Efficient stochastic Hessian estimation for full waveform inversion

Lucas A. Willemsen, Alison E. Malcolm and Russell J. Hewett, Massachusetts Institute of Technology

SUMMARY

In this abstract we present a method that allows arbitrary elements of the approximate Hessian to be estimated simultaneously. Preliminary theoretical and numerical investigations suggest that the number of forward models required for this procedure does not increase with the number of shots. As the number of shots increases this means that the cost of estimating these approximate Hessian entries becomes negligible relative to the cost of calculating the gradient. The most obvious application would be to estimate the diagonal of the approximate hessian. This can then be used as a very inexpensive preconditioner for optimization procedures, such as the truncated Newton method.

INTRODUCTION

The least squares (L_2) misfit functional

$$\chi(m) = \frac{1}{2}||\mathbf{S}u(m) - d||_2^2 \tag{1}$$

has traditionally received a lot of attention in the full waveform inversion (FWI) community. Newton's method

$$\mathbf{H}\delta m = -\mathbf{g} \tag{2}$$

determines the model parameter update δm in terms of the Hessian $\mathbf{H}$ and gradient $\mathbf{g}$ of the misfit functional. The importance of using the Hessian in the inversion process is illustrated by Pratt et al. (1998). However, the Hessian is rarely explicitly computed due to its excessive cost. Inexpensive alternatives have been investigated in the past, for example the 'pseudo-Hessian' presented by Shin et al. (2001). The pseudo-Hessian shares many characteristics with the approximate Hessian and is less costly to compute. Other ways of reducing the cost of Newton's method have also been investigated. The L-BFGS method (Byrd et al., 1995) uses the gradient from the multiple nonlinear iterations to approximate the effect of the inverse Hessian. The inverse Hessian itself is never computed. Another alternative is the truncated Newton method (Métivier et al., 2012). The first and second adjoint wavefields are calculated (Fichtner and Trampert, 2011) which are then used to calculate the action of the Hessian on an arbitrary vector. Krylov subspace methods are then employed for a finite number of iterations to approximate the solution to (2). These methods only require the action of the Hessian on a vector and not explicit computation of the entire Hessian. The authors demonstrate that truncated Newton sometimes outperforms L-BFGS in its ability to recover complicated structures, justifying its higher computational cost.

In this paper we introduce a method for estimating arbitrary elements of the approximate Hessian simultaneously, at low cost. Estimating the diagonal of the approximate Hessian and using that as a cheap preconditioner for (2) may be beneficial.

REVIEW OF THE FREQUENCY DOMAIN L2 FWI-HESSIAN

In continuous mathematics, the constant density acoustic frequency-domain forward model is written as

$$\mathbf{L}(m,x)u(x,\omega) = (-m\omega^2 - \nabla^2)u(x,\omega) = f(x,\omega), \tag{3}$$

where $\mathbf{L}(m,x)$ denotes the Helmholtz operator with model m and Laplacian ∇^2 operating on x, and $f(x,\omega)$ and $u(x,\omega)$ are respectively the source term and the acoustic wavefield at frequency ω and x is the spatial coordinate. The least-squares misfit functional $\chi(m)$ is defined as

$$\begin{aligned}\chi(m) &= \frac{1}{2}||\mathbf{S}u(m) - d||_2^2 = \frac{1}{2}\left\langle \mathbf{S}u(m) - d, \mathbf{S}u(m) - d\right\rangle_d \\ &= \frac{1}{2}\Sigma_s\Sigma_r\Sigma_\omega\left[\left(\mathbf{S}_r u_s(m,x,\omega) - d_{s,r}(\omega)\right)\right. \\ &\left.\times\overline{\left(\mathbf{S}_r u_s(m,x,\omega) - d_{s,r}(\omega)\right)}\right],\end{aligned} \tag{4}$$

where $< \ . \ , \ . >_d$ is the inner product in data space and the overbar is complex conjugation. The operator $\mathbf{S}$ samples the wavefield at the receiver locations; $\mathbf{S}_r$ samples specifically at receiver location r, and the wavefield in model m due to shot s at frequency ω is denoted by $u_s(m,x,\omega)$. Taking two Gâteaux derivatives of the misfit functional (4) gives the Hessian operating on model perturbations $m_1(y)$ and $m_2(y')$ (Virieux and Operto, 2009)

$$\begin{aligned}&\left.\frac{\delta^2\chi}{\delta\mathbf{m}(y)\delta\mathbf{m}(y')}\right|_{m_i}(m_1(y), m_2(y')) = \\ &\mathbf{H}_{\text{appr}}(m_1(y), m_2(y')) + \mathbf{H}_{\text{r}}(m_1(y), m_2(y')),\end{aligned} \tag{5}$$

where

$$\mathbf{H}_{\text{appr}}(m_1(y), m_2(y')) = Re\left\{\left\langle \mathbf{SF}\left[m_1(y)\right], \mathbf{SF}\left[m_2(y')\right]\right\rangle_d\right\}, \tag{6}$$

and $\mathbf{F}$ is the Jacobian operator $\frac{\delta\mathbf{u}_\mathbf{s}(m)}{\delta\mathbf{m}(y)}|_{m_i}$ evaluated at the current model m_i. The combination of operator $\mathbf{F}$ and its argument ($\mathbf{F}\left[m_1(y)\right]$ for instance) corresponds to an inner product in space. The term $\mathbf{F}\left[m_1(y)\right]$ can be interpreted physically as first order scattering of the wavefield in current model m_i due to a perturbation $m_1(y)$. The term of (5) involving $\mathbf{H}_\text{r}$ is often discarded in practice and the approximate Hessian $\mathbf{H}_{\text{appr}}$ is used in (3) instead of the full Hessian. Using adjoint operators, represented by *, the approximate Hessian applied to a single perturbation can be rewritten as

$$\mathbf{H}_{\text{appr}}(m_1(y), (\cdot)) = Re\left\{\mathbf{F}^*\mathbf{S}^*\mathbf{SF}\left[m_1(y)\right]\right\}\left[(\cdot)\right], \tag{7}$$

where $\left[(\cdot)\right]$ is the operator slot where the second perturbation $m_2(y')$ is inserted resulting in an integration over space. The operator $\mathbf{F}^*$ implicitly includes a summation over the shots. In discretized form the operators are replaced by their discretized versions (e.g., $m_1(y)$ becomes a vector). Column k of the Hessian can conveniently be computed by making $m_1(y)$ a unit

perturbation at node location y_k. Since the adjoint operator $\mathbf{F}^*$ includes a summation over the shots, calculating a column of $\mathbf{H}_{\text{appr}}$ requires two forward propagations, for $\mathbf{F}[m_1(y)]$, and one backward propagation, per shot. The total computation per column of $\mathbf{H}_{\text{appr}}$ is $2n_s$ forward propagations and n_s backward propagations, where n_s is the number of sources.

Using our knowledge of the wave operator we can derive an alternative expression for $\mathbf{H}_{\text{appr}}$. The derivative with respect to m of equation (3) is:

$$\mathbf{L}(m_i,x)\frac{\delta \mathbf{u_s}(x,\omega)}{\delta \mathbf{m}(y)} = -\frac{\delta \mathbf{L}(m_i,x)}{\delta \mathbf{m}(y)}u_s(x,\omega). \tag{8}$$

Equation 8 shows that the first order scattered field $\frac{\delta \mathbf{u_s}(x,\omega)}{\delta \mathbf{m}(y)}$ satisfies the wave equation. Using the notion of Green's function (8) can be rewritten as

$$\frac{\delta \mathbf{u_s}(x,\omega)}{\delta \mathbf{m}(y)} = \omega^2 G(x,y,\omega)u_s(y,\omega). \tag{9}$$

Equation 9 is the first order forward scattered field due to shot s and a delta perturbation in m at location y. The waves travel with speed dictated by the unperturbed model m_i, as shown in (8). The approximate Hessian (6) can therefore be written as

$$\mathbf{H}_{\text{appr}}[y,y'] = Re\Big\{\Sigma_s\Sigma_r\Sigma_\omega \mathbf{S}_r\left(\omega^2 G(x,y,\omega)u_s(y,\omega)\right) \times \overline{\mathbf{S}_r\left(\omega^2 G(x,y',\omega)u_s(y',\omega)\right)}\Big\}, \tag{10}$$

where $\mathbf{H}_{\text{appr}}[y,y']$ represents the Hessian entry corresponding to perturbations at nodal locations y and y'. This expression can be simplified when two assumptions are made:

1. the sampling operator $\mathbf{S}_r$ perfectly samples the wavefield at location x_r,
2. the source is a spatial delta function at x_s and all sources have spectrum $W(\omega)$.

Under these assumptions (10) can be rewritten as

$$\mathbf{H}_{\text{appr}}[y,y'] = Re\Big\{\sum_\omega\sum_s\sum_r \omega^4|W(\omega)|^2 \times G(y,x_s,\omega)G(x_r,y,\omega)\overline{G(y',x_s,\omega)G(x_r,y',\omega)}\Big\}. \tag{11}$$

COMPLEXITY REDUCTION WITH WHITE-NOISE VIRTUAL-SOURCE SETS

In this section we rewrite (11) in a form that may be more efficient for computation; there is no additional field data required for the proposed method of computation. We replace the sums over the sources and receivers by using concepts from passive seismic interferometry. When virtual white-noise sources are placed at each of the true receiver locations a virtual receiver at node location y measures the following response, at a specific frequency

$$\tilde{P}_k(y) = \sum_{i=1}^{n_r} G(y,x_i,\omega)N_k(x_i,\omega), \tag{12}$$

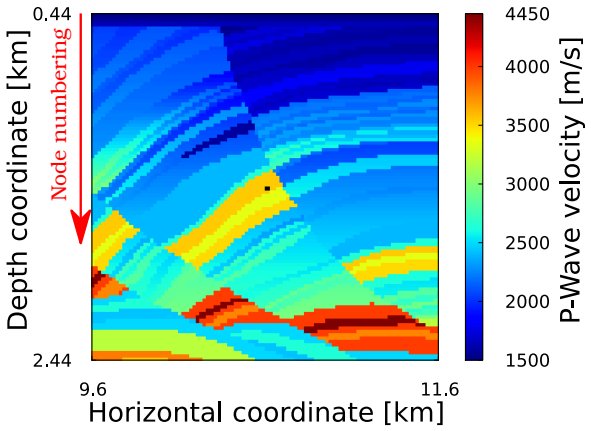

Figure 1: Acoustic P-wave velocity for 100x100 square section of Marmousi2 (Martin et al., 2006) with node spacing 20m. Sources and receivers are placed on all the 100 nodes on top. Part of Hessian column corresponding to black pixel in center is displayed in Figure 3

where $N_k(x_i,\omega)$ is the spectrum of the noise emitted by the virtual source at frequency ω and at the receiver node location x_i, subscript k is a specific realization, and n_r is the number of receivers. The ensemble average of the cross-correlation between the measured pressures $\tilde{P}(y)$ and $\tilde{P}(y')$ over n_k realizations is thus

$$\left\langle \tilde{P}(y)\overline{\tilde{P}(y')}\right\rangle_r = \sum_{i=1}^{n_r}\sum_{j=1}^{n_r} G(y,x_i,\omega)\overline{G(y',x_j,\omega)} \times \frac{1}{n_k}\sum_{k=1}^{n_k} N_k(x_i,\omega)\overline{N_k(x_j,\omega)}. \tag{13}$$

In the limit, as $n_k \to \infty$, the second term of (13) simplifies,

$$\frac{1}{n_k}\sum_{k=1}^{n_k} N_k(x_i,\omega)\overline{N_k(x_j,\omega)} = \mu_N(\omega)\delta_{ij}, \tag{14}$$

where δ_{ij} is the Kronecker delta and $\mu_N(\omega)$ is the expected value of the autocorrelation $N_k(x_i,\omega)\overline{N_k(x_i,\omega)}$ at frequency ω. When n_k is sufficiently large, the ensemble average in (13) reduces to

$$\left\langle \tilde{P}(y)\overline{\tilde{P}(y')}\right\rangle_r \approx \mu_N\sum_{i=1}^{n_r} G(y,x_i,\omega)\overline{G(y',x_i,\omega)}. \tag{15}$$

If instead virtual noise sources are placed at all of the true source locations, a similar expression is obtained

$$\left\langle \tilde{P}(y)\overline{\tilde{P}(y')}\right\rangle_s \approx \mu_N\sum_{i=1}^{n_s} G(y,x_i,\omega)\overline{G(y',x_i,\omega)}. \tag{16}$$

Inserting both (15) and (16) into equation (11) gives

$$\mathbf{H}_{\text{appr}}[y,y'] \approx \mathbf{H}_{\text{s}}[y,y'] = Re\left[\sum_\omega \frac{\omega^4}{\mu_N^2}|W(\omega)|^2\left\langle \tilde{P}(y)\overline{\tilde{P}(y')}\right\rangle_s\left\langle \tilde{P}(y)\overline{\tilde{P}(y')}\right\rangle_r\right], \tag{17}$$

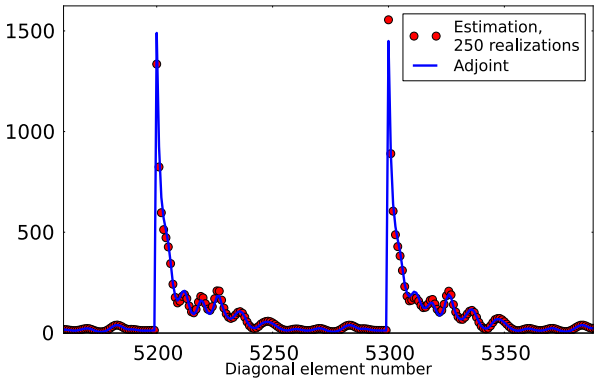

Figure 2: A section of the diagonal extracted from the approximate Hessian calculated with the adjoint method (7) is plotted against the estimate using (17) with the ensemble averages done using 250 realizations.

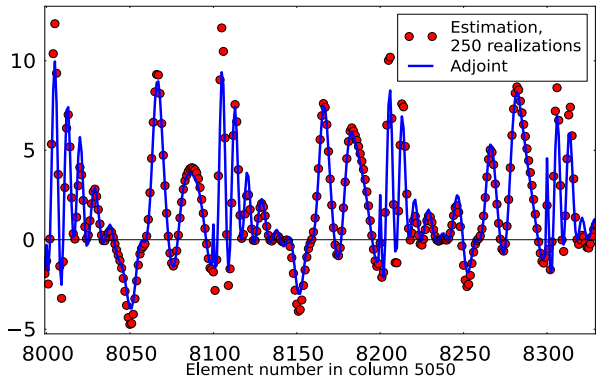

Figure 3: A section of the column for node 5050, in the center of figure 1. The comparison is for $n_k = 250$ in both ensemble averages in equation (17).

where the subscripts s and r indicate whether the virtual noise sources are placed at the source or receiver locations and $\mathbf{H}_s$ is the stochastic approximate Hessian. Equation 17 allows arbitrary elements of the Hessian to be evaluated by simply firing a sufficient number of virtual white noise sources at the frequencies of interest at the true source and true receiver locations. All the $\mathbf{H}_{\text{appr}}$ entries corresponding to specific combinations of y and y' can be approximated simultaneously.

For a finite length white noise sequence, discretely sampled in time, it can be proven that its spectrum at each frequency ω has a normally distributed real and imaginary part with zero mean. So a white noise realization can be generated directly in the frequency domain by letting

$$Re\Big\{N_k(x,\omega)\Big\} \sim \mathcal{N}(0,1), \;\; Im\Big\{N_k(x,\omega)\Big\} \sim \mathcal{N}(0,1), \tag{18}$$

where $\mathcal{N}(\mu,\sigma^2)$ represents a normal distribution with mean μ and variance σ^2. With the probability distributions in (18), $\mu_N(\omega) = 2$.

NUMERICAL EXAMPLE

The derivations presented in the previous section are valid for both 2D and 3D geometries. The synthetic example presented in this section is based on a 100x100 section of the Marmousi2 P-wave velocity model (Martin et al., 2006) surrounded by a 300 m wide PML, see Figure 1. A subsection of the model is investigated so $\mathbf{H}_{\text{appr}}$ can be calculated exactly columnwise in a computationally tractable manner using (7). The stochastic estimation $\mathbf{H}_s$ as given in (17) is compared to $\mathbf{H}_{\text{appr}}$. All simulations are performed using the authors' Python Seismic Inversion Toolbox (PySIT). Sources and receivers are placed on all of the 100 nodes at the top of the domain. It should be mentioned that such dense coverage in source and receivers is not required. The comparison is done at a frequency of 6Hz and both the ensemble averages $<>_s$ and $<>_r$ in (17) use $n_k =$ 250 realizations. Figure 2 compares part of the diagonal of $\mathbf{H}_s$ with $\mathbf{H}_{\text{appr}}$. The node numbering in the grid is top to bottom and then left to right. This gives characteristic peaks with a periodicity of 100 entries along the diagonal. The entries of $\mathbf{H}_{\text{appr}}$ corresponding to nodes close to the surface have a larger amplitude because they experience less geometrical spreading. Figure 2 shows that the diagonal is well approximated using 250 realizations.

Similar results are obtained when off-diagonal elements are estimated. Figure 3 shows column 5050 in the Hessian, corresponding to the node indicated by the black square in Figure 1. Equation 6 shows that each $\mathbf{H}_{\text{appr}}$ entry is a cross-correlation between the first order scattered fields from both perturbations m_1 and m_2 with sums over the sources and receivers. When the distance between the perturbations is small relative to the wavelength this will result in constructive interference of the first order scattered fields at the receiver locations. Because of this, the approximate Hessian contains off-diagonal regions with significant positive values. Figure 3 only shows part of column 5050 corresponding to elements approximately 600 meters to the right of the black pixel in Figure 1. The first order scattered fields will on average only have small correlations. There is also little apparent structure as can be seen in Figure 3. The approximate Hessian entries show large fluctuations although some periodicity of 100 nodes is visible. We see that the approximation from (17) is still quite accurate.

CONVERGENCE AND COMPUTATIONAL COMPLEXITY

The ensemble averages in (17) approach their expected value in the limit when n_k goes to infinity. The approximation is then an equality. The convergence behavior is tested by defining an average relative mismatch between the estimated values (17) and the values calculated through the adjoint method

$$\frac{1}{N_e}\sum_{i=1}^{N_e}\frac{|\text{est}(i)-\text{adj}(i)|}{|\text{adj}(i)|}, \tag{19}$$

where N_e represents the number of Hessian entries compared. The convergence behavior of all the diagonal elements is investigated using (19). We sample $\log_{10}(n_k)$ at constant intervals.

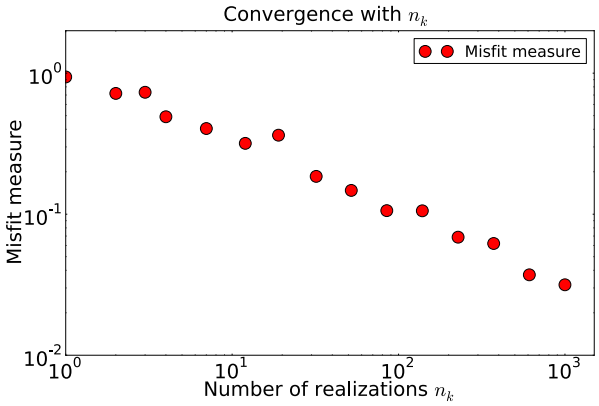

Figure 4: Misfit measure (19) plotted against the number of realizations n_k. The same number of realizations is used for the source and the receiver terms in (17).

After each value n_k the saved realizations are discarded and the process is restarted. The results of this procedure are plotted in Figure 4. Note that the average relative misfit decreases quickly with increasing n_k. There are some oscillations in this trend due to the sampling of a random distribution. Preliminary theoretical and numerical work indicates that this convergence is independent of the number of nodes as well as the number of sources and receivers. Further research is required to confirm this.

In Newton's method we solve the linear system in (2) for the model parameter update δm. The right hand side is the gradient of the misfit functional with respect to the model parameters. Calculating this gradient requires n_s forward- and n_s backward propagations (Plessix, 2006). If our preliminary investigations are correct, the computational cost of estimating the diagonal of the hessian through (17) is independent of the number of sources or receivers. Both ensemble averages in (17) are estimated with n_k forward models each. The value of n_k is determined by balancing the desire for higher accuracy with the need of reducing computational cost. When determining an appropriate value of n_k it is useful to investigate how the misfit (19) responds to an increased number of realizations, as is illustrated by Figure 4.

The cost of computing the gradient scales linearly with the number of shots in a seismic survey. However, the cost of the stochastic approximate Hessian scales independently of the number of shots. The diagonal of $\mathbf{H}_{\text{appr}}$ can be estimated, inverted and applied to the imaging condition to enhance the model parameter update $\delta\mathbf{m}$ at depth as demonstrated by (Shin et al., 2001). But it can also be used as a preconditioner for inversion schemes such as truncated newton (Métivier et al., 2012) since the cost of obtaining it does not seem to scale with the number of shots n_s. This provides an alternative to using the diagonal of the 'pseudo-hessian' of Shin et al. (2001).

We note that we can also get off-diagonal elements this way. This is demonstrated in Figure 5. A 120x120 subsection of the approximate and stochastic Hessian are compared for different numbers of realizations n_k. All of the three plots of $\mathbf{H}_{\text{s}}$

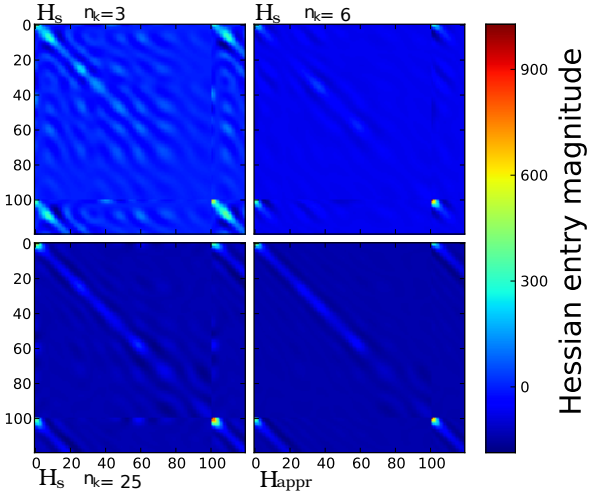

Figure 5: Comparison of $\mathbf{H}_{\text{s}}$ and $\mathbf{H}_{\text{appr}}$ for first 120x120 square of the Hessian. Different numbers of realizations n_k are compared.

are made with different noise realizations. When $n_k = 3$ some of the structure of $\mathbf{H}_{\text{appr}}$ is already visible, but there are many oscillatory artifacts remaining. These gradually become less when n_k increases. The structure of $\mathbf{H}_{\text{s}}$ with $n_k = 25$ already shows great similarity with $\mathbf{H}_{\text{appr}}$. Figure 5 shows that the stochastic Hessian converges towards the approximate Hessian both for diagonal and off-diagonal elements. If n_k is increased further the mismatch would follow a similar trend as displayed in Figure 4. Figure 5 shows that $\mathbf{H}_{\text{appr}}$ has a strongly banded structure. It may prove to be computationally effient to calculate the most significant bands with the stochastic Hessian and then use Newton's method.

CONCLUSIONS

In this abstract we have presented a new method for stochastically estimating both diagonal and off-diagonal elements of the approximate Hessian. Preliminary investigation has shown that the cost of computing the stochastic Hessian does not scale with the number of shots n_s. Because the cost of computing the gradient scales linearly in n_s, the cost of estimating the diagonal of the approximate Hessian becomes relatively inexpensive as the number of shots in the seismic survey becomes large. The inverse of the diagonal can then be used as a preconditioner for Newton's method. Alternatively the most significant bands of the approximate Hessian can be estimated and used in Newtons method.

ACKNOWLEDGEMENTS

We would like to express our gratitude to the ERL founding members consortium. Russell J. Hewett is grateful to Total for support. Also we would like to thank Kris Innanen for useful discussions.

REFERENCES

Byrd, R., P. Lu, J. Nocedal, and C. Zhu, 1995, A limited memory algorithm for bound constrained optimization: SIAM Journal of Scientific Computation, **16**, 1190–1208.

Fichtner, A., and J. Trampert, 2011, Hessian kernels of seismic data functionals based upon adjoint techniques: Geophysical Journal International, **185**, 775–798.

Martin, G., R. Wiley, and K. Marfurt, 2006, Marmousi2 - an elastic upgrade for marmousi: The Leading Edge, **25**, 156–166.

Métivier, L., R. Brossier, J. Virieux, and S. Operto, 2012, Full waveform inversion and the truncated newton method: `http://hal.archives-ouvertes.fr/hal-00763702/`.

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

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

Shin, C., S. Jang, and D.-J. Min, 2001, Improved amplitude preservation for prestack depth migration by inverse scattering theory: Geophysical Prospecting, **49**, 592–606.

Virieux, J., and S. Operto, 2009, An overview of full-waveform inversion in exploration geophysics: Geophysics, **74**.