A scaled gradient projection method for total variation regularized full waveform inversion

Ernie Esser[*], *Tristan van Leeuwen*[†], *Aleksandr Y. Aravkin*[‡], *and Felix J. Herrmann*[*]
[*]*University of British Columbia Dept. of Earth and Ocean Sciences,* [†]*Centrum Wiskunde & Informatica,* [‡]*IBM T.J. Watson Research Center*

SUMMARY

We propose an extended full waveform inversion formulation that includes convex constraints on the model. In particular, we show how to simultaneously constrain the total variation of the slowness squared while enforcing bound constraints to keep it within a physically realistic range. Synthetic experiments show that including total variation regularization can improve the recovery of a high velocity perturbation to a smooth background model.

INTRODUCTION

Acoustic full waveform inversion in the frequency domain can be written as the following PDE constrained optimization problem (Tarantola, 1984; Virieux and Operto, 2009; Herrmann et al., 2013)

$$\min_{m,u} \sum_{sv} \frac{1}{2}\|Pu_{sv} - d_{sv}\|^2 \quad \text{such that} \quad A_v(m)u_{sv} = q_{sv}\,, \tag{1}$$

where $A_v(m)u_{sv} = q_{sv}$ denotes the discretized Helmholtz equation. Let $s = 1,...,N_s$ index the sources and $v = 1,...,N_v$ index frequencies. We consider the model, m, which corresponds to the reciprocal of the velocity squared, to be a real vector $m \in \mathbb{R}^N$, where N is the number of points in the spatial discretization. For each source and frequency the wavefields, sources and observed data are denoted by $u_{sv} \in \mathbb{C}^N$,$q_{sv} \in \mathbb{C}^N$ and $d_{sv} \in \mathbb{C}^{N_r}$ respectively, where N_r is the number of receivers. P is the operator that projects the wavefields onto the receiver locations. The Helmholtz operator has the form

$$A_v(m) = \omega_v^2 \operatorname{diag}(m) + L\,, \tag{2}$$

where ω_v is angular frequency and L is a discrete Laplacian.

The nonconvex constraint and large number of unknowns make (1) a very challenging inverse problem. Since it is not practical to store all the wavefields, and it also is not always desirable to exactly enforce the PDE constraint, it was proposed in (van Leeuwen and Herrmann, 2013b,a) to work with a quadratic penalty formulation of (1), formally written as

$$\min_{m,u} \sum_{sv} \frac{1}{2}\|Pu_{sv} - d_{sv}\|^2 + \frac{\lambda^2}{2}\|A_v(m)u_{sv} - q_{sv}\|^2\,. \tag{3}$$

This is formal in the sense that a slight modification is needed to properly incorporate boundary conditions, and this modification also depends on the particular discretization used for L. As discussed in (van Leeuwen and Herrmann, 2013b), methods for solving the penalty formulation seem less prone to getting stuck in local minima when compared to solving formulations that require the PDE constraint to be satisfied exactly. The unconstrained problem is easier to solve numerically, with alternating minimization as well as Newton-like strategies being directly applicable. Moreover, since the wavefields are decoupled it isn't necessary to store them all simultaneously when using alternating minimization approaches.

The most natural alternating minimization strategy is to iteratively solve the data augmented wave equation

$$\bar{u}_{sv}(m^n) = \arg\min_{u_{sv}} \frac{1}{2}\|Pu_{sv} - d_{sv}\|^2 + \frac{\lambda^2}{2}\|A_v(m^n)u_{sv} - q_{sv}\|^2 \tag{4}$$

and then compute m^{n+1} according to

$$m^{n+1} = \arg\min_{m} \sum_{sv} \frac{\lambda^2}{2}\|L\bar{u}_{sv}(m^n) + \omega_v^2 \operatorname{diag}(\bar{u}_{sv}(m^n))m - q_{sv}\|^2\,. \tag{5}$$

This can be interpreted as a Gauss Newton method for minimizing $G(m) = \sum_{sv} G_{sv}(m)$, where

$$G_{sv}(m) = \frac{1}{2}\|P\bar{u}_{sv}(m) - d_{sv}\|^2 + \frac{\lambda^2}{2}\|A_v(m)\bar{u}_{sv}(m) - q_{sv}\|^2\,. \tag{6}$$

Using a variable projection argument (Aravkin and van Leeuwen, 2012), the gradient of G at m^n can be computed by

$$\begin{aligned}\nabla G(m^n) = \sum_{sv} \operatorname{Re}\Big(&\lambda^2\omega_v^2 \operatorname{diag}(\bar{u}_{sv}(m^n))^* \\ &(\omega_v^2 \operatorname{diag}(\bar{u}_{sv}(m^n))m^n + L\bar{u}_{sv}(m^n) - q_{sv})\Big)\,.\end{aligned} \tag{7}$$

A scaled gradient descent approach (Bertsekas, 1999) for minimizing G can be written as

$$\begin{aligned}\Delta m = \arg\min_{\Delta m \in R^N} &\sum_{sv} \Delta m^T \nabla G_{sv}(m^n) + \frac{1}{2}\Delta m^T H_{sv}^n \Delta m \\ &+ c_n \Delta m^T \Delta m \\ m^{n+1} = m^n + \Delta m\,,&\end{aligned} \tag{8}$$

where H_{sv}^n should be an approximation to the Hessian of G_{sv} and $c_n \geq 0$. Note that this general form includes gradient descent in the case when $H = 0$ and Newton's method when H is the true Hessian and $c = 0$. In (van Leeuwen and Herrmann, 2013b), a Gauss Newton approximation is used, where $c_n = 0$ and

$$H_{sv}^n = \operatorname{Re}(\lambda^2\omega_v^4 \operatorname{diag}(\bar{u}_{sv}(m^n))^* \operatorname{diag}(\bar{u}_{sv}(m^n))\,. \tag{9}$$

Since the Gauss Newton Hessian approximation is diagonal, it can be incorporated into (8) with essentially no additional computational expense. This corresponds to the alternating procedure of iterating (4) and (5) at least for the formal objective that is linear in m, which it may not be in practice depending on how the boundary conditions are implemented.

INCLUDING CONVEX CONSTRAINTS

To make the inverse problem more well posed we can add the constraint $m \in C$, where C is a convex set. For example, a box

constraint on the elements of m could be imposed by setting $C = \{m : m \in [B_1, B_2]\}$. The only modification of (8) is to replace the Δm update by

$$\Delta m = \arg\min_{\Delta m \in R^N} \sum_{sv} \Delta m^T \nabla G_{sv}(m^n) + \frac{1}{2}\Delta m^T H^n_{sv} \Delta m + c_n \Delta m^T \Delta m \text{ such that } m^n + \Delta m \in C \,. \tag{10}$$

This ensures $m^{n+1} \in C$ but makes Δm more difficult to compute. The problem is still tractable if C is easy to project onto or can be written as an intersection of convex constraints that are each easy to project onto. This convex subproblem is unlikely to be a computational bottleneck relative to the expense of solving for $\bar{u}(m^n)$ (4), and in fact it could even speed up the overall method if it leads to fewer required iterations.

If the eigenvalues of the Hessian of $G(m)$ are bounded for $m \in C$, c_n in (10) can be adaptively updated so that it's as small as possible while still guaranteeing that $G(m^n)$ is monotonically decreasing and limit points of $\{m^n\}$ are stationary points (Esser et al., 2013). This approach works for fairly general choices of H^n. It also avoids doing an additional line search at each iteration. A line search can potentially be expensive because evaluating $G(m)$ requires first computing $\bar{u}(m)$. For the Gauss Newton approximation in (9) where $H^n = \sum_{sv} H^n_{sv}$, adaptivity in the choice of c_n is not needed, and we can take c_n to be a fixed small value. This is the version we consider in the remainder of the paper.

Since H^n is diagonal and positive, it is straightforward to add the constraint $m \in [B_1, B_2]$. In fact

$$\Delta m = \arg\min_{\Delta m} \Delta m^T \nabla G(m^n) + \frac{1}{2}\Delta m^T H^n \Delta m + c_n \Delta m^T \Delta m \quad \text{such that } m^n + \Delta m \in [B_1, B_2] \tag{11}$$

has a closed form solution in this case. Even with these bound constraints, the model recovered using the penalty formulation can still contain artifacts and spurious oscillations, as shown in Figure 2. A simple and effective way to reduce oscillations in m via a convex constraint is to constrain its total variation (TV) to be less than some positive parameter τ. TV penalties are widely used in image processing to remove noise while preserving discontinuities (Rudin et al., 1992). It is also a useful regularizer in a wide variety of other inverse problems, especially when solving for piecewise constant or piecewise smooth unknowns. For example, TV regularization has been successfully used for electrical inverse tomography (Chung et al., 2005) and inverse wave propagation (Akcelik et al., 2002). Although the problem is similar, the formulation in (Akcelik et al., 2002) is different that what we are considering here.

TOTAL VARIATION REGULARIZATION

If we represent m as a N_1 by N_2 image, we can define

$$\|m\|_{TV} = \frac{1}{h}\sum_{ij} \sqrt{(m_{i+1,j} - m_{i,j})^2 + (m_{i,j+1} - m_{i,j})^2} = \sum_{ij} \frac{1}{h} \left\| \begin{bmatrix} (m_{i,j+1} - m_{i,j}) \\ (m_{i+1,j} - m_{i,j}) \end{bmatrix} \right\| , \tag{12}$$

which is a sum of the l_2 norms of the discrete gradient at each point in the discretized model. Assume Neumann boundary conditions so that these differences are zero at the boundary. We can represent $\|m\|_{TV}$ more compactly by defining a difference operator D such that Dm is a concatenation of the discrete gradients and $(Dm)_n$ denotes the vector corresponding to the discrete gradient at the location indexed by n, $n = 1, ..., N_1 N_2$. Then we can define

$$\|m\|_{TV} = \|Dm\|_{1,2} := \sum_{n=1}^{N} \|(Dm)_n\| \,. \tag{13}$$

Returning to (8), if we add the constraints $m \in [B_1, B_2]$ and $\|m\|_{TV} \le \tau$, then the overall iterations for solving

$$\min_m G(m) \qquad \text{such that} \qquad m \in [B_1, B_2] \text{ and } \|m\|_{TV} \le \tau \tag{14}$$

have the form

$$\Delta m = \arg\min_{\Delta m} \Delta m^T \nabla G(m^n) + \frac{1}{2}\Delta m^T H^n \Delta m + c_n \Delta m^T \Delta m \quad \text{such that } m^n + \Delta m \in [B_1, B_2] \text{ and } \|m^n + \Delta m\|_{TV} \le \tau$$
$$m^{n+1} = m^n + \Delta m \,. \tag{15}$$

SOLVING THE CONVEX SUBPROBLEMS

An effective approach for solving the convex subproblems in (15) for Δm is to use a modification of the primal dual hybrid gradient (PDHG) method (Zhu and Chan, 2008) discussed in (Esser et al., 2010; Chambolle and Pock, 2011; He and Yuan, 2012; Zhang et al., 2010) that finds a saddle point of

$$\mathscr{L}(\Delta m, p) = \Delta m^T \nabla G(m^n) + \frac{1}{2}\Delta m^T (H^n + 2c_n \mathrm{I}) \Delta m + p^T D(m^n + \Delta m) - \tau \|p\|_{\infty,2} \tag{16}$$

for $m^n + \Delta m \in [B_1, B_2]$. Here, $\|\cdot\|_{\infty,2}$ denotes the dual norm of $\|\cdot\|_{1,2}$ and takes the max instead of the sum of the l_2 norms. The modified PDHG method requires iterating

$$\begin{aligned} p^{k+1} &= \arg\min_p \tau\|p\|_{\infty,2} - p^T D(m^n + \Delta m^k) + \frac{1}{2\delta}\|p - p^k\|^2 \\ \Delta m^{k+1} &= \arg\min_{\Delta m} \Delta m^T \nabla G(m^n) + \frac{1}{2}\Delta m^T (H^n + 2c_n \mathrm{I}) \Delta m \\ &\quad + \Delta m^T D^T (2p^{k+1} - p^k) + \frac{1}{2\alpha}\|\Delta m - \Delta m^k\|^2 \\ &\quad \text{such that } m^n + \Delta m \in [B_1, B_2] \,. \end{aligned} \tag{17}$$

These iterations can be written more explicitly as

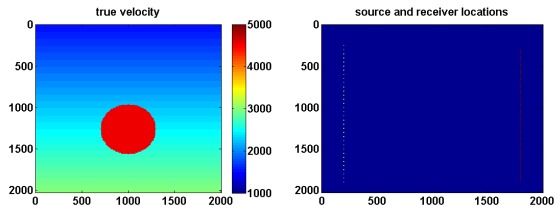
$$\begin{aligned} p^{k+1} &= p^k + \delta D(m^n + \Delta m^k) - \Pi_{\|\cdot\|_{1,2} \le \tau\delta}(p^k + \delta D(m^n + \Delta m^k)) \\ \Delta m^{k+1} &= (H^n + \xi_n \mathrm{I})^{-1} \max\big((H^n + \xi_n \mathrm{I})(B_1 - m^n), \\ &\quad \min\big((H^n + \xi_n \mathrm{I})(B_2 - m^n), \\ &\quad -\nabla G(m^n) + \frac{\Delta m^k}{\alpha} - D^T(2p^{k+1} - p^k)\big)\big) \,, \end{aligned} \tag{18}$$

where $\xi_n = 2c_n + \frac{1}{\alpha}$ and $\Pi_{\|\cdot\|_{1,2} \le \tau\delta}(z)$ denotes the orthogonal projection of z onto the ball of radius $\tau\delta$ in the $\|\cdot\|_{1,2}$ norm.

Computing this projection is of equivalent difficulty as projecting onto a simplex, which can be done efficiently. The step size restriction required for convergence is $\alpha\delta \leq \frac{1}{\|D^T D\|}$. If h is the mesh width, then it suffices to choose positive α and δ such that $\alpha\delta \leq \frac{h^2}{8}$.

NUMERICAL EXPERIMENTS

We consider a 2D synthetic experiment with a roughly 200 by 200 sized model and a mesh width h equal to 10 meters. The synthetic velocity model shown in Figure 1a has a constant high velocity region surrounded by a slower smooth background. We use an estimate of the smooth background as our initial guess m^0. Similar to Example 1 in (van Leeuwen and Herrmann, 2013b), we put $N_s = 34$ sources on the left and $N_r = 81$ receivers on the right as shown in Figure 1b. The sources q_{sv} correspond to a Ricker wavelet with a peak frequency of 30Hz.



(a) (b)

Figure 1: Synthetic velocity model (a) and source and receiver locations (b).

Data is synthesized at 18 different frequencies ranging from 3 to 20 Hertz. We consider both noise free and slightly noisy data. In the noisy case, random Gaussian noise was added to the data d_v independently for each frequency index v and with standard deviations of $\frac{.05\|d_v\|}{\sqrt{N_s N_r}}$. This may not be a realistic noise model, but it can at least indicate that the method is robust to a small amount of noise in the data.

Three different choices for the regularization parameter τ are considered: τ_{opt}, which is chosen to be the total variation of the true slowness squared, $\tau_{\text{large}} = 1000\tau_{\text{opt}}$, which is large enough so that the total variation constraint has no effect, and $\tau_{\text{small}} = .875\tau_{\text{opt}}$, slightly smaller than what would seem to be the optimal choice. Note that by using the Gauss Newton step from (11) as an initial guess, the convex subproblem in (15) converges immediately in the τ_{large} case. The parameter λ for the PDE penalty is fixed at 1 for all experiments.

We loop through the frequencies from low to high in overlapping batches of two, starting with the 3 and 4Hz data, using the computed m as an initial guess for inverting the 4 and 5Hz data and so on. For each frequency batch, we compute 50 outer iterations each time solving the convex subproblem to convergence, stopping when $\max(\frac{\|p^{k+1}-p^k\|}{\|p^{k+1}\|}, \frac{\|\Delta m^{k+1}-\Delta m^k\|}{\|\Delta m^{k+1}\|}) \leq 1\text{e}^{-5}$.

Results of the six experiments are shown in Figure 2. In both the noise free and noisy cases, including TV regularization reduced oscillations in the recovered model and led to better estimates of the high velocity region. Counterintuitively, the locations of the discontinuities were better estimated from noisy data than from noise free data. This likely happened because the noise caused there to be larger discontinuities at the receivers which then strengthened the effect of the TV regularization elsewhere. This is supported by the experiments with τ_{small}, which show that increasing the amount of TV regularization leads to better resolution of the large discontinuities but at the cost of losing contrast and staircasing the smooth background. There is also a risk of completely missing small discontinuities when τ is chosen to be too small.

We also consider a synthetic experiment with simultaneous shots. The true velocity model is a 170 by 676 2D slice from the SEG/EAGE salt model shown in Figure 3a. A total of 116 sources and 676 receivers were placed near the surface. The problem size is reduced by considering $N_{ss} < N_s$ random mixtures of the sources q_{sv} defined by

$$\bar{q}_{jv} = \sum_{s=1}^{N_s} w_{js} q_{sv} \qquad j = 1,\ldots,N_{ss}\ , \tag{19}$$

where the weights $w_{js} \in \mathcal{N}(0,1)$ are drawn from a standard normal distribution. We modify the synthetic data according to $\bar{d}_{jv} = PA_v^{-1}(m)\bar{q}_{jv}$ and use the same strategy to solve the smaller problem

$$\min_{m,u} \sum_{jv} \frac{1}{2}\|Pu_{jv} - \bar{d}_{jv}\|^2 + \frac{\lambda^2}{2}\|A_v(m)u_{jv} - \bar{q}_{jv}\|^2 \qquad \text{such that} \quad m \in [B_1, B_2] \text{ and } \|m\|_{TV} \leq \tau\ . \tag{20}$$

Starting with a good smooth initial model, results using two simultaneous shots with no added noise and for two values of τ are shown in Figure 3. With only two simultaneous shots the number of PDE solves is resuced by almost a factor of 60. TV regularization helps remove some of the artifacts caused by using so few simultaneous shots and in this case mainly reduces noise under the salt.

CONCLUSIONS AND FUTURE WORK

We presented a computationally feasible scaled gradient projection algorithm for minimizing the quadratic penalty formulation for full waveform inversion proposed in (van Leeuwen and Herrmann, 2013b) subject to additional convex constraints. We showed in particular how to solve the convex subproblems that arise when adding total variation and bound constraints on the model. Synthetic experiments suggest that when there is noisy or limited data TV regularization can improve the recovery by eliminating spurious artifacts and by more precisely identifying the locations of discontinuities.

In future work, we still need to find a practical way of selecting the regularization parameter τ. The fact that significant discontinuities can be resolved using small τ suggests that a continuation strategy that gradually increases τ could be effective. It will be interesting to investigate to what extent this strategy can avoid undesirable local minima. A possible numerical framework for this could be along the lines of the

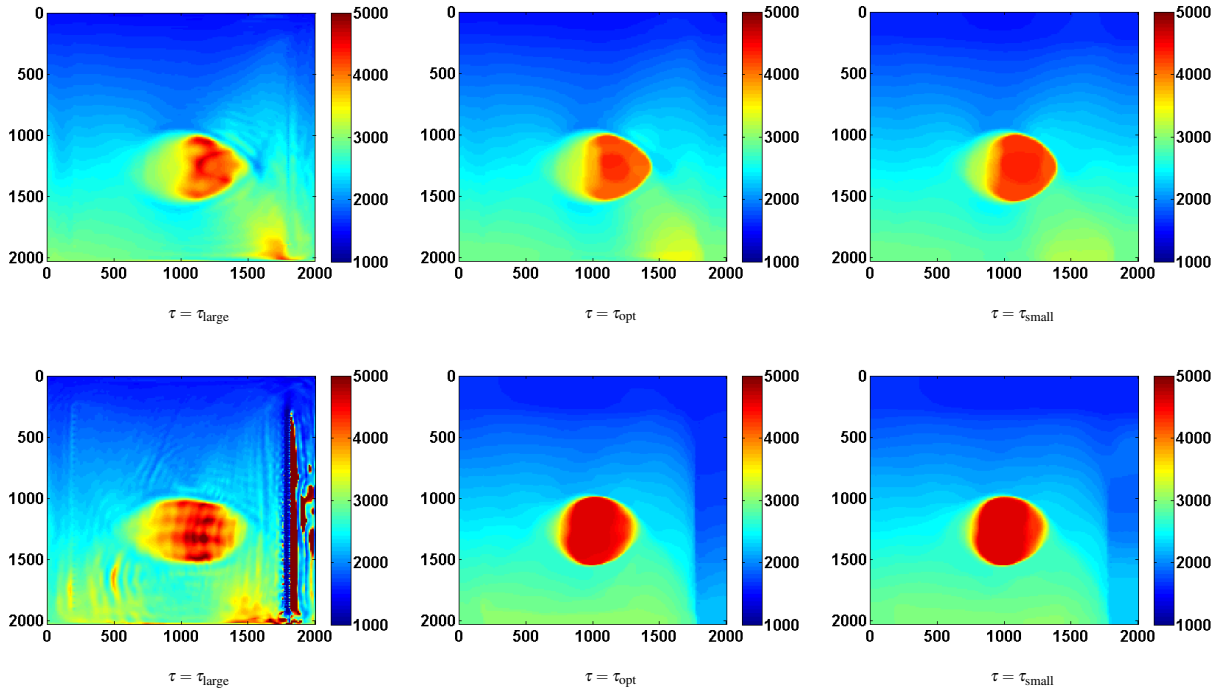


Figure 2: Recovered velocity from noise free data (first row) and noisy data (second row).

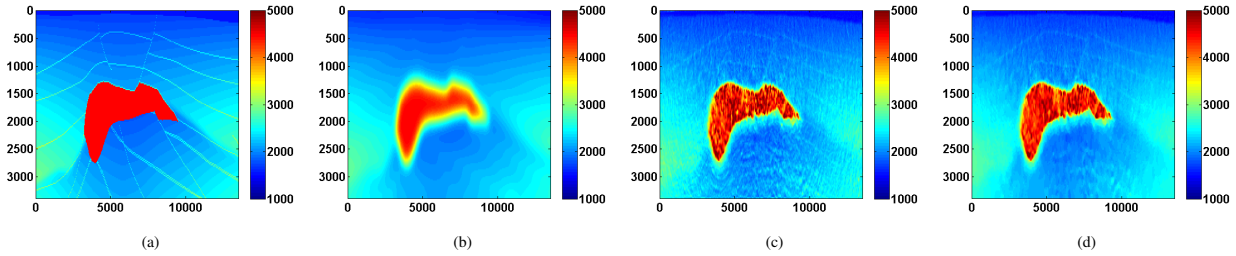


Figure 3: True velocity (a), initial velocity (b), and recovered velocity from noise free data consisting of two simultaneous shots with $\tau = \tau_{\text{large}}$ (c) and $\tau = \tau_{\text{small}}$ (d).

SPOR-SPG method in (van den Berg and Friedlander, 2011). We also intend to study more realistic numerical experiments.

ACKNOWLEDGEMENTS

Thanks to Bas Peters for insights about the formulation and implementation of the penalty method for full waveform inversion, and thanks also to Polina Zheglova for helpful discussions about the boundary conditions. This work was financially supported in part by the Natural Sciences and Engineering Research Council of Canada Discovery Grant (RGPIN 261641-06) and the Collaborative Research and Development Grant DNOISE II (CDRP J 375142-08). This research was carried out as part of the SINBAD II project with support from the following organizations: BG Group, BGP, BP, Chevron, ConocoPhillips, CGG, ION GXT, Petrobras, PGS, Statoil, Total SA, WesternGeco, Woodside.

REFERENCES

Akcelik, V., G. Biros, and O. Ghattas, 2002, Parallel multiscale Gauss-Newton-Krylov methods for inverse wave propagation: Supercomputing, ACM/IEEE . . . , **00**, 1–15.

Aravkin, A. Y., and T. van Leeuwen, 2012, Estimating nuisance parameters in inverse problems: Inverse Problems, **28**, 115016.

Bertsekas, D. P., 1999, Nonlinear Programming: Athena Scientific; 2nd edition.

Chambolle, A., and T. Pock, 2011, A first-order primal-dual algorithm for convex problems with applications to imaging: Journal of Mathematical Imaging and Vision.

Chung, E. T., T. F. Chan, and X.-C. Tai, 2005, Electrical impedance tomography using level set representation and total variational regularization: Journal of Computational Physics, **205**, 357–372.

Esser, E., Y. Lou, and J. Xin, 2013, A Method for Finding Structured Sparse Solutions to Nonnegative Least Squares Problems with Applications: SIAM Journal on Imaging Sciences, **6**, 2010–2046.

Esser, E., X. Zhang, and T. F. Chan, 2010, A General Framework for a Class of First Order Primal-Dual Algorithms for Convex Optimization in Imaging Science: SIAM Journal on Imaging Sciences, **3**, 1015–1046.

He, B., and X. Yuan, 2012, Convergence analysis of primal-dual algorithms for a saddle-point problem: from contraction perspective: SIAM Journal on Imaging Sciences, 1–35.

Herrmann, F. J., I. Hanlon, R. Kumar, T. van Leeuwen, X. Li, B. Smithyman, H. Wason, A. J. Calvert, M. Javanmehri, and E. T. Takougang, 2013, Frugal full-waveform inversion: From theory to a practical algorithm: The Leading Edge, **32**, 1082–1092.

Rudin, L., S. Osher, and E. Fatemi, 1992, Nonlinear total variation based noise removal algorithms: Physica D: Nonlinear Phenomena, **60**, 259–268.

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

van den Berg, E., and M. P. Friedlander, 2011, Sparse Optimization with Least-Squares Constraints: SIAM Journal on Optimization, **21**, 1201–1229.

van Leeuwen, T., and F. J. Herrmann, 2013a, A penalty method for PDE-constrained optimization.

———, 2013b, Mitigating local minima in full-waveform inversion by expanding the search space: Geophysical Journal International, **195**, 661–667.

Virieux, J., and S. Operto, 2009, An overview of full-waveform inversion in exploration geophysics: Geophysics, **74**, WCC1–WCC26.

Zhang, X., M. Burger, and S. Osher, 2010, A Unified Primal-Dual Algorithm Framework Based on Bregman Iteration: Journal of Scientific Computing, **46**, 20–46.

Zhu, M., and T. Chan, 2008, An Efficient Primal-Dual Hybrid Gradient Algorithm For Total Variation Image Restoration: UCLA CAM Report [08-34], 1–29.