Total variation regularization strategies in full waveform inversion for improving robustness to noise, limited data and poor initializations

Ernie Esser[†], Lluís Guasch[¶], Tristan van Leeuwen[§],
Aleksandr Y. Aravkin[#], and Felix J. Herrmann[‡]
[†]Deceased March 8th, 2015;
[¶]Sub Salt Solutions Limited;
[§]Mathematical Institute at Utrecht University;
[#]IBM T.J. Watson Research Center;
[‡]University of British Columbia Dept. of Earth, Ocean and Atmospheric Sciences

Abstract

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, removing artifacts caused by noise and limited data. Total variation-like constraints can make the inversion results significantly more robust to a poor initial model, leading to reasonable results in some cases where unconstrained variants of the method completely fail. Numerical results are presented for portions of the SEG/EAGE salt model and the 2004 BP velocity benchmark.

Disclaimer. *This technical report is ongoing work (and posted as is except for the addition of another author) of the late John "Ernie" Esser (May 19, 1980 - March 8, 2015), who passed away under tragic circumstances. We will work hard to finalize and submit this work to the peer-review literature.* Felix J. Herrmann

1 Introduction

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

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

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 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

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

Their method will be referred to as Wavefield Reconstruction Inversion (WRI) and has been further studied in [Peters et al., 2014]. The objective in (3) 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, Peters et al., 2014], 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 solving

$$\min_m F(m), \tag{6}$$

where

$$F(m) = \sum_{sv} F_{sv}(m) \tag{7}$$

and

$$F_{sv}(m) = \frac{1}{2}\|P\bar{u}_{sv}(m) - d_{sv}\|^2 + \frac{\lambda^2}{2}\|A_v(m)\bar{u}_{sv}(m) - q_{sv}\|^2 \ . \tag{8}$$

Using a variable projection argument [Aravkin and van Leeuwen, 2012], the gradient of F at m^n can be computed by

$$\begin{aligned}\nabla F(m^n) = \sum_{sv} \operatorname{Re}\big(&\lambda^2\omega_v^2 \operatorname{diag}(\bar{u}_{sv}(m^n))^* \\ &(\omega_v^2 \operatorname{diag}(\bar{u}_{sv}(m^n))m^n + L\bar{u}_{sv}(m^n) - q_{sv})\big) \ .\end{aligned} \tag{9}$$

A scaled gradient descent approach [Bertsekas, 1999] for minimizing F can be written as

$$\begin{aligned}\Delta m &= \arg\min_{\Delta m \in R^N} \Delta m^T \nabla F(m^n) + \frac{1}{2}\Delta m^T H^n \Delta m \\ m^{n+1} &= m^n + \Delta m \ ,\end{aligned} \tag{10}$$

where H^n should be a positive definite approximation to the Hessian of F at m^n. This general form includes gradient descent in the case when $H^n = \frac{1}{\Delta t}\mathrm{I}$ for some time step Δt and Newton's method when H^n is the true Hessian. In [van Leeuwen and Herrmann, 2013b], a Gauss Newton approximation is used with

$$H^n = \sum_{sv} H^n_{sv} \ , \tag{11}$$

where

$$H_{sv}^n = \mathrm{Re}(\lambda^2 \omega_v^4 \,\mathrm{diag}(\bar{u}_{sv}(m^n))^* \,\mathrm{diag}(\bar{u}_{sv}(m^n)) \ . \tag{12}$$

Since the Gauss Newton Hessian approximation is diagonal, it can be incorporated into (10) 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.

2 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, and solve

$$\min_m F(m) \quad \text{s.t.} \quad m \in C \ . \tag{13}$$

For example, a box constraint on the elements of m could be imposed by setting $C = \{m : m_i \in [b_i, B_i]\}$. The only modification of (10) is to replace the Δm update by

$$\begin{aligned} \Delta m = \arg \min_{\Delta m \in R^N} \ & \Delta m^T \nabla F(m^n) + \frac{1}{2} \Delta m^T H^n \Delta m \\ \text{s.t.} \quad & m^n + \Delta m \in C \ , \end{aligned} \tag{14}$$

leading to a scaled gradient projection method [Bertsekas, 1999, Bonettini et al., 2009].

The constraint on Δm ensures $m^{n+1} \in C$ but makes Δm more difficult to compute. Note that the simpler iterations $m^{m+1} = \Pi_C(m^n - (H^n)^{-1} \nabla(F(m^n))$ obtained by first taking a scaled gradient descent step and then projecting onto C are not in general guaranteed to converge to a solution of (13) [Bertsekas, 1999]. The problem in (14) 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. As long as the projections can be computed efficiently, 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.

The scaled gradient projection framework includes a variety of methods depending on the choice of H^n. For example, if $H^n = \frac{1}{\Delta t} \mathrm{I}$, (14) becomes a gradient projection iteration with a time step of Δt. Projected Newton like methods are included when H^n is chosen to approximate the Hessian of F at m^n. A good summary of some of the possible methods in this framework can be found in [Schmidt et al., 2012]. In particular, a robust projected quasi Newton method proposed in [Schmidt, 2009] uses a limited memory BFGS approximation of the Hessian and solves the convex subproblems for each update with a spectral projected gradient method.

For the particular application of minimizing the WRI objective subject to additional convex constraints (13), we will prefer to use projected Gauss Newton or Levenberg Marquardt type iterations because the Gauss Newton Hessian approximation for the WRI objective is diagonal. Projected gradient based methods for solving the convex subproblems are reasonable when the constraint sets are easy to project onto. However, since one of the constraints we would like to impose is a bound on the total variation (TV) of the model m, the resulting convex subproblems for computing the updates can be more naturally solved by methods that incorporate operator splitting to simplify the projection. Another consideration when solving (13) is that line search can potentially be expensive because evaluating $F(m)$ requires first solving a large linear system for $\bar{u}(m)$. Instead of doing a line search for each search direction, we prefer to introduce a damping parameter by replacing H^n with $H^n + c_n \mathrm{I}$ and adaptively adjust c_n at each iteration, rejecting iterations that don't lead to a sufficient decrease in the objective.

If ∇F is Lipschitz continuous, so that for some K

$$\|\nabla F(x) - \nabla F(y)\| \leq K \|x - y\| \text{ for all } x, y \in C,$$

and if H is a symmetric matrix, then c can be chosen large enough so that

$$F(m+\Delta m) - F(m) \leq \Delta m^T \nabla F(m) + \frac{1}{2}\Delta m^T (H + c\mathrm{I})\Delta m \tag{15}$$

for any $m \in C$ and Δm such that $m + \Delta m \in C$. So for large enough c_n, $F(m^n + \Delta m) \leq F(m^n)$ when Δm is defined by (14). In particular since

$$F(m+\Delta m) - F(m) \leq \Delta m^T \nabla F(m) + \frac{K}{2}\|\Delta m\|^2 \ ,$$

it follows that

$$F(m+\Delta m) - F(m) \leq \frac{1}{2}(K - \lambda_H^{\min} - c)\|\Delta m\|^2 + \\ \Delta m^T \nabla F(m) + \frac{1}{2}\Delta m^T (H + c\mathrm{I})\Delta m \ ,$$

where $\lambda_H^{\min}$ denotes the smallest eigenvalue of H. So choosing $c > K - \lambda_H^{\min}$ would ensure that (15) is satisfied. This would, however, be an extremely conservative choice of the damping parameter c, which could lead to a slow rate of convergence. We can also adaptively choose c_n to be as small as possible while still leading to iterations that decrease the objective by a sufficient amount, namely such that

$$F(m+\Delta m) - F(m) \leq \sigma(\Delta m^T \nabla F(m) + \frac{1}{2}\Delta m^T (H + c\mathrm{I})\Delta m) \ , \tag{16}$$

for some $\sigma \in (0, 1]$. Using the same framework as in [Esser et al., 2013], the resulting method is summarized in Algorithm 1.

Algorithm 1 A Scaled Gradient Projection Algorithm for (13)

$n = 0$; $m^0 \in C$; $\rho > 0$; $\epsilon > 0$; $\sigma \in (0, 1]$;
H symmetric with eigenvalues between $\lambda_H^{\min}$ and $\lambda_H^{\max}$;
$\xi_1 > 1$; $\xi_2 > 1$; $c_0 > \max(0, \rho - \lambda_H^{\min})$;
while $n = 0$ or $\frac{\|m^n - m^{n-1}\|}{\|m^n\|} > \epsilon$
  $\Delta m = \arg\min_{\Delta m \in C - m^n} \Delta m^T \nabla F(m^n) + \frac{1}{2}\Delta m^T (H^n + c_n \mathrm{I})\Delta m$
  if $F(m^n + \Delta m) - F(m^n) > \sigma(\Delta m^T \nabla F(m^n) + \frac{1}{2}\Delta m^T (H^n + c_n \mathrm{I})\Delta m)$
    $c_n = \xi_2 c_n$
  else
    $m^{n+1} = m^n + \Delta m$
    $c_{n+1} = \begin{cases} \frac{c_n}{\xi_1} & \text{if } \frac{c_n}{\xi_1} > \max(0, \rho - \lambda_H^{\min}) \\ c_n & \text{otherwise} \end{cases}$
    Define H^{n+1} to be symmetric Hessian approximation
      with eigenvalues between $\lambda_H^{\min}$ and $\lambda_H^{\max}$
    $n = n + 1$
  end if
end while

The particular choice of H^n we will use is the Gauss Newton Hessian defined by Equations 11 and 12. Other choices are also possible, but if H^n is a poor approximation to the Hessian of F at m^n, then a larger damping parameter c_n may be needed. Note that c_n will remain bounded because any $c_n > K - \lambda_H^{\min}$ would guarantee a sufficient decrease of F.

In addition to assuming ∇F is Lipschitz continuous, suppose $F(m)$ is coercive on C so that for any $m \in C$, $\{\tilde{m} \in C : F(\tilde{m}) \leq F(m)\}$ is bounded. Then it follows that any limit point m^* of the sequence of iterates $\{m^n\}$ defined by Algorithm 1 is a stationary point of (13) in the sense that $(m - m^*)^T \nabla F(m^*) \geq 0$ for all $m \in C$.

When $H^n + c_n\mathrm{I}$ is diagonal and positive, which it is for the Gauss Newton Hessian defined by (11), it's straightforward to add the spatially varying bound constraint $m_i \in [b_i, B_i]$. In fact

$$\begin{aligned} \Delta m = \arg\min_{\Delta m} \Delta m^T \nabla F(m^n) + \frac{1}{2}\Delta m^T (H^n + c_n\mathrm{I})\Delta m \\ \text{s.t. } m_i^n + \Delta m_i \in [b_i, B_i] \end{aligned} \tag{17}$$

has the closed form solution

$$\Delta m_i = \max\left(b_i - m_i^n, \min\left(B_i - m_i^n, -[(H^n + c_n\mathrm{I})^{-1}\nabla F(m^n)]_i\right)\right) \ .$$

Even with bound constraints, the model recovered using the penalty formulation can still contain artifacts and spurious oscillations, as shown for example in Figure 4. A simple and effective way to reduce oscillations in m via a convex constraint is to constrain its total variation 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 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 in that they directly penalize the total variation of the velocity instead of constraining the total variation of the slowness squared. Total variation regularization has also been successfully used in other recent extensions of full waveform inversion. It is used to regularize the inversion of time lapse seismic data in [Maharramov and Biondi, 2014]. It is also embedded in model updates to encourage blocky models in an FWI method that includes shape optimization in [Guo and de Hoop, 2012].

3 Total Variation Regularization

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

$$\begin{aligned} \|m\|_{TV} &= \frac{1}{h}\sum_{ij}\sqrt{(m_{i+1,j} - m_{i,j})^2 + (m_{i,j+1} - m_{i,j})^2} \\ &= \sum_{ij}\frac{1}{h}\left\|\begin{bmatrix}(m_{i,j+1} - m_{i,j})\\ (m_{i+1,j} - m_{i,j})\end{bmatrix}\right\| \ , \end{aligned} \tag{18}$$

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, $n = 1, ..., N_1N_2$. Then we can define

$$\|m\|_{TV} = \|Dm\|_{1,2} := \sum_{n=1}^{N}\|(Dm)_n\| \ . \tag{19}$$

Returning to (10), if we add the constraints $m_i \in [b_i, B_i]$ and $\|m\|_{TV} \le \tau$, then the overall iterations for solving

$$\min_m F(m) \qquad \text{s.t.} \qquad m_i \in [b_i, B_i] \text{ and } \|m\|_{TV} \le \tau \tag{20}$$

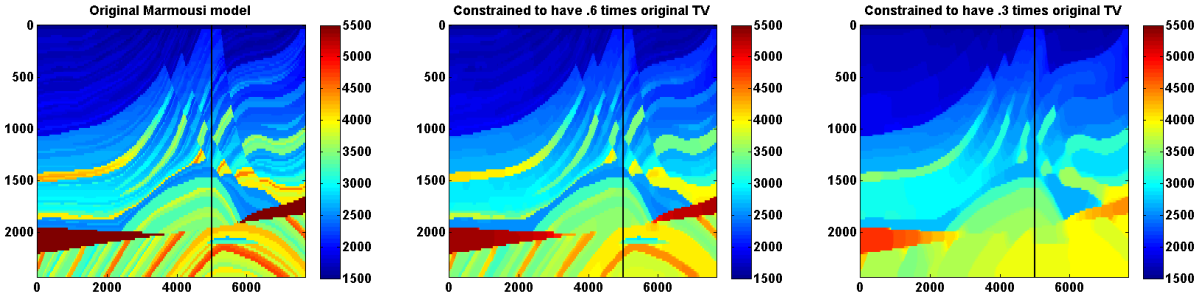
have the form

$$\begin{aligned} \Delta m = \arg\min_{\Delta m} \Delta m^T \nabla F(m^n) + \frac{1}{2}\Delta m^T (H^n + c_n\mathrm{I})\Delta m \\ \text{s.t. } m_i^n + \Delta m_i \in [b_i, B_i] \text{ and } \|m^n + \Delta m\|_{TV} \le \tau \\ m^{n+1} = m^n + \Delta m \ . \end{aligned} \tag{21}$$

To illustrate the effect of the TV constraint, consider projecting the Marmousi model shown in Figure 1a onto two sets with total variation constraints using different values of τ. Let m_0 denote the Marmousi model and let $\tau_0 = \|m_0\|_{TV}$. For the bound constraints, set $B_i = 4.4444 \times 10^{-7}$ everywhere, which corresponds to a lower bound of 1500 on the velocity. Taking advantage of the fact that these constraints can vary spatially, let $b_i = 4.4444 \times 10^{-7}$ in the water layer and $b_i = 3.3058 \times 10^{-8}$ everywhere else, which corresponds to an upper bound of 5500 on the velocity. The orthogonal projection of m_0 onto the intersection of these box and TV constraints is defined by

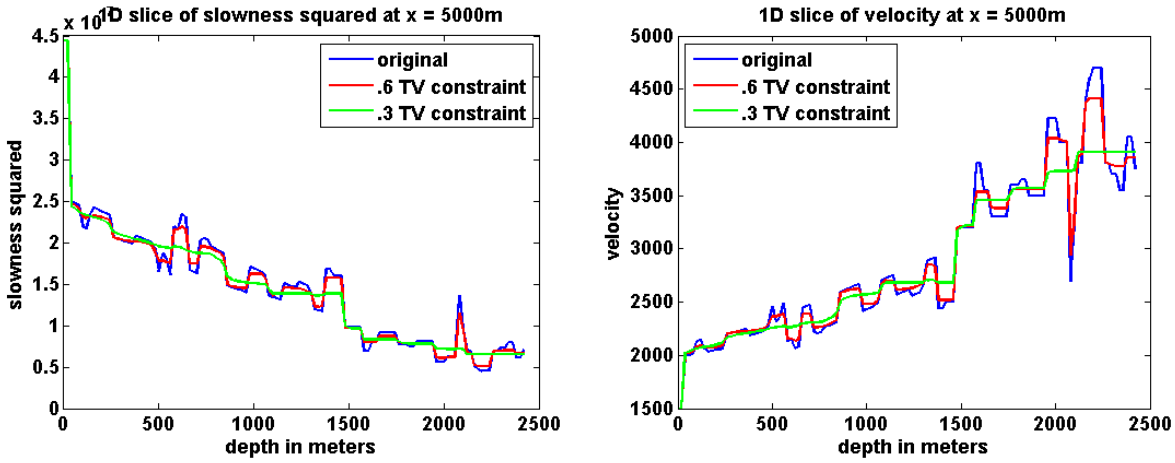
$$\begin{aligned} \Pi_C(m_0) = \arg\min_m \frac{1}{2}\|m - m_0\|^2 \\ \text{s.t. } m_i \in [b_i, B_i] \text{ and } \|m\|_{TV} \leq \tau \ . \end{aligned} \tag{22}$$

Results with $\tau = .6\tau_0$ and $\tau = .3\tau_0$ are shown in Figure 1. The vertical lines at $x = 5000$m indicate the location of the 1D depth slices shown in Figure 2 for both slowness squared and velocity.



(a) (b) (c)

Figure 1: Marmousi model (a) and projected Marmousi model for $\tau = .6\tau_0$ (b) and $\tau = .3\tau_0$ (c).



(a) (b)

Figure 2: Comparison of slices from the Marmousi model and its projections onto different TV constraints both in terms of slowness squared (a) and velocity (b).

A weighted version of the TV constraint can be obtained by replacing D with ΓD for some positive

diagonal matrix Γ. This could for instance be used to make the strength of the TV constraint vary with depth.

4 Solving the Convex Subproblems

An effective approach for solving the convex subproblems in (21) for Δm is to use a modification of the primal dual hybrid gradient (PDHG) method [Zhu and Chan, 2008] studied in [Esser et al., 2010, Chambolle and Pock, 2011, He and Yuan, 2012, Zhang et al., 2010] to find a saddle point of

$$\begin{aligned}\mathcal{L}(\Delta m, p) = \Delta m^T \nabla F(m^n) + \frac{1}{2}\Delta m^T (H^n + c_n \mathrm{I})\Delta m + g_B(m^n + \Delta m) \\ + p^T D(m^n + \Delta m) - \tau \|p\|_{\infty,2}\end{aligned} \tag{23}$$

where g_B is an indicator function for the bound constraints,

$$g_B(m) = \begin{cases} 0 & \text{if } m_i \in [b_i, B_i] \\ \infty & \text{otherwise} \end{cases} .$$

Here, $\|\cdot\|_{\infty,2}$ is using mixed norm notation to denote the dual norm of $\|\cdot\|_{1,2}$. It takes the max instead of the sum of the l_2 norms so that $\|Dm\|_{\infty,2} = \max_n \|(Dm)_n\|$ in the notation of Equation 19. The saddle point problem can be derived from the convex subproblem in (21) by representing the TV constraint as

$$\sup_p p^T D(m^n + \Delta m) - \tau \|p\|_{\infty,2} , \tag{24}$$

which equals the indicator function

$$\begin{cases} 0 & \text{if } \|D(m^n + \Delta m)\|_{1,2} \le \tau \\ \infty & \text{otherwise} \end{cases} .$$

To find a saddle point of (23), the modified PDHG method requires iterating

$$\begin{aligned} p^{k+1} &= \arg\min_p \tau \|p\|_{\infty,2} - p^T D(m^n + \Delta m^k) + \frac{1}{2\delta}\|p - p^k\|^2 \\ \Delta m^{k+1} &= \arg\min_{\Delta m} \Delta m^T \nabla F(m^n) + \frac{1}{2}\Delta m^T (H^n + c_n \mathrm{I})\Delta m \\ &\quad + \Delta m^T D^T (2p^{k+1} - p^k) + \frac{1}{2\alpha}\|\Delta m - \Delta m^k\|^2 \\ &\quad \text{s.t. } m_i^n + \Delta m_i \in [b_i, B_i] . \end{aligned} \tag{25}$$

These iterations can be written more explicitly as

$$\begin{aligned} p^{k+1} &= p^k + \delta D(m^n + \Delta m^k) - \Pi_{\|\cdot\|_{1,2} \le \tau\delta}(p^k + \delta D(m^n + \Delta m^k)) \\ \Delta m_i^{k+1} &= \max\left((b_i - m_i^n), \min\left((B_i - m_i^n),\right.\right. \\ &\quad \left.\left. [(H^n + (c_n + \frac{1}{\alpha})\mathrm{I})^{-1}(-\nabla F(m^n) + \frac{\Delta m^k}{\alpha} - D^T(2p^{k+1} - p^k))]_i\right)\right) , \end{aligned} \tag{26}$$

where $\Pi_{\|\cdot\|_{1,2} \le \tau\delta}(z)$ denotes the orthogonal projection of z onto the ball of radius $\tau\delta$ in the $\|\cdot\|_{1,2}$ norm. Computing this projection involves projecting the vector of l_2 norms onto a simplex, which can be done in linear time [Brucker, 1984]. An easy way to compute the orthogonal projection of a vector z onto the unit simplex $\{x : x_i \ge 0, \sum_i x_i = 1\}$ is to simply use bisection to find the threshold a such that $\sum_i \max(0, z_i - a) = 1$, in which case $\max(0, z_i - a)$ is the ith component of the projection.

The step size restriction required for convergence is $\alpha\delta \leq \frac{1}{\|D^T D\|}$. If h is the mesh width, then the eigenvalues of $D^T D$ are between 0 and $\frac{8}{h^2}$ by the Gershgorin Circle Theorem, so it suffices to choose positive α and δ such that $\alpha\delta \leq \frac{h^2}{8}$. In the weighted TV case, the same bound works as long as the weights are normalized so that they are all less than or equal to one in magnitude. Then the step size restriction $\alpha\delta < \frac{h^2}{8}$ that was used for the unweighted TV constraint will also satisfy $\alpha\delta < \frac{1}{\|\Gamma D\|^2}$, which is the step size restriction when the TV constraint is weighted by the diagonal matrix Γ.

The relative scaling of α and δ can have a large effect on the convergence rate of the method. A reasonable choice of fixed step size parameters is $\alpha = \frac{1}{\max(H^n + c_n \mathrm{I})}$ and $\delta = \frac{h^2 \max(H^n + c_n \mathrm{I})}{8} \leq \frac{\max(H^n + c_n \mathrm{I})}{\|D^T D\|}$. However, this choice may be too conservative. The convergence rate of the method can be improved by using the iteration dependent step sizes proposed in [Chambolle and Pock, 2011]. The adaptive backtracking strategy proposed in [Goldstein et al., 2013] can also be a practical way of choosing efficient step size parameters.

5 Curvelet Sparsity

The total variation constraint penalizes the l_1 norm of the gradient of m in order to promote sparsity of the gradient. Sparsity in other domains can also be encouraged using the same framework. For example, if $\mathcal{C}$ denotes a curvelet transform, we can replace the discrete gradient D in the TV constraint with $\mathcal{C}$ and encourage the curvelet coefficients of m to be sparse via the constraint $\|\mathcal{C}m\|_1 \leq \tau$, which limits the sum of the absolute values of complex curvelet coefficients.

6 One-Sided TV Constraint

Since velocity generally increases with depth, it is natural to penalize downward jumps in velocity. This can be done with a one-sided, one dimensional total variation constraint that penalizes increases in the slowness squared in the depth direction. Such a constraint naturally fits in the same framework previously used to impose TV constraints. Define a forward difference operator D_z that acts in the depth direction so that $D_z m$ is a concatenation of differences of the form $\frac{1}{h}(m_{i+1,j} - m_{i,j})$. To penalize the sum of the positive differences in m, we can include the constraint

$$\|\max(0, D_z m)\|_1 \leq \xi \ , \tag{27}$$

where max is understood in a componentwise sense so that $\|\max(0, D_z m)\|_1 = \sum_{ij} \max(0, \frac{1}{h}(m_{i+1,j} - m_{i,j}))$. This constraint is closely related to the hinge loss penalty commonly used for example in machine learning models such as support vector machines.

It's also possible to include positive depth weights γ_i in the one-sided TV constraint, so that it becomes

$$\sum_{ij} \max(0, \frac{\gamma_i}{h}((m_{i+1,j} - m_{i,j})) \leq \xi \ , \tag{28}$$

which also corresponds to replacing D_z by ΓD_z for a positive, diagonal matrix Γ with γ_i repeated along the diagonal. Using weights that decrease with depth may for example encourage deeper placement of discontinuities that still fit the data.

The constraint in (27) doesn't penalize model discontinuities in the horizontal direction, only in the depth direction. It's therefore likely to lead to vertical artifacts unless combined with additional regularization. It can for example be combined with a TV constraint.

The combination of TV and one-sided TV constraints still fits in same basic framework. Problem (20) becomes

$$\min_m F(m) \ \text{ s.t. } \ m_i \in [b_i, B_i] \ , \ \|m\|_{TV} \leq \tau \text{ and } \|\max(0, D_z m)\|_1 \leq \xi \tag{29}$$

and the convex subproblems (21) that need to be solved are the same except for the added one-sided TV constraint,

$$\begin{aligned}\Delta m = \arg\min_{\Delta m} \Delta m^T \nabla F(m^n) + \frac{1}{2}\Delta m^T (H^n + c_n \mathrm{I})\Delta m \\ \text{s.t. } m_i^n + \Delta m_i \in [b_i, B_i] \ , \ \|m^n + \Delta m\|_{TV} \leq \tau \\ \text{and } \|\max(0, D_z(m^n + \Delta m))\|_1 \leq \xi \ .\end{aligned} \tag{30}$$

The same method for solving the convex subproblems can be used. Analogous to (23), we want to find a saddle point of the Lagrangian

$$\begin{aligned}\mathcal{L}(\Delta m, p_1, p_2) = \Delta m^T \nabla F(m^n) + \frac{1}{2}\Delta m^T (H^n + c_n \mathrm{I})\Delta m + g_B(m^n + \Delta m) \\ + p_1^T D(m^n + \Delta m) - \tau \|p_1\|_{\infty,2} \\ + p_2^T D_z(m^n + \Delta m) - \xi \max(p_2) - g_{\geq 0}(p_2) \ ,\end{aligned} \tag{31}$$

where $g_{\geq 0}$ denotes an indicator function defined by

$$g_{\geq 0}(p_2) = \begin{cases} 0 & \text{if } p_2 \geq 0 \\ \infty & \text{otherwise} \end{cases} \ .$$

The extra terms in this Lagrangian compared to (23) follow from replacing the one-sided TV constraint by

$$\sup_{p_2} p_2^T D_z(m^n + \Delta m) - \xi \max(p_2) - g_{\geq 0}(p_2) \ , \tag{32}$$

which equals the indicator function

$$\begin{cases} 0 & \text{if } \|\max(0, D_z(m^n + \Delta m)\|_1 \leq \xi \\ \infty & \text{otherwise} \end{cases} \ .$$

This can also be seen by noting that $\xi \max(p) + g_{\geq 0}(p)$ is the Legendre transform of the indicator function

$$\begin{cases} 0 & \text{if } \|\max(0, p)\|_1 \leq \xi \\ \infty & \text{otherwise} \end{cases} \ .$$

The modified PDHG iterations are also similar. The update for p_1 is the same as in (26). The update for p_2 is very similar, and there is an extra term involving p_2 in the Δm update. Altogether, the iterations are given by

$$\begin{aligned} p_1^{k+1} &= p_1^k + \delta D(m^n + \Delta m^k) - \Pi_{\|\cdot\|_{1,2} \leq \tau\delta}(p_1^k + \delta D(m^n + \Delta m^k)) \\ p_2^{k+1} &= p_2^k + \delta D_z(m^n + \Delta m^k) - \Pi_{\|\max(0,\cdot)\|_1 \leq \xi\delta}(p_2^k + \delta D_z(m^n + \Delta m^k)) \\ \Delta m_i^{k+1} &= \max\left((b_i - m_i^n), \min\left((B_i - m_i^n),\right.\right. \\ &\quad [(H^n + (c_n + \frac{1}{\alpha})\mathrm{I})^{-1}(-\nabla F(m^n) + \frac{\Delta m^k}{\alpha} - \\ &\quad \left.\left. D^T(2p_1^{k+1} - p_1^k) - D_z^T(2p_2^{k+1} - p_2^k))]_i\right)\right) \ . \end{aligned} \tag{33}$$

The projection $\Pi_{\|\max(0,\cdot)\|_1 \leq \xi\delta}(z)$ can be computed by projecting the positive part of z, $\max(0, z)$, onto the simplex defined by $\{z : z_j \geq 0 \ , \ \sum_j z_j = \xi\delta\}$ if it doesn't already satisfy the constraint.

7 Adjoint State Formulation

All the above constraints can also be incorporated into an adjoint state formulation the PDE constrained problem in (1) in which the Helmholtz equation must be exactly satisfied. The overall problem including a general convex constraint $m \in C$ is given by

$$\min_{m,u} \sum_{sv} \frac{1}{2}\|Pu_{sv} - d_{sv}\|^2 \quad \text{s.t.} \quad A_v(m)u_{sv} = q_{sv} \quad \text{and} \quad m \in C \ . \tag{34}$$

In place of $F(m)$ defined by (8), define

$$\tilde{F}_{sv}(m) = \frac{1}{2}\|P\tilde{u}_{sv}(m) - d_{sv}\|^2 \ , \tag{35}$$

where $\tilde{u}(m)$ solves the Helmholtz equation $A_v(m)\tilde{u}_{sv} = q_{sv}$. The objective can again be written as a sum over frequency and source indices,

$$\tilde{F}(m) = \sum_{sv} \tilde{F}_{sv}(m) \ . \tag{36}$$

Using the adjoint state method, the gradient of $\tilde{F}(m)$ can be written as

$$\nabla \tilde{F}(m) = \sum_{sv} \mathrm{Re}\left[(\partial_m A_v(m))\tilde{u}_{sv}(m)\right]^* \tilde{v}_{sv}(m) \ , \tag{37}$$

where

$$A_v(m)^* \tilde{v}_{sv}(m) = P^T(d_{sv} - P\tilde{u}_{sv}(m)) \ .$$

Formally, with $A_v(m)$ defined by (2),

$$\nabla \tilde{F}(m) = \sum_{sv} \mathrm{Re}(\omega_v^2 \,\mathrm{diag}(\tilde{u}_{sv}(m))^* \tilde{v}_{sv}(m)) \ , \tag{38}$$

but this may require modification depending on how the boundary conditions are implemented. When designing a scaled gradient projection method analogous to (14) for minimizing $\tilde{F}(m)$ subject to $m \in C$ we can choose

$$H^n = \sum_{sv} \mathrm{Re}(\omega_v^4 \,\mathrm{diag}(\tilde{u}_{sv}(m^n))^* \,\mathrm{diag}(\tilde{u}_{sv}(m^n)) \tag{39}$$

even though it no longer corresponds to the Gauss Newton Hessian. Since we expect the structure of this positive diagonal Hessian approximation to be good but possibly scaled incorrectly, we may want to modify how the adaptive damping parameter c_n is implemented so that it automatically finds a good rescaling of H^n. One possibility is to replace every instance of $H^n + c_n\mathrm{I}$ in Algorithm 1 with $c_n(H^n + \nu\mathrm{I})$ for some small positive ν. With this substitution, the convex subproblems are exactly the same as in the WRI quadratic penalty formulation.

8 Numerical Experiments

We consider four 2D numerical FWI experiments based on synthetic data. All examples use a frequency continuation strategy that works with small subsets of frequency data at a time, moving gradually from low to high frequency batches.

In the first example, noisy data is generated based on a simple synthetic velocity model with sources and receivers placed vertically on the left and right sides of the model respectively, analogous to the transmission cross-well example in [van Leeuwen and Herrmann, 2013c]. Here the TV constraint is effective in removing artifacts caused by the added noise.

In the second example, very few simultaneous shots are used to simulate data based on the SEG/EAGE salt model. With a good initial model, the TV constraint is helpful in removing artifacts caused by the use of so few simultaneous shots.

The third example is based on top left portion of the 2004 BP velocity benchmark data set [Billette and Brandsberg-Dahl, 2005]. Due to the large salt body, the inversion results tend to have artifacts even when the initial model is good. In particular, the deeper part of the estimated model tends to have incorrect discontinuities. The TV constraint helps smooth out some of these artifacts while still recovering the most significant discontinuities. It can also help to compute multiple passes through the frequency batches while relaxing the TV constraint. A stronger TV constraint may initially result in an oversmoothed model estimate, but such an estimate can still be a good initial model when computing another pass with a weaker TV constraint.

The fourth example is based on the top middle portion of the same BP velocity model and shows how the one-sided TV constraint can be used to make the inversion results robust to a poor initial model. A continuation strategy that at first strongly discourages downward jumps in the estimated velocity helps prevent the method from getting stuck in bad local minima. By gradually relaxing the constraint, downward jumps in velocity are more tolerated during later passes through the frequency batches.

Before presenting the numerical examples, we first provide additional details about the discretization, boundary conditions and frequency continuation strategy.

8.1 Discretization and Boundary Conditions

The discrete Laplacian L in the Helmholtz operator (2) is defined using a simple five point stencil. We use a Robin boundary condition that can be written in the frequency domain as

$$\nabla u_{sv} \cdot n = -i\omega_v \sqrt{m} u_{sv} \ . \tag{40}$$

To implement this, L is defined with Neumann boundary conditions, which effectively removes the differences at the boundary. Using the Robin boundary condition, these differences are then replaced by $-i\omega\sqrt{m}u$. The boundary condition is incorporated into the objective (3) by replacing the PDE misfit penalties

$$\frac{\lambda^2}{2}\|Lu_{sv} + \omega_v^2 \operatorname{diag}(m) u_{sv} - q_{sv}\|^2$$

with

$$\frac{\lambda^2}{2}\|Lu_{sv} + \omega_v^2 X_{\text{int}} \operatorname{diag}(m) u_{sv} - i\omega_v X_{\text{bnd}} \operatorname{diag}(\sqrt{m}) u_{sv} - q_{sv}\|^2 \ ,$$

where X_{int} is a mask represented as a diagonal matrix with ones corresponding to points in the interior and zeros corresponding to points on the boundary. Similarly, X_{bnd} is a diagonal matrix with values of zero for interior points, $\frac{1}{h}$ for boundary edges and $\frac{2}{h}$ at the corners. This boundary condition is used when computing $\bar{u}_{sv}(m^n)$ (4). The formulas for the gradient (9) and the diagonal Gauss Newton Hessian (11) corresponding to (8) are also modified accordingly. The gradient becomes

$$\begin{aligned} \nabla F(m^n) = \sum_{sv} \operatorname{Re}\big(&\lambda^2 \omega_v \operatorname{diag}(\bar{u}_{sv}(m^n))^* (\omega_v^3 X_{\text{int}} \operatorname{diag}(\bar{u}_{sv}(m^n)) + \\ &\omega_v X_{\text{int}} (L\bar{u}_{sv}(m^n) - q_{sv}) + \frac{i}{2} X_{\text{bnd}} \operatorname{diag}(m^{\frac{-1}{2}}) (L\bar{u}_{sv}(m^n) - q_{sv}) + \\ &\frac{\omega_v}{2} X_{\text{bnd}}^2 \bar{u}_{sv}(m^n))\big) \ , \end{aligned} \tag{41}$$

and the Gauss Newton Hessian becomes

$$\begin{aligned} H^n = \sum_{sv} \operatorname{Re}\big(&\lambda^2 \omega_v^4 X_{\text{int}} \operatorname{diag}(\bar{u}_{sv}(m^n))^* \operatorname{diag}(\bar{u}_{sv}(m^n)) - \\ &\frac{iw\lambda^2}{4} X_{\text{bnd}} \operatorname{diag}(m^{\frac{-3}{2}}) \operatorname{diag}(\operatorname{conj}(\bar{u}_{sv}(m^n))) \operatorname{diag}(L\bar{u}_{sv}(m^n) - q_{sv})\big) \ . \end{aligned} \tag{42}$$

8.2 Frequency Continuation

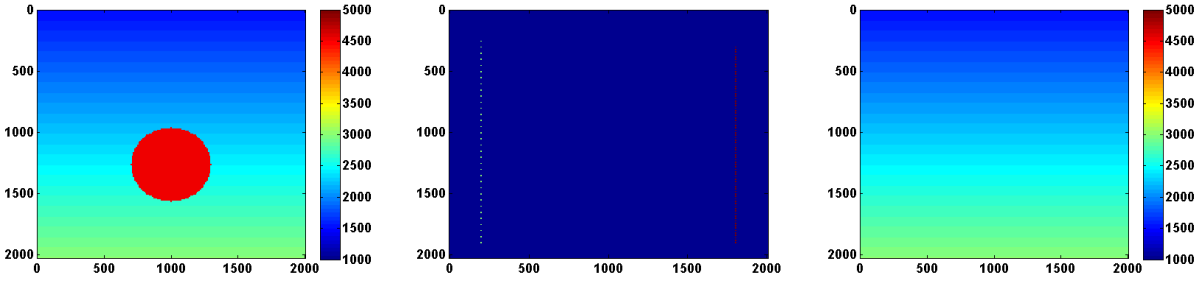
We choose to work with small batches of frequency data at a time, moving from low to high frequencies in overlapping batches of two. This frequency continuation strategy does not guarantee that we solve the overall problem after a single pass from low to high frequencies. But it is far more computationally tractable than minimizing over all frequencies simultaneously, and the continuation strategy of moving from low to high frequencies helps prevent the iterates from tending towards bad local minima.

For example, if the data consists of frequencies starting at 3Hz sampled at intervals at 1Hz, then we would start with the 3 and 4Hz data, use the computed m as an initial guess for inverting the 4 and 5Hz data and so on. For each frequency batch, we will compute at most 25 outer iterations, each time solving the convex subproblem to convergence, stopping when $\max(\frac{\|p^{k+1}-p^k\|}{\|p^{k+1}\|}, \frac{\|\Delta m^{k+1}-\Delta m^k\|}{\|\Delta m^{k+1}\|}) \leq 1 \times 10^{-4}$.

Since the magnitude of the data depends on the frequency, it may be a good idea to compensate for this by incorporating frequency dependent weights in the definition of the objective (7). However, if we only work with small frequency batches in practice, then these weights don't have a significant effect.

8.3 Sequential Shot Example with Noisy Data

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 3a 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 $N_s = 39$ sources on the left and $N_r = 96$ receivers on the right as shown in Figure 3b. The sources q_{sv} correspond to a Ricker wavelet with a peak frequency of 30Hz.



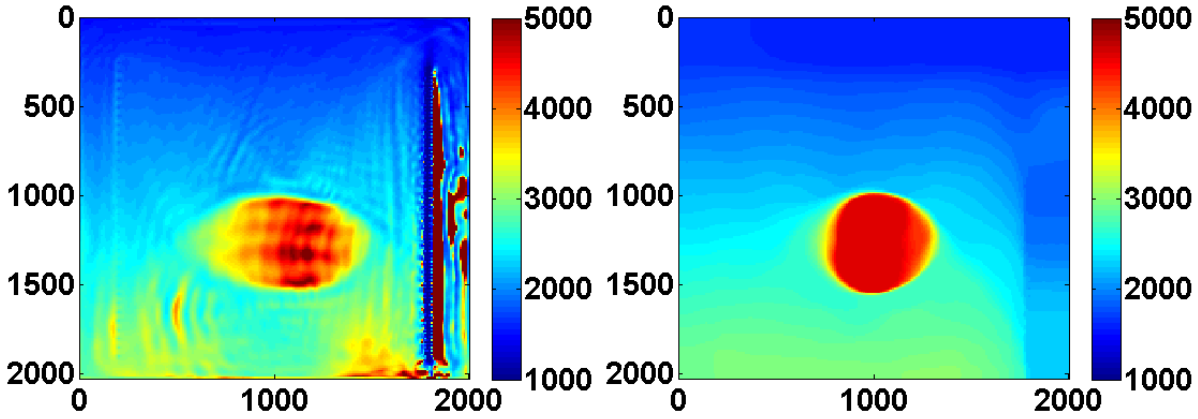
(a) (b) (c)

Figure 3: Synthetic velocity model (a), source and receiver locations (b) and initial velocity (c).

Data is synthesized at 18 different frequencies ranging from 3 to 20 Hertz. Random Gaussian noise is 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.

Two different choices for the regularization parameter τ are considered: $\tau = .875\tau_{\text{true}}$, where τ_{true} is the total variation of the true slowness squared, and $\tau = 1000\tau_{\text{true}}$, which is large enough so that the total variation constraint has no effect. Note that by using the Gauss Newton step from (17) as an initial guess, the convex subproblem in (21) converges immediately in the large τ case. The parameter λ for the PDE penalty is fixed at 1 for all experiments.

Results of the two experiments are shown in Figure 4. Including TV regularization reduced oscillations in the recovered model and led to better estimates of the high velocity region. Choosing τ too large could make the TV constraint have no effect on the result, but there is also a risk of oversmoothing and completely missing small discontinuities when τ is chosen to be too small.

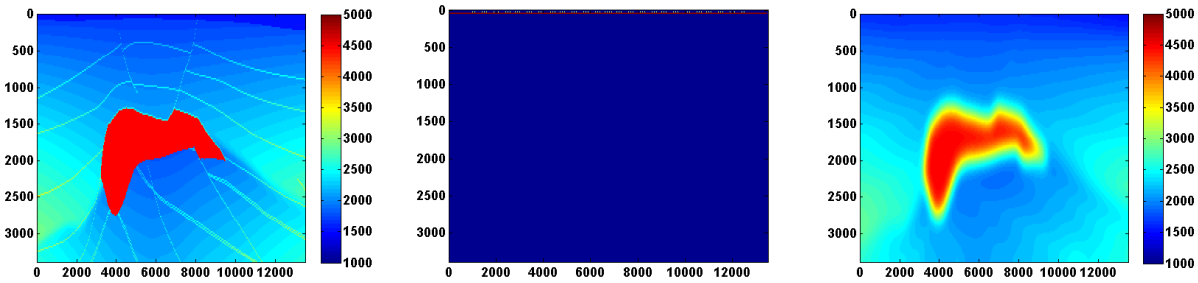


(a) (b)

Figure 4: Results for $\tau = 1000\tau_{\text{true}}$ (a) and $\tau = .875\tau_{\text{true}}$ (b)

8.4 Simultaneous Shot Example with Noise Free Data

Total variation regularization, in addition to making the inversion more robust to noise in he data, can also remove artifacts that arise when using few simultaneous shots. We consider a synthetic experiment with simultaneous shots where the true velocity model is a 170 by 676 2D slice from the SEG/EAGE salt model shown in Figure 5a. A total of 116 sources and 676 receivers are placed near the surface, and a very good smooth initial model is used. We use frequencies between 3 and 33 Hz and the same frequency continuation strategy as before that works with overlapping batches of two.



(a) (b) (c)

Figure 5: Synthetic velocity model (a), source and receiver locations (b) and initial velocity (c).

The problem size is reduced by considering $N_{ss} < N_s$ random mixtures of the sources q_{sv} defined by

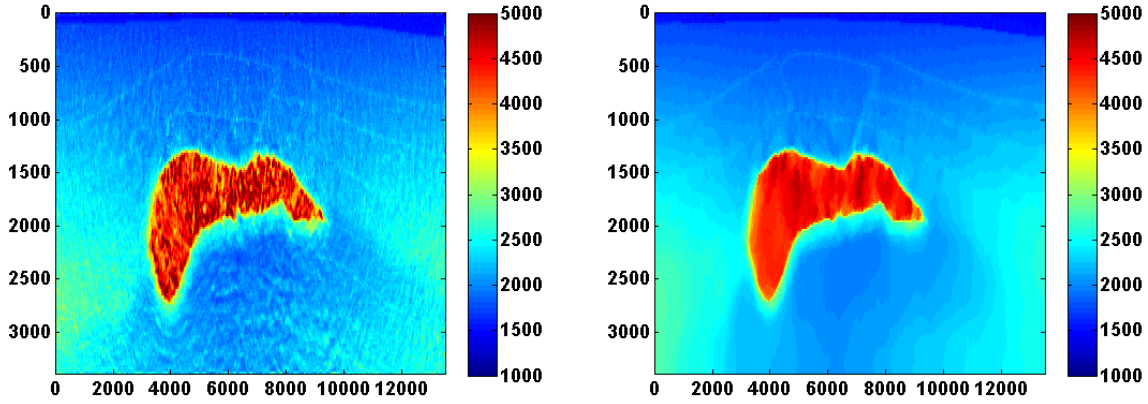
$$\bar{q}_{jv} = \sum_{s=1}^{N_s} w_{js} q_{sv} \qquad j = 1, ..., N_{ss} \ , \tag{43}$$

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

$$\begin{aligned} \min_{m,u} \sum_{jv} \frac{1}{2}\|Pu_{jv} - \bar{d}_{jv}\|^2 + \frac{\lambda^2}{2}\|A_v(m)u_{jv} - \bar{q}_{jv}\|^2 \\ \text{s.t.} \qquad m_i \in [b_i, B_i] \text{ and } \|m\|_{TV} \leq \tau \ . \end{aligned} \tag{44}$$

Results using two simultaneous shots with no added noise and for two values of τ are shown in Figure 6. With only two simultaneous shots the number of PDE solves is reduced by almost a factor of 60. TV regularization helps remove some of the artifacts caused by using so few simultaneous shots. In this case it mainly reduces noise under the salt and produces a smoother estimate of the salt body.



(a) (b)

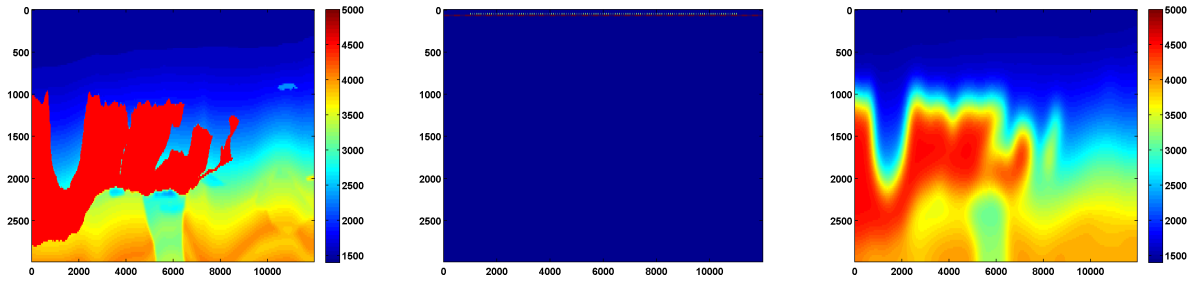
Figure 6: Recovered velocity from noise free data consisting of two simultaneous shots with $\tau = 1000\tau_{\text{true}}$ (a) and $\tau = .5\tau_{\text{true}}$ (b).

8.5 Benefit of Multiple Passes

The results of the WRI method with frequency continuation can be improved by computing multiple passes through the frequencies, each time using the model estimated after the previous pass as a warm start [CITE BAS'S BG EXAMPLE]. The constrained WRI method also benefits from multiple passes. [INCLUDE SECOND PASS FOR PREVIOUS EXAMPLE] To illustrate this, consider the 150 by 600 model shown in Figure 7a, which corresponds to a downsampled upper left portion of the 2004 BP velocity benchmark data set [Billette and Brandsberg-Dahl, 2005]. We will start with the smooth model shown in 7c and again loop through the frequencies ranging from 3 to 20 Hertz from low to high in overlapping batches of two. The bound constraints on the slowness squared are defined to correspond to minimum and maximum velocities of 1400 and 5000 m/s respectively. The τ parameter for the TV constraint is chosen to be .9 times the TV of the ground truth model. To reduce computation time, we again use two simultaneous shots, but now with new Gaussian weights w_{sj} redrawn every time the model is updated. The estimated model after computing 25 outer iterations per frequency batch is shown in Figure 8a. Using this as a warm start for a second pass through the frequency batches from low to high yields the improved result in Figure 8b.

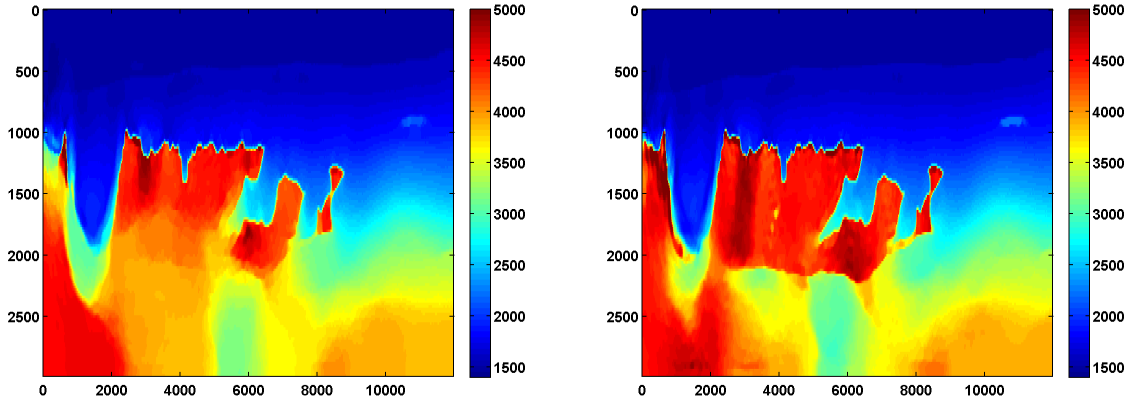
We run the same experiment using a weighted TV constraint with weights that decrease in depth as shown in Figure 9. The results shown in Figure 10 are similar, although some differences are apparent due to the weighted TV constraint being weaker in the deeper part of the model.

For comparison, two passes through the frequency batches are computed without any TV constraint, just the bound constraints. The results shown in Figure 11 appear much noisier. The estimated result after just



(a) (b) (c)

Figure 7: Top left portion of BP 2004 velocity model (a), source and receiver locations (b) and initial velocity (c).



(a) (b)

Figure 8: Recovered velocity with a TV constraint from noise free data and a good smooth initial model after one pass with $\tau = .9\tau_{\text{true}}$(a) and a second pass with $\tau = .99\tau_{\text{true}}$ (b) through small frequency batches from 3 to 20 Hz.

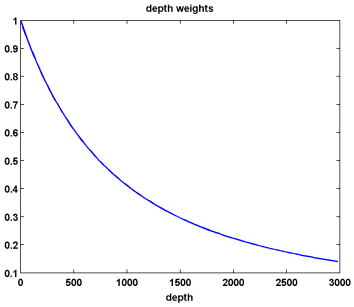
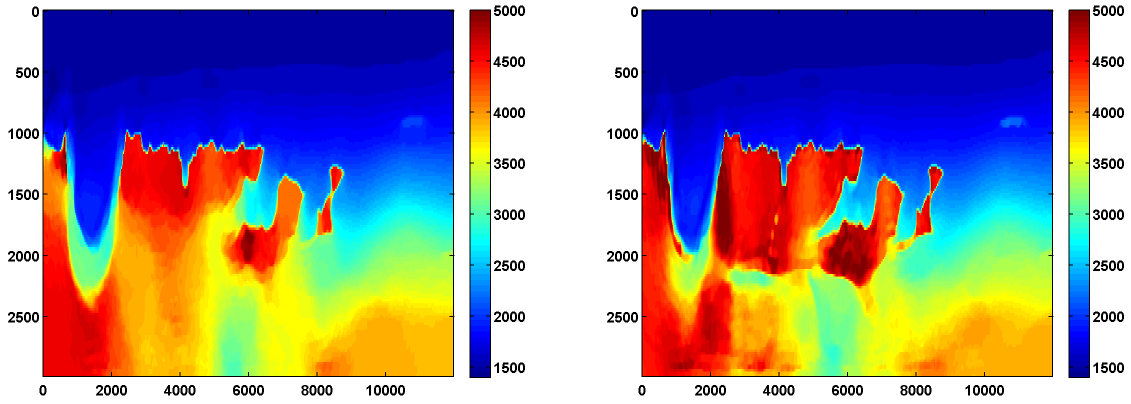


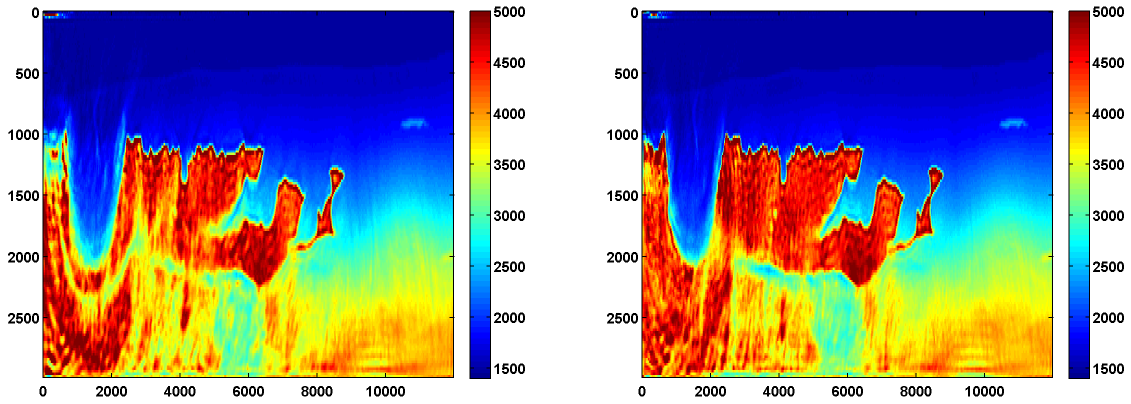
Figure 9: Decreasing depth weights used in the weighted TV experiments.



(a) (b)

Figure 10: Recovered velocity with a depth weighted TV constraint from noise free data and a good smooth initial model after one pass with $\tau = .9\tau_{\text{true}}$(a) and a second pass with $\tau = .99\tau_{\text{true}}$ (b) through small frequency batches from 3 to '20" Hz.

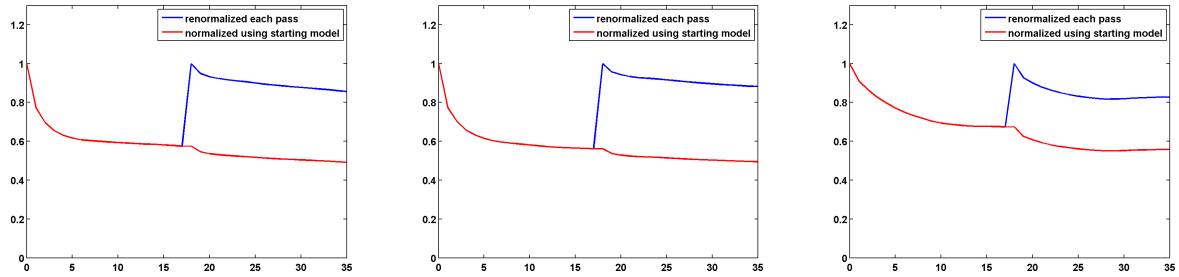
one pass is notable because it has an oscillation just below the top of the salt on the left side. This mostly disappears after the second pass but isn't present at all after one pass with either the TV or weighted TV constraints.



(a) (b)

Figure 11: Recovered velocity without TV constraint from noise free data and a good smooth initial model after one pass (a) and after two passes (b) through small frequency batches from 3 to 20 Hz.

Figure 12 shows a comparison of the relative model errors defined by $\frac{\|m-m_{\text{true}}\|}{\|m-m_{\text{init}}\|}$. The TV constrained examples both have lower model error than the unconstrained example.

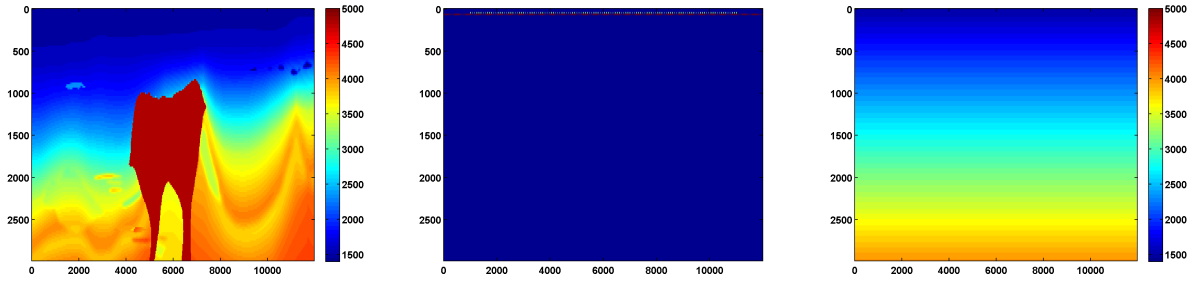


(a) (b) (c)

Figure 12: Relative model error versus frequency batch over two passes through frequency batches with a TV constraint (a), a depth weighted TV constraint (b) and no TV constraint (c).

8.6 One-Sided TV Constraint for Recovering Salt Bodies from Poor Initial Models

The inversion results shown in Figures 6b and 8b relied heavily on having good initial models, which for those examples were defined by smoothing the ground truth models. Consider the 150 by 600 velocity model shown in Figure 13a, which is the top middle portion of the same 2004 BP velocity benchmark data set, also downsampled. The same strategy that worked well for the top left portion of the model again works well if we initialize with a smoothed version of the true model. However, if we instead start with the poor initial model shown in Figure 20a, then the method appears to get stuck near a local minimum for which the inversion result is poor.



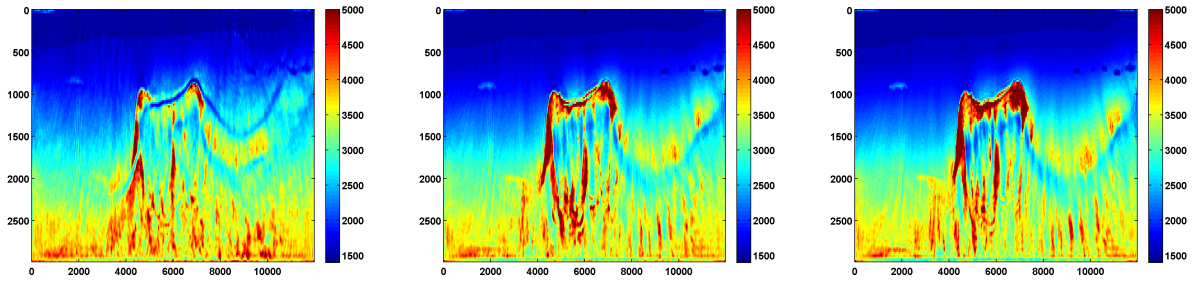
(a) (b) (c)

Figure 13: Top middle portion of BP 2004 velocity model (a), source and receiver locations (b) and initial velocity (c).

As shown in Figure 14, the WRI method with just bound constraints yields a very noisy result and fails to improve significantly even after multiple passes through the frequency batches.

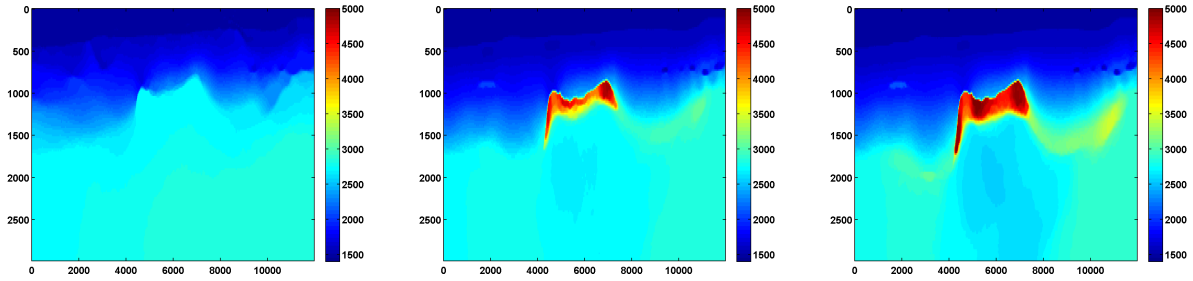
With a TV constraint added, the method still tends to get stuck at a poor solution, even with multiple passes and different choices of τ. Figure 15 shows the estimated velocity models after three passes, where increasing values of τ were used so that the TV constraint was weakened slightly after each pass.

Using a depth weighted TV constraint with the weights defined as in Figure 9 can slightly improve the solution because weights that decrease in depth prefer placing discontinuities in deeper parts of the model where the jumps are not as penalized, thus encouraging the lower boundary of the salt to move deeper as long at it continues to fit the data. The depth weighted TV result using Figure 15c as a starting model



(a) (b) (c)

Figure 14: Recovered velocity with a no TV constraint from noise free data and a poor initial model after one pass (a) after a second pass (b) and after a third pass (c) through small frequency batches from 3 to 20 Hz.



(a) (b) (c)

Figure 15: Recovered velocity with a TV constraint from noise free data and a poor initial model after one pass with $\tau = .75\tau_{\text{true}}$(a), a second pass with $\tau = .825\tau_{\text{true}}$ (b) and a third pass with $\tau = .9\tau_{\text{true}}$.

and $\tau = .9\tau_{\text{true}}$ is shown in Figure 16. While it's an improvement, it doesn't significantly update the initial model.

To discourage getting stuck in local minima that have spurious downward jumps in velocity, we consider adding the one-sided TV constraint and implementing a continuation strategy in the ξ parameter that starts with a small value and gradually increases it for each successive low to high pass through the frequency batches. This encourages the initial velocity estimate to be nearly monotonically increasing in depth. As ξ increases, more downward jumps are allowed. Again starting with the poor initial model in Figure 20a, Figure 17 shows the progression of velocity estimates over eight passes. The sequence of ξ parameters as a fraction of $\|\max(0, D_z m_{\text{true}})\|_1$ is chosen to be $\{.01, .05, .10, .15, .20, .25, .40, .90\}$. The τ parameter is fixed at $.9\tau_{\text{true}}$ throughout. Although small values of ξ cause vertical artifacts, the continuation strategy is surprisingly effective at preventing the method from getting stuck at a poor solution. As ξ increases, more downward jumps in velocity are allowed, and the model error continues to significantly decrease during each pass.

Figure 18 shows a comparison of the relative model errors for the previous three examples. Only the one-sided TV continuation results continue improving significantly after several passes.

Even with a poor initial model, the one-sided TV constraint with continuation is able to recover the main features of the ground truth model. Poor recovery near the left and right boundaries is expected because the sources along the surface start about 1000m away from these boundaries.

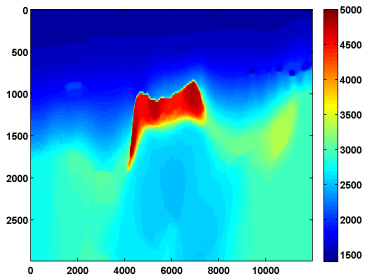


Figure 16: Recovered velocity using depth weights from Figure 9, the starting model from Figure 15c and $\tau = .9\tau_{\text{true}}$.

8.7 Constrained Adjoint State Method Comparisons

The TV and one-sided TV constraints can also be used to improve the results of the adjoint state method. We still apply Algorithm 1 to the adjoint state formulation of (34), but with $H^n + c_n \mathrm{I}$ replaced by $c_n(H^n + \nu \mathrm{I})$ where H^n is defined by (39). Since this doesn't approximate the Hessian as well as the Gauss Newton Hessian used in the quadratic penalty approach, some iterations are occasionally rejected as c_n is updated.

If we consider again the model in Figure 13a and start with the initial model in Figure 20a, the adjoint state methods with only bound constraints fails completely. The first update moves the model away from the ground truth, and the method quickly stagnates at the very poor result shown in Figure 19

On the other hand, if we include TV and one-sided TV constraints and use the same continuation strategy used to generate the results in Figure 17, the results are nearly as good. The constrained adjoint state results are shown in Figure 20. Compared to the constrained WRI method, the results are visually slightly worse near the top and sides of the model. Additionally, the model error doesn't decrease as much, but it's encouraging that it once again continues to decrease during each pass instead of stagnating at a poor solution. Continuation in the ξ parameter for the one-sided TV constraint appears to be a promising strategy for preventing both the constrained WRI and adjoint state methods from stagnating in a poor solution when starting from a bad initial model.

Figure 21 shows a comparison of the relative model errors for the previous two adjoint state examples. As in Figure 18 the one-sided TV continuation results continue improving significantly after several passes.

9 Ongoing Work

Many more experiments are required to answer lingering questions.

The numerical examples should be recomputed without inverse crime. Currently the same modeling operator was used both for inversion and for generating the synthetic data. In combination with the simplistic Robin boundary condition, this may conspire to make the inverse problem easier. However, the comparisons between methods and the relative benefits of the TV constraints are still meaningful. An interesting test would be to simply generate the data using more sophisticated modeling and boundary conditions while continuing to use the simple variants for inversion.

To remove any effect of randomness, the examples should be recomputed using sequential shots instead of simultaneous shots with the random weights redrawn every model update. This will be more computationally expensive, but it's doable and is not expected to significantly alter the results.

The numerical examples using the one-sided TV continuation should be recomputed without the depth weighting. This is expected to have only a small effect on the results. It's possible, however, that different parameter choices may be required.

It may be possible to improve the results shown in Figure 15 with different continuation strategies in the τ parameter. Experiments that started with very small values of τ were not initially promising, but so few

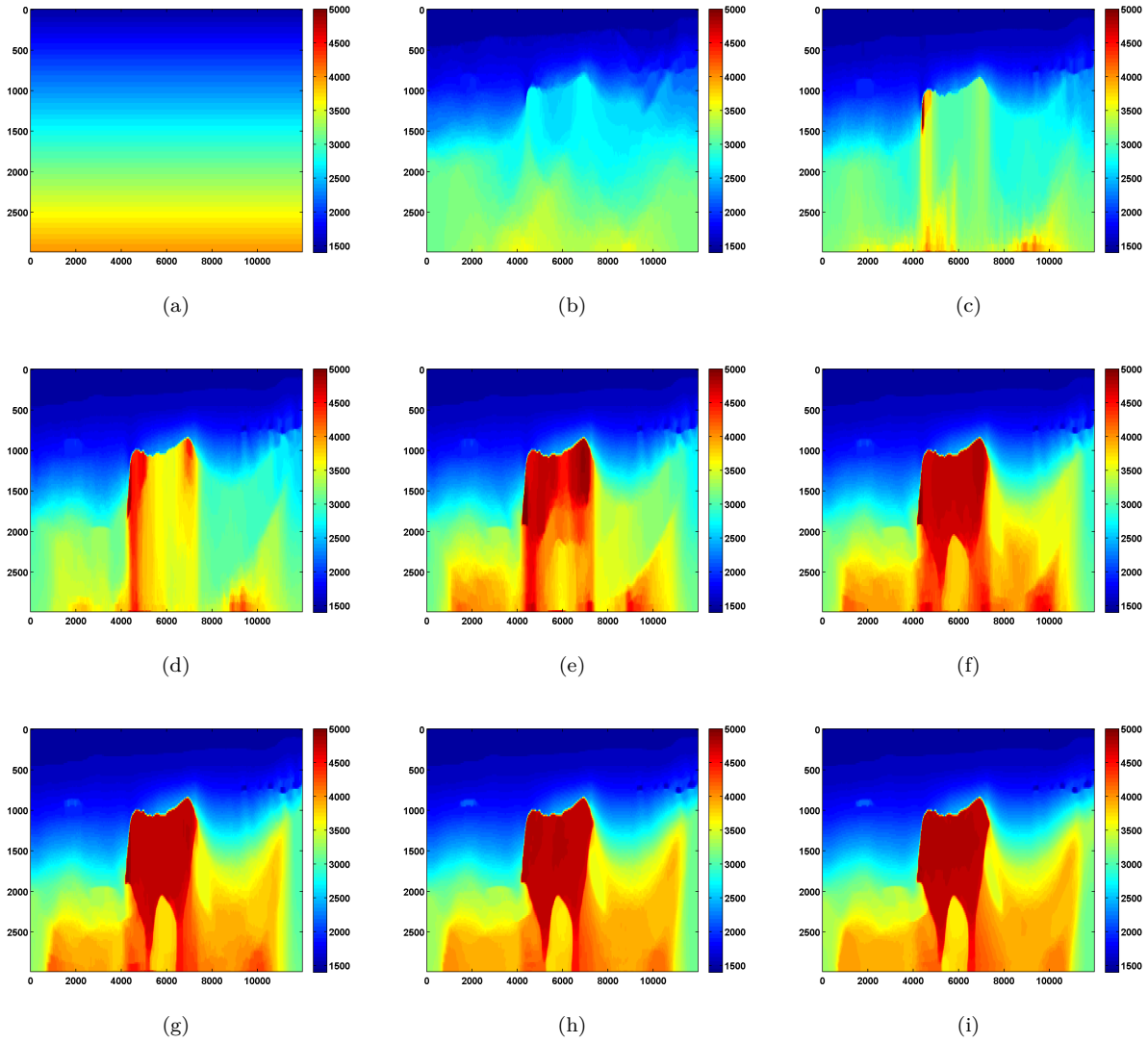
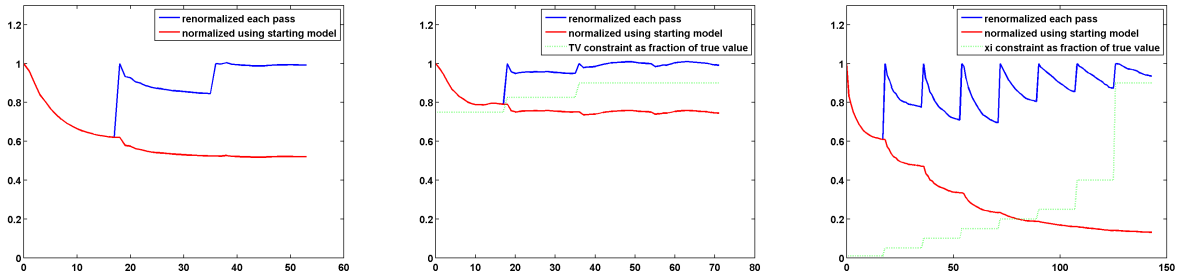


Figure 17: Initial velocity (a) and recovered velocity with one-sided TV continuation corresponding to $\frac{\xi}{\xi_{\text{true}}} =$ .01 (a), .05 (b), .10 (c), .15 (d), .20 (e), .25 (f), .40 (g), .90 (h).

experiments have been computed that the parameter selections are certainly not optimal.

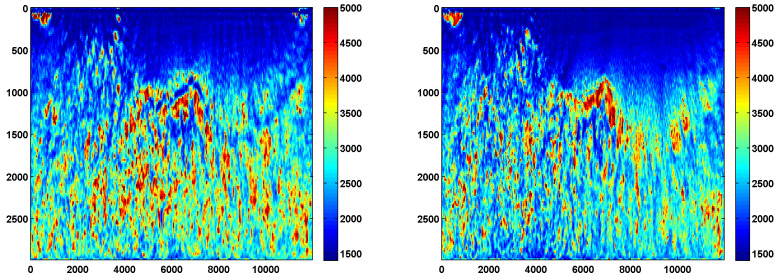
More practical methods for selecting the parameters and more principled ways of choosing continuation strategies are needed. Currently the τ and ξ parameters are chosen to be proportional to values corresponding to the true solution. Although the true solution clearly isn't known in advance, it may still be reasonable base the parameter choices on estimates of its total variation (or one-sided TV). It would be even better to develop continuation strategies that don't rely on any assumptions about the solution but that are still effective at regularizing early passes through the frequency batches to prevent the method from stagnating at poor solutions.

There is a lot of room for improvement in the algorithm used to solve the convex subproblems. There are for example methods such as the one in [Chambolle and Pock, 2011] that have better theoretical rates of convergence and are straightforward to implement.



(a) (b) (c)

Figure 18: Relative model error versus frequency batch without TV constraints (a), with TV constraints (including depth weights in the last pass) (b) and with both TV and one-sided TV constraints (including depth weights in all passes) (c).



(a) (b)

Figure 19: Recovered velocity using the adjoint state method with no TV constraints from noise free data and a poor initial model after one pass (a) and after a second pass through small frequency batches from 3 to 20 Hz.

It will be important to study how the method scales to larger problems. One effect is that the convex subproblems are likely to require more iterations. If they become too computationally expensive, then it may be better to replace the adaptive step size strategy in Algorithm 1 with a more standard line search. It would be good in any case to compare the efficiency of some other line search tecnhiques.

Continuation strategies that take advantage of the ability to enforce spatially varying bound constraints have still not been explored here. It should be possible for example to control what parts of the model are allowed to update. Other constraints can be considered that keep the updates closer to the initial model in places where it is more trusted.

If possible, it would also be good to experiment with different frequency continuation and sampling strategies as well as methods that attempt to minimize over all frequencies simultaneously.

10 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 TV, one-sided TV and bound constraints on the model. The proposed framework is general, and the

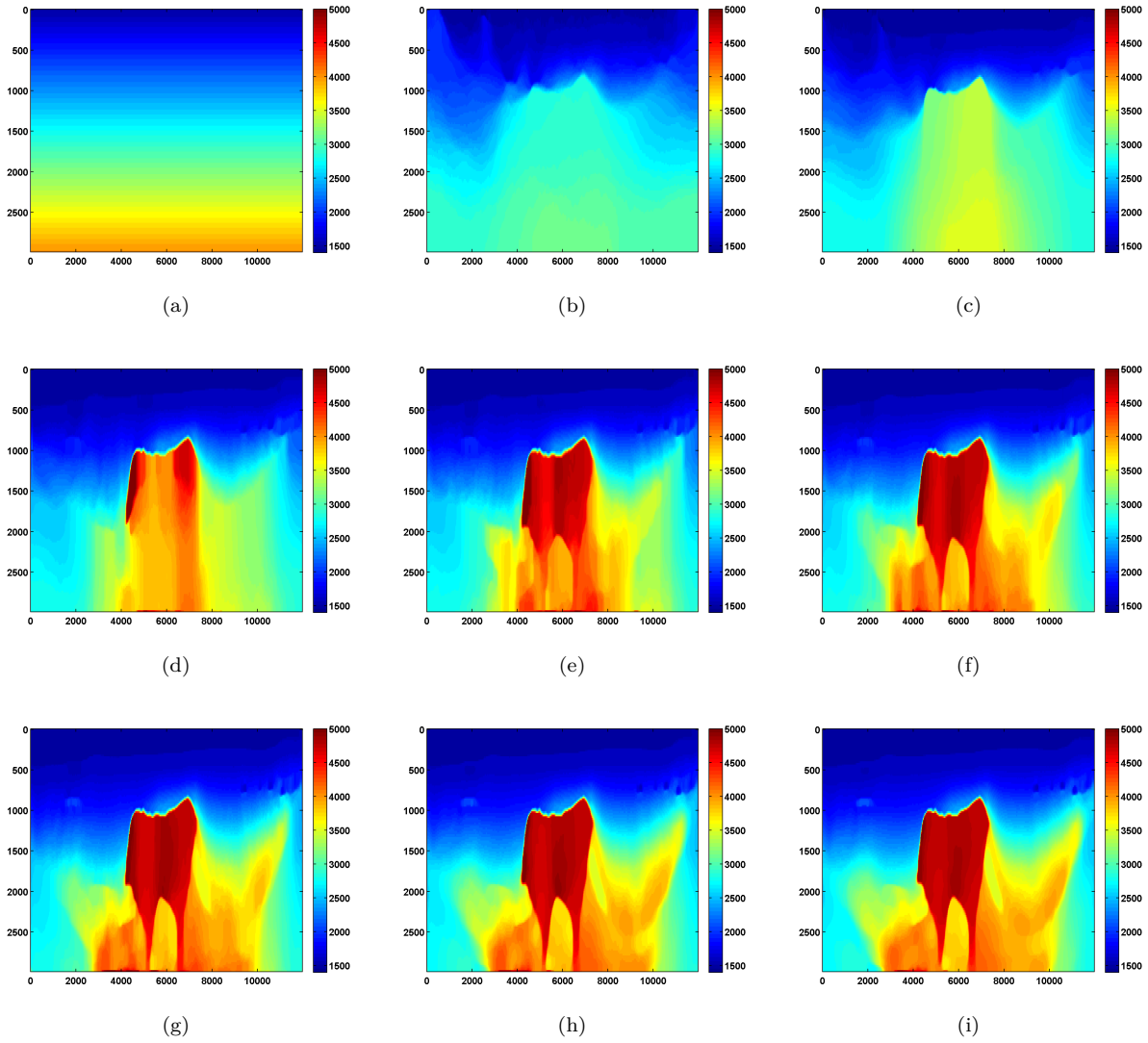
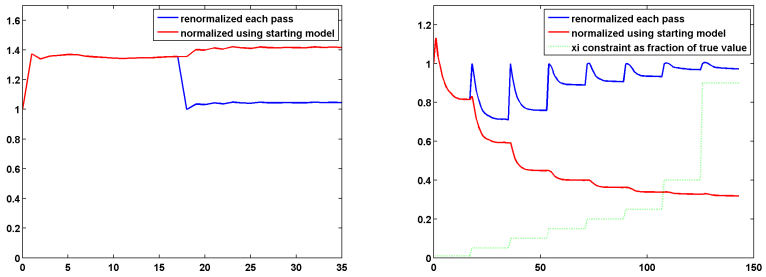


Figure 20: Initial velocity (a) and recovered velocity for the constrained adjoint state method with one-sided TV continuation corresponding to $\frac{\xi}{\xi_{\text{true}}} = .01$ (a), .05 (b), .10 (c), .15 (d), .20 (e), .25 (f), .40 (g), .90 (h).

TV constraint can for instance be straightforwardly replaced by an l_1 constraint on the curvelet coefficients of the model or combined with other convex constraints.

Synthetic experiments suggest that when there is noisy or limited data, TV regularization can improve the recovery by eliminating spurious artifacts. Experiments also show that a one-sided TV constraint that encourages velocity to increase with depth helps recover salt structures from poor initial models. In combination with a continuation strategy that gradually weakens the one-sided TV constraint, it can prevent the method from getting prematurely stuck at a bad solution when starting with a poor initialization.

In future work, we want to find a more practical way of selecting the regularization parameters τ and ξ as well as more principled continuation strategies for adjusting them during multiple passes through frequency batches. It would be interesting if a numerical framework along the lines of the SPOR-SPG method in [van den Berg and Friedlander, 2011] could be extended to this application. Another direction we want to



(a) (b)

Figure 21: Relative model error versus frequency batch without TV constraints (a) and with both TV and one-sided TV constraints (b) for the adjoint state examples.

pursue is to consider nonconvex sparsity penalties on the gradient of the model, penalizing for instance the ratio of the l_1 and l_2 norms of the gradient. We also intend to study more realistic numerical experiments and investigate how to better take advantage of the proposed constrained WRI framework.

11 Acknowledgements

Thanks to Lluis Guasch [ASK TO ADD AS AUTHOR] for suggesting the 2004 BP dataset to better show the merits of the TV constrained WRI method and for designing the numerical examples shown in Figures 7 and 8. Thanks to Bas Peters for insights about the formulation and implementation of the penalty method for full waveform inversion including the strategies for frequency continuation and computing multiple passes with warm starts. Thanks also to Polina Zheglova for helpful discussions about the boundary conditions.

References

Volkan Akcelik, George Biros, and Omar Ghattas. Parallel multiscale Gauss-Newton-Krylov methods for inverse wave propagation. *Supercomputing, ACM/IEEE ...*, 00(c):1–15, 2002. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1592877.

Aleksandr Y Aravkin and Tristan van Leeuwen. Estimating nuisance parameters in inverse problems. *Inverse Problems*, 28(11):115016, June 2012. URL http://arxiv.org/abs/1206.6532.

Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific; 2nd edition, 1999. ISBN 1886529000. URL http://www.amazon.com/Nonlinear-Programming-Dimitri-P-Bertsekas/dp/1886529000/ref=sr_1_2?ie=UTF8&qid=1395180760&sr=8-2&keywords=nonlinear+programming.

FJ Billette and S Brandsberg-Dahl. The 2004 BP velocity benchmark. *67th EAGE Conference & Exhibition*, (June):13–16, 2005. URL http://www.earthdoc.org/publication/publicationdetails/?publication=1404.

S Bonettini, R Zanella, and L Zanni. A scaled gradient projection method for constrained image deblurring. *Inverse Problems*, 25(1):015002, January 2009. ISSN 0266-5611. doi: 10.1088/0266-5611/25/1/015002. URL http://stacks.iop.org/0266-5611/25/i=1/a=015002?key=crossref.31896d5825318968df4855dada3c1552.

P Brucker. An O (n) algorithm for quadratic knapsack problems. *Operations Research Letters*, 3(3):163–166, 1984. URL http://www.sciencedirect.com/science/article/pii/0167637784900105.

Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 2011. URL http://link.springer.com/article/10.1007/s10851-010-0251-1.

Eric T. Chung, Tony F. Chan, and Xue-Cheng Tai. Electrical impedance tomography using level set representation and total variational regularization. *Journal of Computational Physics*, 205(1):357–372, May 2005. ISSN 00219991. doi: 10.1016/j.jcp.2004.11.022. URL http://linkinghub.elsevier.com/retrieve/pii/S0021999104004711.

Ernie Esser, Xiaoqun Zhang, and Tony F. Chan. A General Framework for a Class of First Order Primal-Dual Algorithms for Convex Optimization in Imaging Science. *SIAM Journal on Imaging Sciences*, 3(4):1015–1046, January 2010. ISSN 1936-4954. doi: 10.1137/09076934X. URL http://epubs.siam.org/doi/abs/10.1137/09076934X.

Ernie Esser, Yifei Lou, and Jack Xin. A Method for Finding Structured Sparse Solutions to Nonnegative Least Squares Problems with Applications. *SIAM Journal on Imaging Sciences*, 6(4):2010–2046, January 2013. ISSN 1936-4954. doi: 10.1137/13090540X. URL http://epubs.siam.org/doi/abs/10.1137/13090540X.

T Goldstein, E Esser, and R Baraniuk. Adaptive primal-dual hybrid gradient methods for saddle-point problems. *arXiv preprint arXiv:1305.0546*, 2013. URL http://arxiv.org/abs/1305.0546.

Z Guo and M.V. de Hoop. SHAPE OPTIMIZATION IN FULL WAVEFORM INVERSION WITH SPARSE BLOCKY MODEL REPRESENTATIONS. *gmig.math.purdue.edu*, 1:189–208, 2012. URL http://gmig.math.purdue.edu/pdfs/2012/12-12.pdf.

Bingsheng He and X Yuan. Convergence analysis of primal-dual algorithms for a saddle-point problem: from contraction perspective. *SIAM Journal on Imaging Sciences*, pages 1–35, 2012. URL http://epubs.siam.org/doi/abs/10.1137/100814494.

Felix J. Herrmann, Ian Hanlon, Rajiv Kumar, Tristan van Leeuwen, Xiang Li, Brendan Smithyman, Haneet Wason, Andrew J. Calvert, Mostafa Javanmehri, and Eric Takam Takougang. Frugal full-waveform inversion: From theory to a practical algorithm. *The Leading Edge*, 32(9):1082–1092, September 2013. ISSN 1070-485X. doi: 10.1190/tle32091082.1. URL http://library.seg.org/doi/abs/10.1190/tle32091082.1.

Musa Maharramov and Biondo Biondi. Robust joint full-waveform inversion of time-lapse seismic data sets with total-variation regularization. *arXiv preprint arXiv:1408.0645*, 2014. URL http://arxiv.org/abs/1408.0645.

B. Peters, F.J. Herrmann, and T. van Leeuwen. Wave-equation Based Inversion with the Penalty Method - Adjoint-state Versus Wavefield-reconstruction Inversion. 2014. doi: 10.3997/2214-4609.20140704. URL http://www.earthdoc.org/publication/publicationdetails/?publication=75576.

LI Rudin, S Osher, and E Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60:259–268, 1992. URL http://www.sciencedirect.com/science/article/pii/016727899290242F.

Mark Schmidt, Dongmin Kim, and S Sra. 11 Projected Newton-type Methods in Machine Learning. *Optimization for Machine Learning*, (1), 2012. URL http://books.google.com/books?hl=en&lr=&id=JPQx7s2L1A8C&oi=fnd&pg=PA305&dq=Projected+Newton-type+Methods+in+Machine+Learning&ots=vccfyhm5Jb&sig=1JHG-pEdkJv_Yur0SVJQ8j_bw6Ahttp://books.google.com/books?hl=en&lr=&id=JPQx7s2L1A8C&oi=fnd&pg=PA305&dq=11+Projected+Newton-type+Methods+in+Machine+Learning&ots=vccfyhm8Bg&sig=zz6vgxjlcGhjMDFgaDnf774W0xU.

MW Schmidt. Optimizing costly functions with simple constraints: A limited-memory projected quasi-newton algorithm. *International …*, 5:456–463, 2009. URL http://machinelearning.wustl.edu/mlpapers/paper_files/AISTATS09_SchmidtBFM.pdf.

Albert Tarantola. Inversion of seismic reflection data in the acoustic approximation. *Geophysics*, 49(8):1259–1266, 1984. URL http://link.springer.com/article/10.1007/s10107-012-0571-6http://arxiv.org/abs/1110.0895http://library.seg.org/doi/abs/10.1190/1.1441754.

Ewout van den Berg and Michael P. Friedlander. Sparse Optimization with Least-Squares Constraints. *SIAM Journal on Optimization*, 21(4):1201–1229, October 2011. ISSN 1052-6234. doi: 10.1137/100785028. URL http://epubs.siam.org/doi/abs/10.1137/100785028.

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

T. van Leeuwen and F. J. Herrmann. Mitigating local minima in full-waveform inversion by expanding the search space. *Geophysical Journal International*, 195(1):661–667, 2013b. URL http://gji.oxfordjournals.org/cgi/doi/10.1093/gji/ggt258.

Tristan van Leeuwen and Felix J. Herrmann. Fast waveform inversion without source-encoding. *Geophysical Prospecting*, 61(2010):10–19, June 2013c. ISSN 00168025. doi: 10.1111/j.1365-2478.2012.01096.x. URL http://doi.wiley.com/10.1111/j.1365-2478.2012.01096.x.

J. Virieux and S. Operto. An overview of full-waveform inversion in exploration geophysics. *Geophysics*, 74(6):WCC1–WCC26, November 2009. ISSN 0016-8033. doi: 10.1190/1.3238367. URL http://library.seg.org/doi/abs/10.1190/1.3238367.

Xiaoqun Zhang, Martin Burger, and Stanley Osher. A Unified Primal-Dual Algorithm Framework Based on Bregman Iteration. *Journal of Scientific Computing*, 46(1):20–46, August 2010. ISSN 0885-7474. doi: 10.1007/s10915-010-9408-8. URL http://link.springer.com/10.1007/s10915-010-9408-8.

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