estimated models obtained using the three different methods are very similar. The one obtained with the l-BFGS method is presented in figure 1.

The convergence curves (fig 2) are plotted as a function of the total number of the forward-problem resolutions. As expected, the steepest-descent algorithm converges very slowly, while the l-BFGS and the TrN method provide faster convergence. Note that the l-BFGS method is the fastest.

3.2. A near-surface imaging test case

We consider next a near-surface seismic imaging problem. Detecting and correctly imaging two concrete structures buried in the subsurface at few meters depth (fig. 4) is a challenge. The

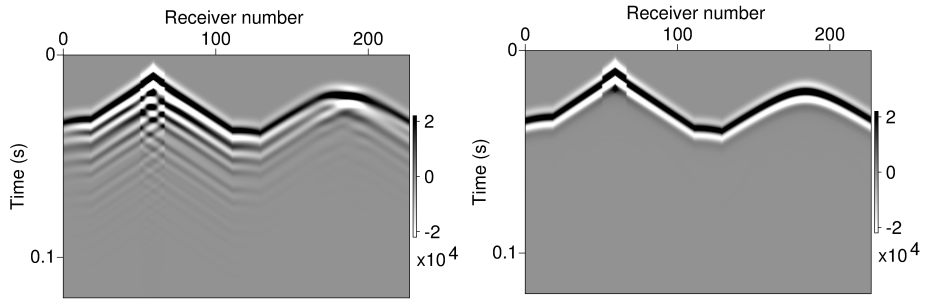

Figure 3. Dataset associated with the exact model (left), data associated with the initial homogeneous domain (right)

depth of investigation is limited to 3 m, the width of the exact model is 15 m. The discretization step is set to 0.15 m. We use a full acquisition system: four lines of sources/receivers are located on each side of the domain. We compute 9 datasets, from 100 Hz to 300 Hz each 25 Hz.

The very high velocity contrast between the background (300 m.s^{-1}) and the concrete foundations (4000 m.s^{-1}) generates energetic reflections. In addition, the close distance between the two structures is responsible for important multiple scattering. This is illustrated in figure 3 where two datasets in the time domain are presented[1], computed using the exact model and the homogeneous background model. The background model correctly predicts only the first-arrival waves. The signal below the first arrival on the left figure, corresponding to the multi-scattered waves, is not predicted. Starting from this background model, we invert simultaneously the 9 datasets, using the l-BFGS algorithm, and the TrN method. The convergence curves and the corresponding results are presented in figures 2 and 4 (the steepest-descent result is not presented since it is very similar to the result provided by the l-BFGS method). While the

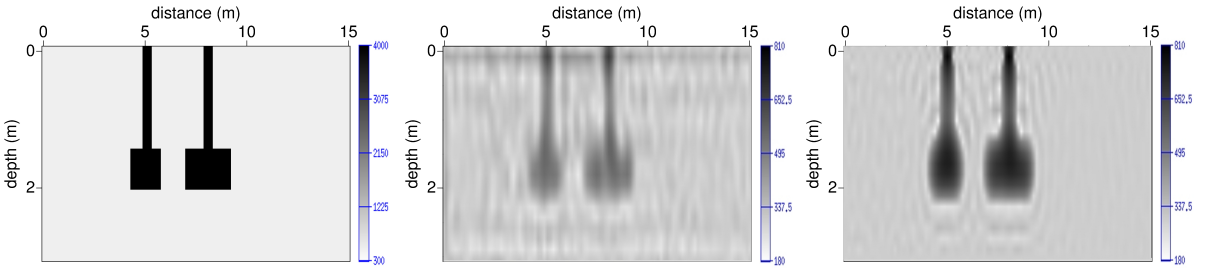

Figure 4. Pressure wave velocities. Exact model (left), l-BFGS estimation (middle), TrN estimation (right).

steepest-descent and the l-BFGS method stop after few iterations, trapped in a local minimum, the TrN method further reduces the misfit function. Why in this case the TrN method performs better than the l-BFGS method? Consider the Hessian operator expression:

$$\nabla^2 f(p) = \sum_{s=1}^{N_s} J_s^\dagger(p) R_s^\dagger R_s J_s(p) + \sum_{s=1}^{N_s} R_s^\dagger \left(R_s u_s(p) - d_s \right) \partial_{pp} u_s(p). \tag{15}$$

[1] These two datasets are generated using a Ricker source located at the surface between the two concrete structures at $x = 6.5$ m, centered on the frequency 200 Hz.

The first term of the Hessian operator is positive definite by construction, while the second one is indefinite, related to the double-scattered wavefield $\partial_{pp} u_s(p)$ and the residuals $R_s u_s(p) - d_s$. When single scattered waves dominate the data, gradient-based methods converge quickly from the first iterations. Therefore, the residuals decrease, and the Hessian operator tends to become positive definite. The l-BFGS methods builds a symmetric definite positive approximation of the inverse Hessian along the minimization process, and therefore improves the convergence speed of standard gradient methods. Since the Marmousi II model does not present high velocity contrasts, the amplitude of the multiscattered wavefield for this test case is weak, and the l-BFGS method is efficient.

When multiscattered waves cannot be neglected, as in our second test case, gradient-based methods face difficulties to fit the data, since the gradient direction only accounts for single scattered waves. Therefore, for the first iterations, the l-BFGS estimation only slowly decreases the residuals as the inverse Hessian approximation is close from the identity. As a consequence, the true Hessian operator stays indefinite while the l-BFGFS method build a definite positive approximation of its inverse. This explains why the l-BFGS method fails in this case. Conversely, the TrN method better accounts for the Hessian operator and is able to converge.

4. Conclusion and perspectives

Second-order adjoint formulas and the Eisenstat stopping criterion [2] yield an efficient implementation of the TrN method. Within this framework, the comparison of the TrN method with the l-BFGS method demonstrates the interest of using the TrN method when the seismic data is dominated by multi-scattered waves. In this case, the second-order information embedded in the Hessian operator, which is better accounted for by the TrN method, is crucial. The authors now look forward for the application of the TrN method to more general context, such as multi-parameter inversion and elastic FWI. The possibility of computing the inverse Hessian locally is also investigated in order to compute the posterior covariance matrix and provides an uncertainty estimation of the parameter reconstruction. Finally, the use of matrix-free preconditionners in order to speed-up the convergence of the CG algorithm will also be investigated.

Acknowledgments

This research is funded by the SEISCOPE consortium sponsored by BP, CGG-VERITAS, ENI, EXXON MOBIL, PETROBRAS, SAUDI ARAMCO, SHELL, STATOIL and TOTAL. The linear systems were solved with the MUMPS package. This work was performed by accessing to the high-performance computing facilities of CIMENT (Université de Grenoble, France) and to the HPC resources of GENCI-CINES under Grant 2011-046091.

References

[1] R. Byrd, P. Lu, and J. Nocedal, *A limited memory algorithm for bound constrained optimization*, SIAM Journal on Scientific and Statistical Computing, 16 (1995), pp. 1190–1208.

[2] S. C. Eisenstat and H. F. Walker, *Choosing the forcing terms in an inexact newton method*, SIAM Journal on Scientific Computing, 17 (1994), pp. 16–32.

[3] I. Epanomeritakis, V. Akçelik, O. Ghattas, and J. Bielak, *A Newton-CG method for large-scale three-dimensional elastic full waveform seismic inversion*, Inverse Problems, 24 (2008), pp. 1–26.

[4] A. Fichtner and J. Trampert, *Hessian kernels of seismic data functionals based upon adjoint techniques*, Geophysical Journal International, 185 (2011), pp. 775–798.

[5] S. G. Nash, *A survey of truncated Newton methods*, Journal of Computational and Applied Mathematics, 124 (2000), pp. 45–59.

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

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