FWI with DFP optimization using an approximate low-resolution Hessian

Scott Keating and Kris Innanen

ABSTRACT

Quasi-Newton optimization methods have been shown to provide considerable improvements in the convergence rate of full waveform inversion (FWI). These methods use approximate Hessian matrices to generate an update direction. The DFP method uses an approximate Hessian which predicts observed changes in the gradient, while remaining nearest in a least-squares sense to the previous approximate Hessian. By using a low-resolution approximation to the Hessian to initialize the method, we attempt to increase the rate of convergence. This work is ongoing, currently there are significant challenges in using the inverse of the Hessian on the low-resolution scale to provide a meaningful inverse on the high-resolution scale.

INTRODUCTION

Full waveform inversion (FWI) is a computationally demanding approach for recovering a model of the sub-surface (Lailly, 1983; Tarantola, 1984). The driver of the FWI problem is a numerical optimization procedure, wherein measured data and synthetic data generated by the estimated model are made to match as closely as possible. A vital step in most numerical optimization procedures is the calculation of the gradient, and this calculation typically dominates the cost of FWI (Virieux and Operto, 2009). FWI can easily be formulated using a steepest descent approach to this optimization, where the model is simply updated by the gradient multiplied by a scalar at each iteration. It has been shown, however, that the inversion can be made more efficient by using more sophisticated numerical optimization techniques, such as conjugate gradients (e.g. Kamei and Pratt, 2013), l-BFGS (e.g. Hak and Mulder, 2011) and truncated Newton optimization (e.g. Métivier et al., 2015). Quasi-Newton methods like l-BFGS modify the gradient direction with an approximation of the inverse Hessian before updating the model. This introduces an additional cost at each iteration, but typically leads to faster convergence, and thus fewer required iterations.

A major factor in the efficiency of the l-BFGS method is the starting estimate of the inverse Hessian. Estimates which better capture the important features of the inverse Hessian lead to faster convergence in the numerical optimization procedure. Unfortunately, it can be very difficult to obtain a reasonable estimate of the inverse Hessian. A closely related method, DFP optimization (named for Davidon, Fletcher and Powell), employs an estimate of the Hessian matrix, which may be simpler to approximate. We investigate here the idea of using a low resolution estimate of the Hessian matrix in initializing DFP in FWI.

THEORY

In exact Newton optimization, the descent direction at each iteration is calculated by solving

$$Hp = -g, \tag{1}$$

where H is the Hessian, g is the gradient, and p is the descent direction. Broyden class optimization methods create a descent direction by applying an approximation to the inverse Hessian Q to the gradient, so equation 1 is replaced with

$$p = -Qg. \tag{2}$$

Information about the Hessian is introduced to Q by requiring that it successfully predicts the last observed change in the gradient (Nocedal and Wright, 2006). For a linear problem, this change is given by

$$\Delta g = H\Delta m, \tag{3}$$

where Δm is the change in the model, and Δg is the change in the gradient. The matrix Q is then required to satisfy

$$Q^{-1}\Delta m = \Delta g. \tag{4}$$

This condition alone does not sufficiently constrain Q, so additional conditions are required. In the BFGS method, the Q chosen is that which satisfies equation 4 while minimizing the Frobenius norm of the difference from the Q used at the previous iteration (Nocedal and Wright, 2006). This approach can be seriously affected by the initial Q set in the inversion. If the input Q is very close to the true inverse Hessian in a least squares sense, the BFGS update will both predict the changes in the gradient and be close in a least squares sense to the true Hessian. Ideally, this leads to an update which is close to the exact Newton update. Unfortunately, the inverse of the Hessian matrix can be very difficult to predict.

The DFP method is another Broyden class method, and imposes the same requirement of predicting the last observed change in the gradient. The DFP method replaces the requirement that Q be nearest to the previous Q with the requirement that Q^{-1} be nearest to the previous Q^{-1} (Nocedal and Wright, 2006). This effectively means that the change to the approximate Hessian is minimized, rather than its inverse, and that the initial Q should be the inverse of an approximate Hessian. The DFP method can then be expected to perform well if provided with an approximate Hessian close to the true one, while the BFGS method should perform well with an approximate inverse Hessian close to the true one. This distinction may be notable: matrices which are close may have inverses that are far apart.

The DFP approximation to the inverse Hessian is given by

$$Q_{k+1} = Q_k - \frac{Q_k \Delta g_k \Delta g_k^T Q_k}{\Delta g_k^T Q_k} + \frac{\Delta m_k \Delta m_k^T}{\Delta g_k^T \Delta m_k} \tag{5}$$

In Quasi-Newton methods, the approximate inverse Hessian matrix is typically not computed directly or stored, due to the large size of the matrix. Direct calculation is not necessary because it is the product of this matrix with the gradient that is required in solving equation 2. An iterative procedure for the calculation of Qg can be created based on 5 (Nocedal and Wright, 2006). This requires only the storage of the previous k differences Δg_k and Δm_k, and the initial estimate Q_0.

An approximate Hessian matrix

Creating approximations close to the Hessian matrix may be considerably easier than generating approximations close to the inverse Hessian. This is because while the entire Hessian may be costly to compute and store, it is relatively simple to generate an expression for an element of the Hessian. By making approximations to these expressions, we can generate an approximation to the Hessian. By contrast, it is very difficult to generate an expression for an element of the inverse Hessian, given the need for a matrix inversion. As previously discussed, the inverse of a close approximation to the Hessian does not necessarily generate a close approximation to the inverse Hessian.

For example, in the an-acoustic FWI problem, the elements of the Hessian matrix relating to derivatives with respect to, for instance, velocity, are given by

$$H_{cc}(r,r') = \sum_{r_g,r_s} \int d\omega \omega^4 [1+\beta s_{c_o}(r)]^2 \mathcal{G}(r_g,r,r',r_s), \tag{6}$$

where

$$\mathcal{G}(r_g,r,r',r_s) = G_0^*(r_g,r')G_0^*(r',r_s)G_0(r_g,r)G_0(r,r_s), \tag{7}$$

$$\beta = i - \frac{2}{\pi}\log\left(\frac{\omega}{\omega_0}\right), \tag{8}$$

$s_{c_o} = \frac{1}{c^2}$, ω_0 is a reference frequency, and G_0 are Green's functions in the current medium.

The matrix $\sum_{r_g,r_s} \mathcal{G}(r_g,r,r',r_s)$ is very expensive to calculate for large models, and is typically too large to store. One approach to approximating the Hessian matrix is to replace G_0 with $P^T G_0$, where P^T is a matrix projecting from the full dimension of the inversion problem onto a smaller grid. $[1+\beta s_{c_o}(r)]$ is likewise replaced with $P^T[1+\beta s_{c_o}(r)]$. The result is an approximation to the matrix $P^T H P$, where P and P^T are applied to the Green's functions before they are multiplied, rather than after. This effectively changes the order of a multiplication and a summation, valid only when all elements of a sum are equal. Thus, this approximation effectively assumes that the wavefield is slowly varying on the scale of the coarser grid. An approximate version of the Hessian on the original grid is then given by PBP^T, where B is the approximation to $P^T H P$.

The advantage of this approach is that for a model of dimension N, and a coarse grid of dimension M, the cost of calculating the approximate Hessian reduces to $\mathcal{O}(M^3)$ from $\mathcal{O}(N^3)$ operations, and the memory required to store the approximate Hessian is $\mathcal{O}(M^2)$ instead of $\mathcal{O}(N^2)$. If this approximate Hessian can then be shown to be close to the true Hessian, it provides an appealing starting matrix for a DFP type optimization.

An example segment of an exact Gauss-Newton Hessian matrix and the corresponding approximation for an example QFWI problem are shown in figure 1. The Frobenius norm of the difference between the exact and approximate Hessian is about 40.9% the norm of the exact Hessian matrix, so these matrices are fairly close to one another for the purposes of the DFP optimization approach. The inverses of these matrices after adding a small stabilization term λI are shown in figure 2. The Frobenius norm of the difference of these

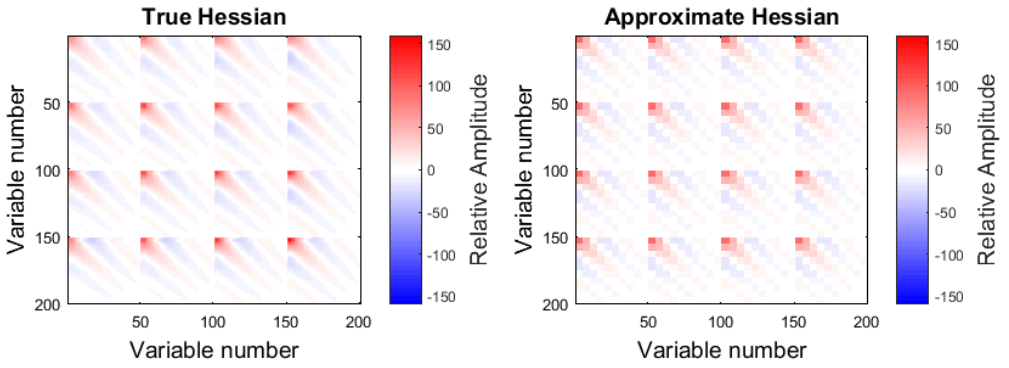

FIG. 1.
Exact (left) and approximate (right) Hessian for a subset of variables in a QFWI problem. The difference between the exact and approximate Hessian has a relatively small Frobenius norm.

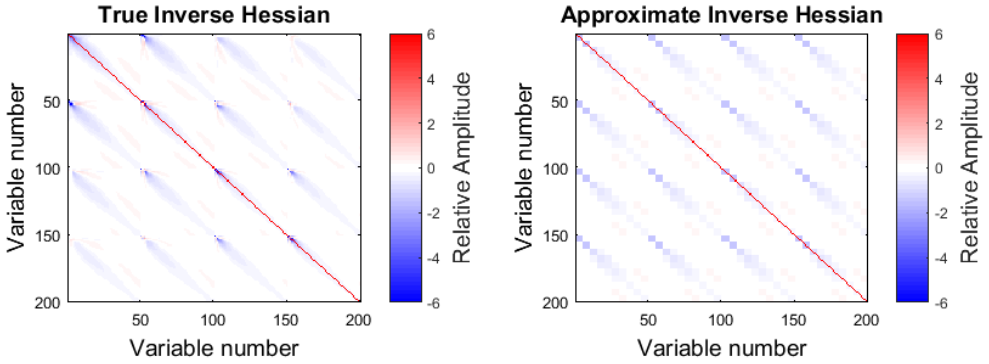

FIG. 2.
Inverse of the exact (left) and approximate (right) Hessian after stabilization for a subset of variables in a QFWI problem. The difference between the exact and approximate inverse Hessian has a large Frobenius norm.

matrices is about 98.4% the norm of the inverse of the exact Hessian matrix, making this approximation a poor starting point for the BFGS optimization approach.

An inverse to the approximate Hessian

The DFP method requires the inverse of the approximate Hessian to initialize the inversion. Inverting a matrix of N^2 elements is a process requiring $\mathcal{O}(N^3)$ operations, which is generally infeasible. Much less expensive is obtaining the inverse of the matrix B, which contains only M^2 elements. Challenges arise in the proposed method when attempting to use this inverse to provide a meaningful value of Q. Even with a calculated B^{-1} such that $BB^{-1} = I_M$, it is difficult to calculate the inverse of the approximate Hessian, PBP^T. While it is easy to generate an estimate Q_M such that

$$PBP^TQ_MP^T = PI_MP^T, \tag{9}$$

the inverse we require is Q, which satisfies

$$PBP^TQ = I_N. \tag{10}$$

Several problems arise in determining such a Q. First, the smaller dimensionality of B means that PBP^T is not an invertible matrix. This obstacle is not insurmountable, a sparse stabilizing term can be added to PBP^T to make this term invertible. If the stabilizing term

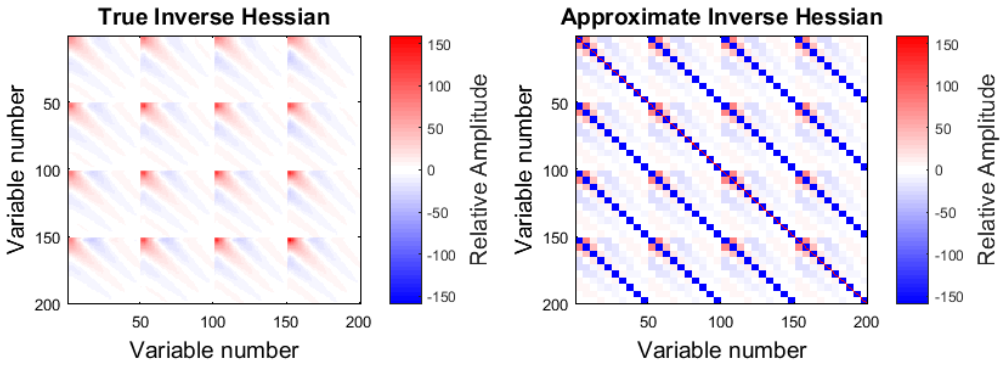

FIG. 3.
Exact Hessian (left) and inverse of Q as defined in equation 11 (right) for a subset of variables in a QFWI problem. Despite significant similarities, the difference between these matrices is very large.

is sufficiently small, this should also preserve the L_2 closeness to the true Hessian. A more daunting challenge is the necessity for Q to be N dimensional to satisfy equation 10. This introduces significant memory requirements unless Q can be efficiently generated from a smaller dimensional term, as the Hessian approximation PBP^T is. A similar approach to that which generates the Hessian approximation can be followed, using

$$Q = PQ_MP^T + \lambda I_N, \tag{11}$$

where λ is a small scalar. Unfortunately, the inverse of the Q provided in equation 11 is not a good approximation of the Hessian, as illustrated in figure 3.

CONCLUSIONS

The approach discussed in this report produces an approximation to the Hessian matrix which is very close to the true Hessian. This approximation can be generated with small memory and computational cost requirements, and provides an appealing initial estimate for a DFP optimization procedure. Unfortunately, the inverse of the approximate Hessian is challenging to compute using the same memory and computational cost requirements. This challenge is not specific to this particular approximation to the Hessian, approaches to generating a low memory approximation to the Hessian still require that the inverse be evaluated, which may require much more memory and computation. This will likely present a problem for any non-diagonal approximate Hessian.

ACKNOWLEDGMENTS

The authors thank the sponsors of CREWES for continued support. This work was funded by CREWES industrial sponsors and NSERC (Natural Science and Engineering Research Council of Canada) through the grant CRDPJ 461179-13. Scott Keating was also supported by the Earl D. and Reba C. Griffin Memorial Scholarship.

REFERENCES

Hak, B., and Mulder, W., 2011, Seismic attenuation imaging with causality: Geophysical Journal International, **184**, 439–451.

Kamei, R., and Pratt, R., 2013, Inversion strategies for visco-acoustic waveform inversion: Geophysical Journal International, **194**, 859–884.

Lailly, P., 1983, The seismic inverse problem as a sequence of before stack migrations: Conference on Inverse Scattering, Theory and Application, Society for Industrial and Applied Mathematics, Expanded Abstracts, 206–220.

Métivier, L., Brossier, R., Operto, S., and Virieux, J., 2015, Acoustic multi-parameter FWI for the reconstruction of P-wave velocity, density and attenuation: preconditioned truncated Newton approach: SEG Expanded Abstracts, 1198–1203.

Nocedal, J., and Wright, P. S., 2006, Numerical Optimization: Springer, 2nd edn.

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

Virieux, J., and Operto, S., 2009, An overview of full-waveform inversion in exploration geophysics: Geophysics, **74**, No. 6, WCC1.