A trust-region Newton method for frequency-domain full waveform inversion

Xavier Adriaens, Université de Liège, F.R.S-FNRS; Ludovic Métivier, Univ. Grenoble Alpes, CNRS, LJK, ISTerre; Christophe Geuzaine, Université de Liège*

SUMMARY

Exploiting Hessian information greatly enhances the convergence of full waveform inversion. A theoretically simple way to incorporate these second-order derivatives is to minimize the misfit using Newton methods. In practice however the pure Newton method is too computationally intensive to implement, because it requires inverting the Hessian operator. In addition, the misfit is not necessarily quadratic, thus the exact Newton direction is not necessarily appropriate. Consequently, it is natural to turn to inexact Newton methods, where the search direction is constructed iteratively to approximate the pure Newton direction. The bottleneck of these methods lies in the compromise to find between a direction built in few iterations, but which hardly takes the Hessian into account and a nearly exact direction which is very expensive to compute. In this work we present an inexact Newton method based on a particular trust-region algorithm, in the context of frequency-domain full waveform inversion. A numerical test is performed on the Marmousi model to compare convergence speeds with a line search based inexact Newton algorithm. This illustrates that the trust-region method is more robust and provides faster convergence for an adequate choice of trust-region parameters.

INTRODUCTION

Full waveform inversion is a data fitting technique whose aim is to recover some model parameters by minimizing the discrepancy between recorded data and data simulated by solving wave propagation problems. By nature these data are oscillatory and consequently the misfit quantifying the discrepancy features local minima (Mulder and Plessix, 2008). Global optimization techniques should ideally be used but the typically very high dimensions of the search space prohibits their use and only local optimization tools can practically be used, with care. A straightforward direction to iteratively update the model properties is of course the (preconditioned) gradient, *i.e.* the direction of steepest decrease. However it is well-known that the inverse Hessian plays a crucial role in the reconstruction in addition to offering the possibility to account for coupling effects between parameter classes for multi-parameter inversions (Pratt and Shin, 1998). While state-of-the-art methods rely on the quasi-Newton l-BFGS algorithm, which implicitly builds an approximation of the inverse Hessian operator from l saved previous gradients and model parameters, it has been illustrated that on some specific cases, inexact Newton methods can provide faster convergence. These methods compute the descent direction through few iterations of a linear system involving the Hessian operator (the Newton system). One advantage over l-BFGS is the locality of the quadratic approximation: such methods do not rely on the convergence history of the algorithm, which might yield inaccurate inverse Hessian approximation for non quadratic misfit functions.

To implement an inexact Newton scheme, one can rely either on line search algorithms, or on trust-region methods. In the former case, once a direction is chosen, the outer iteration is completed by finding the optimal length of the step that should be performed along that direction. Among the non linear optimization community, it is sometimes argued however that line search is not well suited with Newton directions, especially when the Hessian is nearly singular. Indeed when the Hessian is nearly singular, the Newton direction becomes excessively long such that the quadratic approximation implicitly made when computing it ceases to hold. Much computational effort must then be made by the line search procedure to reduce the step size (Nocedal et al., 2006). Stopping the iterative solution of the Newton system earlier appears as a solution to this problem. For example Eisenstat and Walker (1996) proposes to relax its convergence requirements such that they reflect the accuracy of the quadratic approximation.

In this contribution we propose a trust-region method, which instead limits the length of the Newton direction, also depending on the quadratic approximation accuracy.

THEORY

Full waveform inversion consists in finding the optimal model parameter m^* whose corresponding wave field u, defined through a wave propagation operator F, matches the recorded data set d after a projection R onto receivers

$$Ru = d \text{ with } F(m^*)u = f, \tag{1}$$

through the minimization of a distance $\text{dist}(\cdot,\cdot)$

$$m^* = \arg\min J(m), \text{ with } J(m) := \text{dist}(Ru(m), d). \tag{2}$$

Local optimization techniques are based on a local expansion of the misfit J around the current model estimate

$$J(m+\delta m) \approx J(m) + \{D_m J\}(\delta m) + \frac{1}{2}\{D^2_{mm} J\}(\delta m, \delta m). \tag{3}$$

This expansion can also be written in terms of the gradient j' and the Hessian operator H once an inner product $\langle\cdot,\cdot\rangle_M$ is chosen for the model space M

$$J(m+\delta m) \approx J(m) + \langle j', \delta m\rangle_M + \frac{1}{2}\langle H\delta m, \delta m\rangle_M . \tag{4}$$

Resulting from this expansion, the pure Newton direction p_N is defined as the solution of the linear system

$$Hp_N = -j'. \tag{5}$$

The large-scale nature of this linear system requires the use of Hessian-free iterative methods. The Hessian operator being symmetric, the conjugate gradient method is the ideal candidate. An additional safeguard is however added to exit prematurely when directions of negative curvature are encountered. Such directions exist because the full Hessian is not necessarily positive definite, especially far from the global minimum.

The choice of the inner product plays a central role in the inversion as it defines both gradients and Hessians and is actually

equivalent to preconditioning them (Zuberi and Pratt, 2017). Basically, a different choice for the inner product does not modify the pure Newton direction, but does modify the subspace constructed by the conjugate gradient method. A good choice can thus lead to better convergence or to better directions if the convergence criterion cannot be met, for example because of negative curvature or trust region violation.

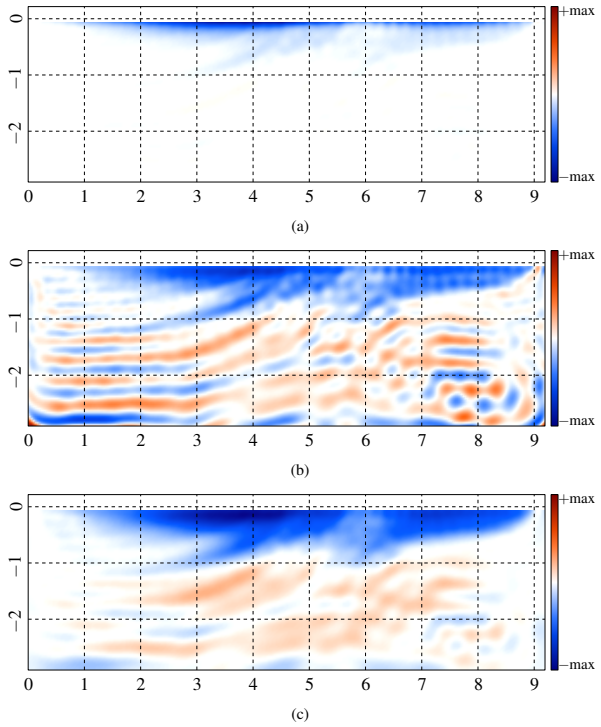

Figure 1: Gradient for different inner product choices: (a) conventional; (b) first term of (6); (c) both terms of (6).

Here we chose an inner product composed of two terms

$$\langle \cdot, \cdot \rangle_M := \left\langle \tilde{h}_{\text{diag}}^{1/2} \cdot, \tilde{h}_{\text{diag}}^{1/2} \cdot \right\rangle + \varepsilon \left\langle \vec{\nabla} \cdot, \vec{\nabla} \cdot \right\rangle. \tag{6}$$

The first term is related to the diagonal part of the Gauss-Newton Hessian $\tilde{h}_{\text{diag}}^{1/2}$ and compensates for the geometrical spreading (Pan et al., 2017) while the second term, related to spatial derivation, increases the norm of directions that are rapidly varying and prevents the inner product from being insensitive in regions where the diagonal Hessian is close to zero. This inner product is very similar to the one introduced by Zuberi and Pratt (2017), except that the Gauss-Newton diagonal Hessian is used in addition. The stabilizing parameter ε can be expressed in terms of a characteristic length l_c

$$\varepsilon := \tilde{h}_{\text{thres}} (l_c / 2\pi)^2, \tag{7}$$

with $\tilde{h}_{\text{thres}}$ a threshold value for the diagonal Hessian. In regions where the diagonal Hessian is close to the threshold, then directions with details smaller than this characteristic length l_c are penalized with respect to smoother directions. From the point of view of preconditioning, this inner product generates a depth-scaling and a Laplacian filtering. The effect of this inner product is illustrated in Figure 1 on the first gradient computed during the inversion described in the application section.

Line search

Newton methods can be combined with a line search procedure. In that case a direction p is first found by solving the Newton system approximately with the conventional conjugate gradient method (Algorithm 1). Over-solving is here avoided through the forcing term η, which is not systematically close to zero but which is instead chosen to reflect the accuracy of the second-order expansion. Eisenstat and Walker (1996) described and studied three possible choices for this sequence. These three choices were then compared in the context of acoustic imaging by Métivier et al. (2013), who advise to use the forcing sequence

$$\eta = \frac{\| j'(m_n) - j'(m_{n-1}) - \gamma_{n-1} H(m_{n-1}) p_{n-1} \|_M}{\| j'(m_{n-1}) \|_M}. \tag{8}$$

If the accuracy of the local quadratic approximation is good then this forcing term is close to zero and the Newton system is solved accurately. If not, then iterations are truncated sooner. Additional safeguards are also added to prevent this forcing term to decrease too fast.

Algorithm 1 Conventional conjugate gradient

$p \leftarrow 0$, $r \leftarrow j'$, $q \leftarrow -j'$
loop
  if $\langle Hq, q \rangle_M \leq 0$ **then return** p
  $\xi \leftarrow \langle r, r \rangle_M$
  $\alpha \leftarrow \frac{\xi}{\langle Hq, q \rangle_M}$, $\quad p \leftarrow p + \alpha q$, $r \leftarrow r + \alpha Hq$
  if $\|r\|_M < \eta \|j'\|_M$ **then return** p
  $\beta \leftarrow \frac{\langle r, r \rangle_M}{\xi}$, $\quad q \leftarrow -r + \beta q$
end loop

An appropriate length γ is then given to this direction p, ideally the global minimum along the line $m + \gamma p$. In practice however less stringent satisfactory conditions are used instead (*e.g.* Wolfe conditions) to spare expensive wave problem resolutions (Nocedal et al., 2006). The outer loop is finally obtained by repeating these two steps until convergence.

Trust region

When the Newton method is associated with a trust-region technique, both the direction and its length are found simultaneously by solving the Newton system with an additional constraint

$$H p_N = -j' \text{ with } \|p_N\|_M \leq \Delta. \tag{9}$$

This new problem can be solved approximately with a slightly modified version of the conjugate gradient method (Algorithm 2) due to Steihaug (1983). Basically there are only two modifications compared to Algorithm 1. First, the inner iterations are cropped to the trust region radius Δ when the unconstrained solution increases beyond it. Second, directions of negative curvature are followed up to the boundary of the trust region while these directions are never investigated in the conventional version. The convergence criterion is unchanged but here the forcing term is kept constant.

The size of the trust region is actually controlled by the outer iterations. The decision of modifying the trust region is based on the accuracy of the second-order expansion. When the expansion is accurate but the updates are limited by the length constraint, then the trust region radius is increased. At the opposite, when the updates are out of the range of validity of the expansion, then the trust region radius is decreased.

Algorithm 2 Steihaug conjugate gradient

$p \leftarrow 0$, $r \leftarrow j'$, $q \leftarrow -j'$
loop
  if $\langle Hq, q\rangle_M \leq 0$ **then**
    Find $\tau^* > 0$ such that $\|p+\tau^* q\|_M = \Delta$
    return $p+\tau^* q$
  end if
  $\xi \leftarrow \langle r, r\rangle_M$
  $\alpha \leftarrow \frac{\xi}{\langle Hq, q\rangle_M}$
  if $\|p+\alpha q\|_M \geq \Delta$ **then**
    Find $\tau^* > 0$ such that $\|p+\tau^* q\|_M = \Delta$
    return $p+\tau^* q$
  end if
  $p \leftarrow p+\alpha q$, $r \leftarrow r+\alpha Hq$
  if $\|r\|_M < \eta \|j'\|_M$ **then return** p
  $\beta \leftarrow \frac{\langle r, r\rangle_M}{\xi}$, $\quad q \leftarrow -r+\beta q$
end loop

The quality of the expansion is quantified by the ratio ρ between the actual decrease $\delta J_{\mathrm{a}} := J(m+p) - J(m)$ and the decrease predicted by the second-order expansion $\delta J_{\mathrm{p}} := \langle j', p\rangle_M + 0.5 \langle Hp, p\rangle_M$

$$\rho := \frac{\delta J_{\mathrm{a}}}{\delta J_{\mathrm{p}}}. \tag{10}$$

This ratio is close to one when the expansion is accurate. It plays a similar role as the forcing sequence (8). It is however based on an expansion of the misfit while the forcing sequence (8) is based on an expansion of the gradient. Standard trust-region methods directly control the radius Δ. However it is an absolute quantity, in the sense that it is compared to $\|p\|_M$, which depends on the inner product. Thus, it seems more natural to control this radius relatively to the gradient norm, which provides a length reference for the Newton system. In this way, even when the Newton system changes scale from one iteration to another, the trust region remains relevant. This particular variant (Algorithm 3) has been first introduced by Fan et al. (2016).

Algorithm 3 Fan trust region

Require: $0 \leq \rho_0 < \rho_1 < 1$ and $0 < c_0 < 1 < c_1$
  $\mu \leftarrow 1$
  loop
    $p \leftarrow$ Algorithm 2 with $\Delta = \mu \|j'\|_M$
    $\delta J_{\mathrm{a}} = J(m+p) - J(m)$
    $\delta J_{\mathrm{p}} = \langle j', p\rangle_M + 0.5 \langle Hp, p\rangle_M$
    $\rho = \delta J_{\mathrm{a}} / \delta J_{\mathrm{p}}$
    if $\rho \geq \rho_0$ **then** $m \leftarrow m+p$ **else** $m \leftarrow m$
    if $\rho < \rho_1$ **then** $\mu \leftarrow c_0 \mu$
    else if $\rho \geq \rho_1$ **and** $\|p\|_M > 0.5\Delta$ **then** $\mu \leftarrow c_1 \mu$
    else then $\mu \leftarrow \mu$
  end loop

APPLICATION

In this section we present a standard numerical test case to which both methods presented above are applied. Final details of their implementation are also introduced.

Numerical tests are performed on the 2D Marmousi acoustic model (Versteeg, 1994) (Figure 2(a)) in the frequency domain. It is here chosen that the subsurface is described by the slowness squared s^2 [s^2/km^2], thus the forward operator writes

$$F(p, s^2) = \Delta p + \omega^2 s^2 p. \tag{11}$$

Three frequencies (4, 6 and 8 [Hz]) are inverted sequentially from the lowest to the highest to avoid local minima (Bunks et al., 1995). Their spacing is chosen following the guidelines from Sirgue and Pratt (2004) regarding wavenumber coverage. We used three stabilizing parameters ε, corresponding to three characteristic lengths l_c (0.8 [km], 0.5 [km] and 0.4 [km]), one for each frequency. A surface acquisition system composed of 122 equally spaced (72 [m]) emitters-receivers is used and the misfit J is chosen as the conventional least-square distance between simulated and recorded data

$$J(s^2) = \frac{1}{2} \sum_e \sum_r \left| p_e(x_r; s^2) - d_{er} \right|^2. \tag{12}$$

For each frequency, outer iterations are stopped when the convergence criterion $J(s^2) < 3 \times 10^{-3}$ is reached. A smoothed version of the exact Marmousi model is used as an initial guess. This initial model is computed with a Laplacian filter $s^2_{\mathrm{init}} = (1 + (l_c/2\pi)^2 \Delta)^{-1} s^2_{\mathrm{exact}}$ with $l_c = 2$ [km] (Figure 2(b)).

Slowness squared and pressure fields at the three frequencies are discretized on a square grid (36 [m]) by hierarchical finite elements, respectively of order 1 and of order 2, 3, 4. A water layer (216 [m]) is also added at the top of the model but it is kept constant during the inversion. The model is spatially truncated by Sommerfeld boundary conditions (Schot, 1992). Any gradient or the Hessian vector product is computed using the (second-order) adjoint state method (Métivier et al., 2013). Recorded data are generated synthetically using the same hierarchical finite elements setting than for the inversion, to reduce numerical errors.

Line search
We choose a line search algorithm that satisfies strong Wolfe conditions and accepts steps very easily (Algorithm 3.2 from Nocedal et al. (2006) with $c_1 = 10^{-4}$, $c_2 = 0.9$ and $\alpha_0 = 1$).

Trust region
Three sets of values for ρ_0, ρ_1, c_0, and c_1 have been tested. The first one (a) is very similar to what was originally proposed by Fan et al. (2016). The other two (b,c) are more cautious because they modify the radius more rarely and when they do, it increases by a smaller factor. These parameters sets are given in Table 1. The forcing term is constant for all trust-region methods ($\eta = 0.4$).

	ρ_0	ρ_1	c_0	c_1
(a)	10^{-4}	0.25	0.2	5.
(b)	10^{-4}	0.75	0.25	2.
(c)	10^{-4}	0.9	0.5	2.

Table 1: Parameter sets for Fan trust-region algorithm

RESULTS

The squared slowness estimated with trust-region (c) method is shown in Figure 2(c). From a relatively low resolution initial guess (Figure 2(b)), full waveform inversion indeed provides a high resolution estimation of the exact model (Figure 2(a)). The images obtained with the other methods do not differ significantly.

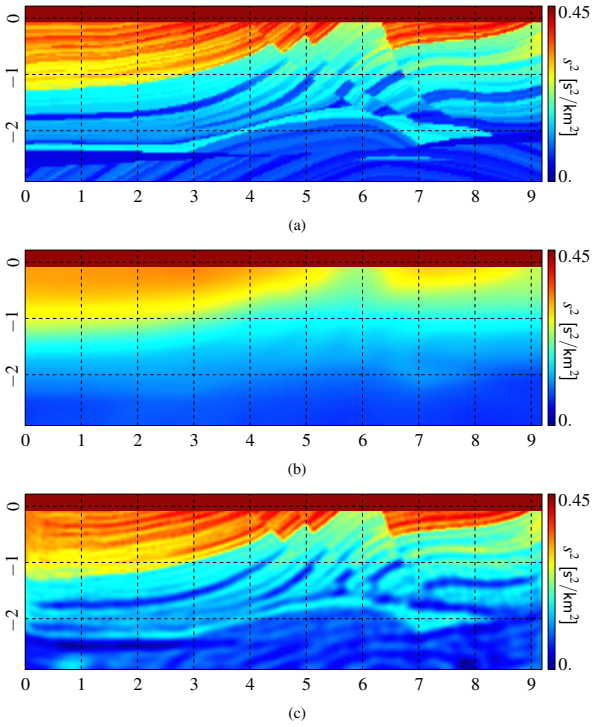

Figure 2: (a) Marmousi model (b) Initial guess (c) Inversion results (trust-region (c))

Some interesting quantities concerning the optimization algorithms are given in Table 2. For the sake of comparison, results obtained with a standard preconditioned gradient descent are also reported.

	gd	ls	tr (a)	tr (b)	tr (c)
Wave sol. (tot)	1303	432	400	310	340
Outer it. (tot)	630	42	42	33	41
Inner it. (avg)	(1.)	3.81	3.76	3.7	3.15
Rejected (%)	.04	.24	.21	.03	.07
Constrained (%)	-	-	.83	.85	.93

Table 2: Statistics related to the different inversion algorithms.

Among these indicators lies the percentage of direction refusal (rejected). Such a denial yields an additional cost because either other step length γ must be tried (line search) or the entire outer iteration must be restarted with a smaller radius Δ (trust-region). The overall computational time is not proportional to the number of outer iterations (outer it.) because the computational cost of these iterations is highly dependent on the number of conjugate gradient iterations (inner it.). Instead each outer iteration is quantified by the number of wave propagation problems it requires to solve (wave sol.). As detailed in Métivier et al. (2013), for a given model, a misfit evaluation requires one wave solution, the associated gradient one more and any Hessian vector product still requires two more. Therefore the additional cost of each inner iteration is only two wave propagation problems. The misfit is plotted against this measure of computational complexity in Figure 3.

Not surprisingly second-order methods converge orders of magnitude faster than the preconditioned gradient descent, while the average number of inner iterations is not much higher than one. Line search and trust region (a) methods give comparable results because both actually reject directions equally often. A rejected direction is potentially a heavy efficiency loss if lots of inner iterations were necessary to compute it. Plateaus appearing in the convergence curves are a consequence of these refusals. Trust-region (b) shows the best convergence, closely followed by trust-region (c). Their good performance is due to the fact that both almost do not reject any step and therefore do not waste computing time. In effect, the convergence slope looks the same for all second-order methods if plateaus are omitted, and the advantage of the trust-region method lies in reducing the number of rejected directions and thus the number of plateaus.

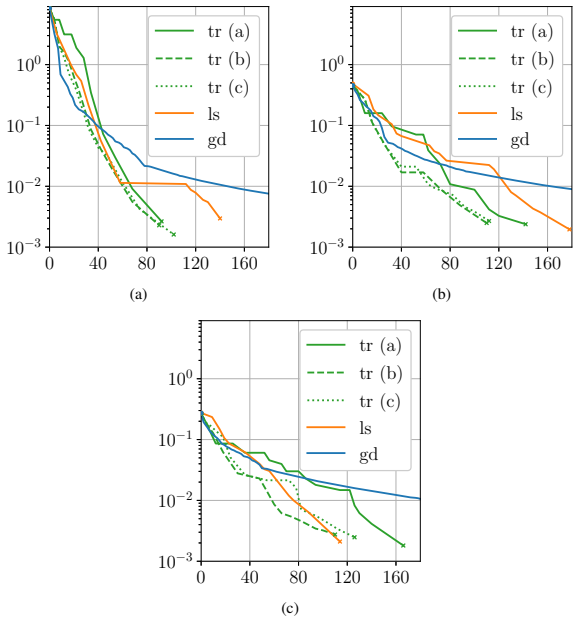

Figure 3: Misfit against the number of wave problem solved. (a) 4 [Hz] (b) 6 [Hz] (c) 8 [Hz].

CONCLUSION

We introduced a trust-region Newton method and compared its computational performance with a line search Newton method, in the context of full waveform inversion in the frequency domain. In particular we showed that the trust-region method significantly reduces over-solving and thus yields faster convergence.

ACKNOWLEDGMENTS

The authors would like to thank Anthony Royer for his help on the finite element solver used in this work (Royer et al., 2020). This research was funded by the Fonds de la Recherche Scientifique de Belgique (F.R.S.-FNRS) and by the ARC “WAVES” grant 15/19-03 from the Wallonia-Brussels Federation of Belgium. Computational resources were provided by the Consortium des Équipements de Calcul Intensif (CÉCI), funded by the Fonds de la Recherche Scientifique de Belgique (F.R.S.-FNRS) and by the Walloon Region.

REFERENCES

Bunks, C., F. M. Saleck, S. Zaleski, and G. Chavent, 1995, Multiscale seismic waveform inversion: GEOPHYSICS, **60**, 1457–1473.

Eisenstat, S. C., and H. F. Walker, 1996, Choosing the forcing terms in an inexact Newton method: SIAM Journal of Scientific Computing, **17**, 16–32.

Fan, J., J. Pan, and H. Song, 2016, A Retrospective Trust Region Algorithm with Trust Region Converging to Zero: Journal of Computational Mathematics, **34**, 421–436.

Métivier, L., R. Brossier, J. Virieux, and S. Operto, 2013, Full Waveform Inversion and the Truncated Newton Method: SIAM Journal on Scientific Computing, **35**, B401–B437.

Mulder, W., and R.-E. Plessix, 2008, Exploring some issues in acoustic full waveform inversion: Geophysical Prospecting, **56**, 827–841.

Nocedal, J., S. J. Wright, and S. M. Robinson, 2006, Numerical Optimization: Springer New York. Springer Series in Operations Research and Financial Engineering.

Pan, W., K. A. Innanen, and W. Liao, 2017, Accelerating Hessian-free Gauss-Newton full-waveform inversion via l-BFGS preconditioned conjugate-gradient algorithm: GEOPHYSICS, **82**, R49–R64.

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

Royer, A., B. Eric, and C. Geuzaine, 2020, GmshFEM: an efficient finite element library based on Gmsh: 14th World Congress on Computational Mechanics (WCCM), 19–24.

Schot, S. H., 1992, Eighty years of Sommerfeld's radiation condition: Historia Mathematica, **19**, 385–401.

Sirgue, L., and R. G. Pratt, 2004, Efficient waveform inversion and imaging: A strategy for selecting temporal frequencies: GEOPHYSICS, **69**, 231–248.

Steihaug, T., 1983, The Conjugate Gradient Method and Trust Regions in Large Scale Optimization: SIAM Journal on Numerical Analysis, **20**, 626–637.

Versteeg, R., 1994, The Marmousi experience: Velocity model determination on a synthetic complex data set: The Leading Edge, **13**, 927–936.

Zuberi, M. A., and R. G. Pratt, 2017, Mitigating nonlinearity in full waveform inversion using scaled-Sobolev pre-conditioning: Geophysical Journal International, **213**, 706–725.