Incorporating optimal transport in Tomographic Full Waveform Inversion : Theory

Rahul Sarkar and Biondo Biondi

ABSTRACT

In recent years, a great deal of work has gone into devising optimization strategies that are more robust to cycle skipping as compared to the conventional Full Waveform Inversion (FWI) objective function. Tomographic full waveform inversion (TFWI) is one such technique that involves the use of a non-physical extension of the whole velocity model and an extended modeling operator that is better capable of modeling the observed data. Inversion algorithms based on this concept have been shown to converge to reasonable models on field data, but with slow convergence rate. Another interesting direction of research that has emerged in this field is based on the use of optimal transport objective functions. These formulations have also been shown to be capable of converging to correct models from relatively inaccurate starting models, leveraging the fact that objective functions based on optimal transport distances are much less susceptible to local minima compared to their L^2 counterparts. In this report, we investigate an inversion strategy based on the idea of velocity model extension that attempts to leverage optimal transport distances to overcome the challenges of TFWI convergence. We derive the problem formulation and present some preliminary results in 2D for the acoustic case.

INTRODUCTION

The non-convexity of the acoustic FWI problem leads to gradient-descent based inversion schemes based getting stuck in local minima, unless one starts very close to the true velocity model. Even in the 2D case, the high dimensionality of the problem precludes brute force stochastic optimization techniques for typical model sizes of interest in the seismic industry. Several strategies, which can be broadly grouped into three categories, have been proposed and experimented with over the past few decades to overcome these problems.

The first category consists of frequency continuation (Pratt, 1999) and multiscale (Bunks et al., 1995) methods where the FWI inversion is performed in overlapping frequency bands, starting from low frequencies and progressively moving to higher frequencies. These schemes are motivated by the fact that for any velocity model, the modeled data and observed data will stop cycle skipping at sufficiently low frequencies. This type of approach works nicely if there is sufficient low frequencies in the data with high enough signal to noise ratio. However, due to limitations imposed by the physics of seismic sources, this requirement is quite often not fulfilled and valuable low frequency content is missing in the observed data.

The second category involves constructing objective functions which have a larger basin of attraction around the global minima, and are thus more "convex" with respect to the velocity model. Objective functions based on cross-correlation (Luo and Schuster, 1991), dynamic-warping (Ma and Hale, 2013) and deconvolution (Luo and Sava, 2011) are examples that fall in this class of methods. However, the inverse problem of estimating model parameters from surface measurements is intrinsically non-convex because of the oscillatory nature of the source wavelet, and it may not be possible to design an objective function that is convex over the whole feasible set of velocities. However, these methods demonstrate that it may be possible to achieve some form of convexity over a subset of the feasible set that is still larger than that with conventional FWI. Another interesting development that belongs to this class of finding increasingly convex objective functions comes from the use of optimal transport based misfit functions (Engquist and Froese, 2014; Métivier et al., 2016; Yang et al., 2016). These misfit functions are more commonly known as Wasserstein distance functions, and they have been reported to be much more robust at avoiding local minima and also have better convergence properties.

The third category of methods involve extending the reflectivity model, such as extending subsurface offset and/or time lag, in the migrated image domain. The idea uses an extended imaging condition to form migrated images along the extended dimension and then attempts to focus the energy at zero along the extended axes with a misfit function. This is the concept behind approaches such as differential semblance optimization (Symes and Carazzone, 1991; Shen and Symes, 2008) which tries to focus the image at zero subsurface offset, or an equivalent formulation that attempts to flatten the image gathers in scattering angle (Sava and Biondi, 2004a,b; Zhang and Biondi, 2013). More recently, an interesting alternative has been proposed that involves extension of the whole velocity model (Symes, 2008; Biondi and Almomin, 2014) and is named tomographic full waveform inversion (TFWI). However, the speed of convergence of these methods can be quite slow for large scale problems (Almomin and Biondi, 2014).

In this paper, we develop the theoretical formulation of a new optimization scheme based on the model extension approach but one that attempts to overcome some of the convergence problems of TFWI, by leveraging the properties of the optimal transport objective function. The theory is developed for the 2D case, and its generalization to 3D is straightforward but computationally challenging. The new algorithm is motivated by the original TFWI algorithm proposed by Biondi and Almomin (2014) and follows it closely, but has some important differences in the way it attempts to focus the extended model at the origin of the extended model axis. Along the way, we illustrate the concept with some motivating examples and preliminary numerical results.

BACKGROUND

We first provide a concise introduction to the extended modeling operator in 2D for acoustic wave propagation that defines the forward problem. We will introduce some of the intuition behind such a formulation. Next, we briefly review Kolmogorov spectral factorization that is used to find the minimum phase equivalent of signal. We finish this section with a basic introduction to optimal transportation theory, with emphasis on the quadratic Wasserstein distance.

The acoustic wave equation

Let us represent the 2D spatial variables by $\mathbf{x} = (x, z)$ and time by t. We consider the acoustic wave equation in a constant density medium, written in the following form:

$$u_{tt}(\mathbf{x}, t) - v^2(\mathbf{x}) \nabla^2 u(\mathbf{x}, t) = f(t) \delta(\mathbf{x} - \mathbf{x}_s) \;, t \geq 0 \;. \tag{1}$$

Here $u(\mathbf{x}, t)$ denotes the pressure at position $\mathbf{x}$ and time t, and ∇^2 denotes the 2D Laplacian operator $(\partial_x^2 + \partial_z^2)$. The source term $f(t)\delta(\mathbf{x} - \mathbf{x}_s)$ denotes a point source located at position $\mathbf{x}_s$ in space with a time dependent source function $f(t)$. The velocity of the medium is heterogenous and is denoted by $v(\mathbf{x})$. To compress notation going forward, we will drop the explicit dependence of the quantities appearing in equation (1) on space and time variables, and simply use u to denote $u(\mathbf{x}, t)$, m to denote the square of the velocity $v^2(\mathbf{x})$, and f to denote the source function $f(t)\delta(\mathbf{x} - \mathbf{x}_s)$. The dependence on space and time variables will be understood from context, and explicit dependence on these variables will only be given in cases deemed absolutely necessary. We will also refer to the quantity m as the "acoustic model parameter".

The solution u to equation (1) is given by the convolution of the source function with the Green's function (the solution to equation (1) when $f(t) = \delta(t)$), and is linear with respect to f. However the Green's function is not needed to understand what follows, and so we will simplify notation and introduce the solution operator $\mathcal{L}(m)$ that can be used to write u as indicated below:

$$u = \mathcal{L}(m) f \;. \tag{2}$$

In (2) we have abused notation slightly and written $\mathcal{L}(m)f$ to denote $\mathcal{L}(m)(f)$, where the first bracket denotes the fact that the solution operator depends on m, while the second bracket denotes the argument on which the solution operator $\mathcal{L}(m)$ acts on, namely the source function. We will be using both these notations interchangeably depending upon context. For a fixed value of m, $\mathcal{L}(m)$ is a linear operator on the space of source functions.

The Born linearization operator

A natural question that arises in the study of hyperbolic systems such as equation (1), is how the solution u changes upon perturbations in the acoustic model m for a fixed source function f. Under some assumptions about convergence, this question is conveniently answered using perturbation theory. We perturb the acoustic model by a small amount δm, parameterized by a parameter ϵ, so that the perturbed model is given by $m' = m + \epsilon \delta m$. We now seek a new solution u' of the equation $u'_{tt} - m' \nabla^2 u' = f$ in the power series expansion in ϵ

$$u' = u + \sum_{n=1}^{\infty} \epsilon^n \delta u_n \ , \tag{3}$$

assuming such a series exists and converges uniformly to u'.

Under the stated assumptions of existence and convergence, this series is known as the Born scattering series. Substituting the expressions for u' and m' into the wave equation:

$$(u + \sum_{n=1}^{\infty} \epsilon^n \delta u_n)_{tt} - (m + \epsilon \delta m) \nabla^2 (u + \sum_{n=1}^{\infty} \epsilon^n \delta u_n) = f \ . \tag{4}$$

Next, following perturbation theory, we equate similar order terms in powers of ϵ appearing in equation (4) and finally obtain the system of equations

$$\begin{aligned} Zero^{th}\ order: &\quad u_{tt} - m\nabla^2 u = f \ , \\ 1^{st}\ order: &\quad (\delta u_1)_{tt} - m \nabla^2 (\delta u_1) = \delta m \nabla^2 u \ , \text{ and} \\ n^{th}\ order: &\quad (\delta u_n)_{tt} - m \nabla^2 (\delta u_n) = \delta m \nabla^2 (\delta u_{n-1}) \ , \ n \geq 2 \ . \end{aligned} \tag{5}$$

The $Zero^{th}$ order equation in (5) is the same as equation (1), and represents the acoustic wave equation satisfied by the unperturbed solution u, which is also called the zero order wavefield. The 1^{st} order and all subsequent higher order equations give us a recursive formula to compute the n^{th} order wavefield using the $(n-1)^{th}$ order wavefield. Using the solution operator $\mathcal{L}(m)$ introduced in equation (2), we can symbolically write the solution to the system of equations in (5) as

$$\begin{aligned} Zero^{th}\ order: &\quad u = \mathcal{L}(m) f \ , \\ 1^{st}\ order: &\quad \delta u_1 = \mathcal{L}(m)(\delta m \nabla^2 u) \ , \text{ and} \\ n^{th}\ order: &\quad \delta u_n = \mathcal{L}(m)(\delta m \nabla^2 \delta u_{n-1}) \ , \ n \geq 2 \ . \end{aligned} \tag{6}$$

It is also clear that the solution operator allows us to write the perturbed solution directly as $u' = \mathcal{L}(m + \epsilon \delta m) f$. Hence, equating the Born series representation of the perturbed wavefield with $\mathcal{L}(m + \epsilon \delta m) f$ and using the relations derived in equation

(6) we obtain:

$$\mathcal{L}(m+\epsilon\delta m)f-\mathcal{L}(m)f=\epsilon\mathcal{L}(m)(\delta m\nabla^2 u)+\sum_{n=2}^{\infty}\epsilon^n\mathcal{L}(m)(\delta m\nabla^2\delta u_{n-1})$$

which taking limit $\epsilon\to 0$ becomes, $\mathcal{L}(m+\epsilon\delta m)f-\mathcal{L}(m)f=\epsilon\mathcal{L}(m)(\delta m\nabla^2 u)+\mathcal{O}(\epsilon^2)$. (7)

Equation (7) says that to first order the perturbation in the wavefield, which is the difference between the perturbed wavefield and zero order wavefield is linear in δm. Moreover, the Frechet derivative of the solution u with respect to m also exists as the following holds:

$$\lim_{\epsilon\to 0}\frac{||\mathcal{L}(m+\epsilon\delta m)f-\mathcal{L}(m)f-\epsilon\mathcal{L}(m)(\delta m\nabla^2 u)||}{||\epsilon\delta m||}=0 \tag{8}$$

with any suitably defined norm on the respective spaces.

The existence of the Frechet derivative gives us the *Born linearization operator* $L(m)$ that, for fixed m and f linearly maps small perturbations in the model parameter to perturbations in the wavefield. For this operator we assume $||\epsilon\delta m||$ is small, drop quadratic and higher order terms in ϵ in equation (7), and finally absorb ϵ into the definition of δm to obtain:

$$\delta u=u'-u=\mathcal{L}(m+\delta m)f-\mathcal{L}(m)f=\mathcal{L}(m)(\nabla^2 u)(\delta m)=L(m)(\delta m)\ . \tag{9}$$

In the above equation, δu denotes the change in the wavefield, and $L(m)$ is the composition of two operators $\mathcal{L}(m)\circ\nabla^2 u$, where $\mathcal{L}(m)$ is the solution operator introduced earlier and $\nabla^2 u$ denotes another linear operator that produces a secondary source by multiplying the physical quantity $\nabla^2 u$ point-wise in space with the model perturbation δm for each time instance t. Also, the notation $L(m)$ signifies the dependence of the Born linearization operator on m which is the point in the model space around which linearization is being carried out.

The extended modeling operator

We are now in a position to introduce the extended model and the extended modeling operator. In seismic inversion problems, most of the time the starting models are simple models with only the low wavenumber components in them. In this case, the higher order terms in equation (7) are significant and cannot be neglected. This condition is also typically the case when we have cycle skipping between the observed data and the modeled data due to the presence of large time shifts, and in this regime the Born linearization operator is inadequate to model the perturbation in the wavefield δu. However, it is conceivable that in the latter scenario, it may possible to model the perturbation δu by creating a superposition of the secondary source wavefield $\nabla^2 u(\delta m)$ at different time delays or *"lags"*, and then using the solution operator $\mathcal{L}(m)$ to model the data using this superimposed secondary source. This idea is the intuition behind the extended Born modeling operator introduced in the context of TFWI by Biondi and Almomin (2014).

In what follows, we will use a tilde notation to differentiate the extended model and the extended operators from their non-extended counterparts. We start by defining the extended model perturbation which is a model perturbation that depends on a time lag parameter τ, and denote it $\delta\widetilde{m}(\tau)$. For conciseness, we will drop the explicit dependence on τ, and write it as simply $\delta\widetilde{m}$. It is convenient to think of $\delta\widetilde{m}$ as a model perturbation that depends on "time" where time is allowed to be positive or negative. The extended modeling operator $\widetilde{\mathcal{L}}(m)$ is then defined as follows:

$$\begin{aligned} \widetilde{\mathcal{L}}(m)(\delta\widetilde{m}, f) &= \mathcal{L}(m)f + \widetilde{L}(m)(\delta\widetilde{m}) = u + \widetilde{L}(m)(\delta\widetilde{m}) \ , \\ \text{where, } \widetilde{L}(m)(\delta\widetilde{m}) &= \mathcal{L}(m)(\nabla^2(\mathcal{L}(m)f) \star_\tau \delta\widetilde{m}) = \mathcal{L}(m)(\nabla^2 u \star_\tau \delta\widetilde{m}) \ , \\ \text{and, } (\nabla^2 u \star_\tau \delta\widetilde{m})(t) &= \int_\tau \nabla^2 u(\tau + t)\delta\widetilde{m}(\tau) d\tau \ . \end{aligned} \tag{10}$$

The operator $\widetilde{\mathcal{L}}(m)$ has two arguments $\delta\widetilde{m}$ and f, and is linear in the argument f. The output consists of the sum of the modeled wavefield u using the non-extended solution operator $\mathcal{L}(m)$, and the extended Born operator $\widetilde{L}(m)$ which is linear in the argument $\delta\widetilde{m}$. Notice that when $\delta\widetilde{m} = \delta(\tau)\delta m$, then $(\nabla^2 u \star_\tau \delta\widetilde{m})(t) = \nabla^2 u(\delta m)$, and then the extended Born operator reduces to the Born operator as $\widetilde{L}(m)(\delta\widetilde{m}) = L(m)(\delta m)$.

Kolmogorov spectral factorization

We now proceed to provide a very brief introduction to the task of estimating a minimum phase signal from a given spectrum. In effect, this process can also be used to construct the minimum phase equivalent of any given signal, by first computing its amplitude spectrum. We will limit ourselves to continuous real valued signals in time, and refer the reader to excellent references (Kolmogorov, 1939; Robinson and Treitel, 1980; Claerbout, 1985) for an in-depth treatment of this topic. We will also assume that the signal belongs to the Schwarz class of functions, which implies that it is *"nicely behaved"*, and we can Fourier transform it without needing to worry about existence, stability and convergence issues.

Let us first define a few quantities like the Fourier transform and its inverse, and state some important properties that hold in special cases, which are the ones we will need. Let $h(t)$ be a real valued signal in time. Then the Fourier transform $\hat{h}(\omega)$ is defined as:

$$\hat{h}(\omega) = \int_{\mathbb{R}} h(t) e^{-2\pi i \omega t} dt \ . \tag{11}$$

By the Fourier inversion formula, we can recover the function $h(t)$ from $\hat{h}(\omega)$ as follows:

$$h(t) = \int_{\mathbb{R}} \hat{h}(\omega) e^{2\pi i \omega t} d\omega \ . \tag{12}$$

A very special case occurs when $h(t)$ is an even function of t, i.e, $h(t) = h(-t)$. In this case, the Fourier transform $\hat{h}(\omega)$ is also an even function of ω, and so we have $\hat{h}(\omega) = \hat{h}(-\omega)$. The Fourier formulas in this case become:

$$\begin{aligned} \hat{h}(\omega) &= 2\int_{\mathbb{R}^+} h(t)\cos(2\pi\omega t) dt \ \text{, and} \\ h(t) &= 2\int_{\mathbb{R}^+} \hat{h}(\omega)\cos(2\pi\omega t) d\omega \ . \end{aligned} \tag{13}$$

The quantity $|\hat{h}(\omega)|$ is called the amplitude spectrum. Let us denote it by $A(\omega)$ and note that it satisfies $A(\omega) = \sqrt{\hat{h}(\omega)\hat{h}^*(\omega)}$, where $\hat{h}^*(\omega)$ is the complex conjugate of $\hat{h}(\omega)$. In the case that $h(t)$ is real, it follows from equation (11) that $A(\omega)$ is an even function of ω. This is a key property that is crucial in deriving the minimum phase signal that has the same amplitude spectrum $A(\omega)$. We also have by definition $A(\omega) \geq 0$. In addition, let us assume that the amplitude spectrum is strictly positive, i.e, $A(\omega) > 0$.

Under these assumptions, Kolmogorov spectral factorization performs the following steps in sequence:
(i) Take log of the amplitude spectrum $A(\omega)$ to get $B(\omega)$ which is even
$B(\omega) = \log(A(\omega))$.
(ii) Inverse Fourier transform $B(\omega)$ to time, and note that $B(t)$ is also even
$B(t) = \int_{\mathbb{R}} B(\omega) e^{2\pi i\omega t} d\omega = 2\int_{\mathbb{R}^+} B(\omega)\cos(2\pi\omega t) d\omega$.
(iii) Create a causal signal from $B(t)$
$B_c(t) = 2B(t)H(t)$, where $H(t)$ is the Heaviside step function.
(iv) Fourier transform $B_c(t)$ to frequency domain
$B_c(\omega) = \int_{\mathbb{R}} B_c(t) e^{-2\pi i\omega t} dt = 2\int_{\mathbb{R}} B(t)H(t) e^{-2\pi i\omega t} dt = 2\int_{\mathbb{R}^+} B(t) e^{-2\pi i\omega t} dt$.
(v) Restore amplitude spectrum by exponentiation, and get Fourier transform of minimum phase signal
$\widetilde{h}(\omega) = e^{B_c(\omega)}$.
(vi) Recover the minimum phase signal by inverse Fourier transform
$\widetilde{h}(t) = \int_{\mathbb{R}} \widetilde{h}(\omega) e^{2\pi i\omega t} d\omega$.

Note that steps (i), (iv) and (v) use equation (13) to ensure that the minimum phase signal has the same amplitude spectrum as $A(\omega)$, as shown below:

$$\begin{aligned} B_c(\omega) &= 2\int_{\mathbb{R}^+} B(t) e^{-2\pi i\omega t} dt = B(\omega) - 2i\int_{\mathbb{R}^+} B(t)\sin(2\pi\omega t) dt \ , \\ \implies |\widetilde{h}(\omega)| &= |e^{B_c(\omega)}| = e^{B(\omega)} = A(\omega) \ . \end{aligned} \tag{14}$$

It can be easily seen that $\widetilde{h}(\omega) \neq 0$, and hence $1/\widetilde{h}(\omega) = e^{-B_c(\omega)} \neq 0$. The signal obtained this way is front loaded, i.e., has most of the energy concentrated at the beginning of the signal. In certain cases, if the amplitude spectrum has zeros then the above algorithm can still be applied with a small thresholding parameter $\epsilon > 0$ which is added to $A(\omega)$ before performing step (i).

In the discrete case, the above algorithm goes through verbatim with the only exception that step (iii) is modified to assign the $t = 0$ sample a weight of 1, when forming the causal signal $B_c(t)$. Also the discrete Fourier transform and discrete inverse Fourier transform are used in the above calculations. The thresholding step is usually mandatory in the discrete case to achieve numerical stability. We will be using the discrete equivalent in the new inversion problem to be defined later.

Optimal transport

We finish this section by introducing the concept of optimal transport and the quadratic Wasserstein distance in 1D, which play a critical role in the new inverse problem formulation. We also modify the quadratic Wasserstein distance definition for non-positive signals that do not obey the *"mass conservation"* requirement. These concepts are introduced and explained next.

We follow a minimalist approach here and completely avoid the language of measure theory in this brief introduction to optimal transport. However, this approach keeps things simple and is more than sufficient for our case as we only deal with 1D optimal transport problem. The mathematically inclined reader is directed to excellent texts on this topic in the references (Villani, 2003, 2008; Engquist et al., 2016; Yang et al., 2016).

Consider two functions $v(x)$ and $w(x)$ defined on $\mathbb{R}$, and a positive integer N, which satisfy the properties:

$$\begin{aligned} &(i) \;\; v(x) \geq 0 \text{ and } w(x) \geq 0, \; \forall \, x \in \mathbb{R} \, , && \text{(non-negativity)} \\ &(ii) \;\; \int_{\mathbb{R}} v(x) dx = \int_{\mathbb{R}} w(x) dx \neq 0 \, , \text{ and} && \text{(mass conservation)} \\ &(iii) \;\; \int_{\mathbb{R}} |x|^p v(x) dx < \infty, \text{ and } \int_{\mathbb{R}} |x|^p w(x) dx < \infty, \; \forall \; 0 \leq p \leq N \; .\text{(finite moments)}. \end{aligned} \tag{15}$$

In optimal transport literature, the mass conservation condition is stated to read $\int_{\mathbb{R}} v(x)dx = \int_{\mathbb{R}} w(x)dx = 1$, as it is always possible to do so by rescaling the functions $v(x)$ and $w(x)$, and this is the convention that we will also adopt. In fact, when this is true, these functions can be interpreted as probability measures on $\mathbb{R}$. As we will only be interested in the quadratic Wasserstein distance, it will also suffice for us to require $N = 2$.

The optimal transport problem is concerned with finding a transport plan to rearrange v to w given a cost function $c(x,y)$, such that the transportation cost is minimum. The cost function can be thought of as a penalty to move one unit of mass from position x to position y. For some rearrangement strategy T, called the transport plan, the transportation cost is given by:

$$\text{Cost}(T) = \int_{\mathbb{R}} c(x, T(x)) v(x) dx \ . \tag{16}$$

We can now define the *quadratic Wasserstein distance* between the functions v and w (Villani, 2003), using the quadratic cost function $|x-y|^2$ (hence the name quadratic Wasserstein distance):

$$W_2(v,w) = \left(\inf_{T \in \mathcal{M}} \int_{\mathbb{R}} |x - T(x)|^2 v(x) dx \right)^{1/2} , \tag{17}$$

where $\mathcal{M}$ is the set of all possible maps that rearrange v to w. It can be shown that W_2 satisfies all the conditions of a metric, and hence is justifiably referred to as a "distance" function.

In 1D, the minimization problem in equation (17) can be solved exactly (see Villani (2003)). Moreover, this can also be done very quickly on a computer. This is a key feature that motivated us to consider W_2 in our work in this paper. The solution can be given in terms of the cumulative distribution functions of v and w, namely $V(x)$ and $W(x)$, which are defined below:

$$V(x) = \int_{-\infty}^{x} v(x) dx \ \ , \ \ W(x) = \int_{-\infty}^{x} w(x) dx \ . \tag{18}$$

With these quantities defined, the optimal transportation cost can be evaluated as follows (see Villani (2003) for details):

$$W_2^2(v,w) = \int_0^1 |V^{-1}(\gamma) - W^{-1}(\gamma)|^2 d\gamma = \int_{\mathbb{R}} |x - W^{-1}(V(x))|^2 v(x) dx \ . \tag{19}$$

Any of the formulas in equation (19) can be used for the computation of $W_2(v,w)$. It is also easy to see that the optimal transportation plan is given by the composite map $W^{-1} \circ V$. However, we will be trying to apply the quadratic Wasserstein distance to seismic inversion problems, where neither the positivity condition nor the mass conservation condition is satisfied. To get around this problem, we adopt the normalization technique introduced by Yang et al. (2016). The first step involves choosing a parameter $\delta \geq 0$, such that both the functions $v + \delta \geq 0$ and $w + \delta \geq 0$. An obvious such choice is $\delta = -\min(0, \min_{x \in \mathbb{R}}(v(x), w(x)))$. Next, once we have satisfied the non-negativity condition, we can scale both functions $v + \delta$ and $w + \delta$ to have area 1. We will refer to these normalized functions as $\widetilde{v}$ and $\widetilde{w}$. The W_2 metric can now be applied to these normalized functions. However, these operations in general no longer give rise to a metric with respect to the original functions v and w. For this reason, we will still refer to it as the quadratic Wasserstein distance (with abuse of the term "distance"), but to distinguish it from a true distance function we will denote it by $\widetilde{W_2}(v, w)$. Thus we have:

$$\widetilde{W_2}(v, w) = W_2(\widetilde{v}, \widetilde{w}) \; . \tag{20}$$

THE INVERSE PROBLEM

We now define the inverse problem mathematically and also explain the important differences between standard TFWI and our new formulation based on optimal transport. It is best to convert the problem to the discrete setting, as this operation is a necessary step in solving the problem on a computer. All parameters and operators will be represented in boldface to emphasize the fact they are defined on finite dimensional vector spaces after the discretization. Also, we will represent the field quantities like wavefield, source function and all model parameters (extended and non-extended) as vectors using the canonical basis in their respective dimensions, and similarly the linear operators will become matrices represented with respect to the canonical basis of their arguments.

The discrete setting

We discretize space and time using $\Delta x, \Delta z,$ and Δt corresponding to the x, y and z dimensions respectively. We also assume that the number of grid points along these dimensions are respectively $N_x, N_z,$ and N_t. For the time lag axis τ, we discretize it using the same interval for time t, and thus $\Delta \tau = \Delta t$. To cover all possible time lags we will choose the number of grid points along τ as $N_\tau = 2N_t - 1$, with the N_t sample corresponding to $\tau = 0$. The model and its perturbation will be represented by $\mathbf{m}$ and $\boldsymbol{\delta}\mathbf{m}$, and belong to a vector space of dimension $N_x N_z$. The extended model perturbation will be denoted by $\boldsymbol{\delta}\widetilde{\mathbf{m}}$ that belongs to a vector space of dimension $N_x N_z N_\tau$. Source functions, wavefields and wavefield perturbations will be represented by $\mathbf{f}$, $\mathbf{u}$ and $\boldsymbol{\delta}\mathbf{u}$ respectively, and belong to a vector space of dimension $N_x N_z N_t$. The i^{th} order wavefield in the Born scattering series will be denoted by $\boldsymbol{\delta}\mathbf{u}_i$ and live in the same vector space as $\mathbf{u}$. All the operators are defined similarly in boldface, and we write their modeling equations as

$$\begin{aligned}
&\mathbf{u} = \boldsymbol{\mathcal{L}}(\mathbf{m})\mathbf{f} \; , \\
&\boldsymbol{\delta}\mathbf{u} = \boldsymbol{\mathcal{L}}(\mathbf{m})(\boldsymbol{\nabla}^2\mathbf{u})(\boldsymbol{\delta}\mathbf{m}) = \mathbf{L}(\mathbf{m})(\boldsymbol{\delta}\mathbf{m}) \; , \text{ and} \\
&\widetilde{\boldsymbol{\mathcal{L}}}(\mathbf{m})(\boldsymbol{\delta}\widetilde{\mathbf{m}}, \mathbf{f}) = \boldsymbol{\mathcal{L}}(\mathbf{m})\mathbf{f} + \boldsymbol{\mathcal{L}}(\mathbf{m})(\boldsymbol{\nabla}^2\mathbf{u} \star_\tau \boldsymbol{\delta}\widetilde{\mathbf{m}}) = \boldsymbol{\mathcal{L}}(\mathbf{m})\mathbf{f} + \widetilde{\mathbf{L}}(\mathbf{m})(\boldsymbol{\delta}\widetilde{\mathbf{m}}) \; .
\end{aligned} \tag{21}$$

The operators $\boldsymbol{\mathcal{L}}(\mathbf{m})$, $\mathbf{L}(\mathbf{m})$ and $\widetilde{\mathbf{L}}(\mathbf{m})$ are linear operators and thus can be represented as matrices with sizes easily understood from their range and domain. The elements of the matrices depend on the model $\mathbf{m}$. We recognize that they also depend on

$\mathbf{f}$ but we will develop the inverse formulation for a fixed shot in the seismic survey and hence assume $\mathbf{f}$ to be fixed. For the sake of brevity, we will drop the explicit dependence of these matrices on $\mathbf{m}$ and simply write $\boldsymbol{\mathcal{L}}$, $\mathbf{L}$ and $\widetilde{\mathbf{L}}$ to refer to them respectively. We will denote their adjoints by $\boldsymbol{\mathcal{L}}^\dagger$, $\mathbf{L}^\dagger$ and $\widetilde{\mathbf{L}}^\dagger$ which we will need in the next section. The adjoints are nothing but the transposes of the matrices for each operator. We will also be using a few other adjoints which we will introduce later.

FWI and TFWI inverse problems

We can now formulate the inverse problem. For a seismic survey the objective function that we will develop is *additive* in each shot, hence all derived quantities like gradients and Hessians are also additive in the shots. It thus suffices for us to develop an inverse problem formulation for a single seismic shot.

Let the source be located at the point $\mathbf{x}_s = (x_s, z_s)$, and assume we have N_r receivers for the seismic shot records each with its own location $\mathbf{x}_r = (x_r, z_r)$. We denote the recorded data at all the receivers by $\mathbf{d}_r$, which is a vector with dimension $N_r N_t$. We also introduce the restriction operator $\mathbf{R}$ (also a matrix as it is a linear operator) which samples quantities such as the wavefield $\mathbf{u}$ at the receiver locations. We define the true model by $\mathbf{m}_{true}$, and thus we have $\mathbf{d}_r = \mathbf{R}\boldsymbol{\mathcal{L}}(\mathbf{m}_{true})\mathbf{f}$.

For reference, the standard FWI objective function $J_{FWI}(\mathbf{m})$ and the expression for its gradient $\frac{\partial J_{FWI}(\mathbf{m})}{\partial \mathbf{m}}$ are:

$$\begin{aligned} J_{FWI}(\mathbf{m}) &= \frac{1}{2}||\mathbf{R}\boldsymbol{\mathcal{L}}\mathbf{f} - \mathbf{d}_r||_2^2 \ , \ \text{and} \\ \frac{\partial J_{FWI}(\mathbf{m})}{\partial \mathbf{m}} &= \mathbf{L}^\dagger \mathbf{R}^\dagger (\mathbf{R}\boldsymbol{\mathcal{L}}\mathbf{f} - \mathbf{d}_r) \ , \ \text{noting that } \frac{\partial(\boldsymbol{\mathcal{L}}\mathbf{f})}{\partial \mathbf{m}} = \mathbf{L} \ . \end{aligned} \tag{22}$$

FWI attempts to recover $\mathbf{m}_{true}$ by minimizing J_{FWI} starting from some initial model $\mathbf{m}_0$. However, as is well known, if $||\mathbf{m}_{true} - \mathbf{m}_0||$ is too large, then FWI gets trapped in local minima.

Now, we examine the TFWI objective function introduced in Biondi and Almomin (2014). Their objective function is

$$J_{TFWI}(\mathbf{m}, \boldsymbol{\delta}\widetilde{\mathbf{m}}) = \frac{1}{2}||\mathbf{R}\widetilde{\boldsymbol{\mathcal{L}}}(\boldsymbol{\delta}\widetilde{\mathbf{m}}, \mathbf{f}) - \mathbf{d}_r||_2^2 + \epsilon|| \, |\tau| \boldsymbol{\delta}\widetilde{\mathbf{m}}||_2^2 \ . \tag{23}$$

The data-fitting term in the TFWI objective function $||\mathbf{R}\widetilde{\boldsymbol{\mathcal{L}}}(\boldsymbol{\delta}\widetilde{\mathbf{m}}, \mathbf{f}) - \mathbf{d}_r||$ uses the extended modeling operator to match the recorded data. This is possible in part due to the introduction of additional variables into the problem in the form of the extended model $\boldsymbol{\delta}\widetilde{\mathbf{m}}$. The regularization term in the objective function $|| \, |\tau| \boldsymbol{\delta}\widetilde{\mathbf{m}}||$ penalizes components of the extended model at $\tau \neq 0$. The goal of this term is to eventually drive the extended model to the approximate form $\boldsymbol{\delta}\widetilde{\mathbf{m}} \approx \delta(\tau)\boldsymbol{\delta}\mathbf{m}$. When this happens, i.e, both the data fitting and regularization terms are small simultaneously for some $\mathbf{m}$

and $\boldsymbol{\delta}\mathbf{m}$, it means we are in a regime where Born linearization is accurate and we thus have $J_{FWI}(\mathbf{m}+\boldsymbol{\delta}\mathbf{m}) \to 0$. The iterative inversion is carried out in both variables $\mathbf{m}$ and $\boldsymbol{\delta}\widetilde{\mathbf{m}}$ with scale mixing.

A new inverse problem formulation

Suppose we have a model $\mathbf{m}_0$ and we want to calculate $\boldsymbol{\delta}\widetilde{\mathbf{m}}$ that minimizes the data fitting term of TFWI $\frac{1}{2}||\mathbf{R}\widetilde{\boldsymbol{\mathcal{L}}}(\boldsymbol{\delta}\widetilde{\mathbf{m}},\mathbf{f})-\mathbf{d}_r||_2^2$. Let us start from $\boldsymbol{\delta}\widetilde{\mathbf{m}}=\mathbf{0}$, and calculate the gradient of the data fitting term with respect to $\boldsymbol{\delta}\widetilde{\mathbf{m}}$ by

$$\begin{aligned} &\frac{\partial}{\partial(\boldsymbol{\delta}\widetilde{\mathbf{m}})}\left(\frac{1}{2}||\mathbf{R}\widetilde{\boldsymbol{\mathcal{L}}}(\boldsymbol{\delta}\widetilde{\mathbf{m}},\mathbf{f})-\mathbf{d}_r||_2^2\right)=\widetilde{\mathbf{L}}^\dagger\mathbf{R}^\dagger(\mathbf{R}\widetilde{\boldsymbol{\mathcal{L}}}(\boldsymbol{\delta}\widetilde{\mathbf{m}},\mathbf{f})-\mathbf{d}_r)\ ,\ \text{as}\ \frac{\partial\widetilde{\boldsymbol{\mathcal{L}}}(\boldsymbol{\delta}\widetilde{\mathbf{m}},\mathbf{f})}{\partial(\boldsymbol{\delta}\widetilde{\mathbf{m}})}=\widetilde{\mathbf{L}}\ ,\\ \implies &\frac{\partial}{\partial(\boldsymbol{\delta}\widetilde{\mathbf{m}})}\left(\frac{1}{2}||\mathbf{R}\widetilde{\boldsymbol{\mathcal{L}}}(\boldsymbol{\delta}\widetilde{\mathbf{m}},\mathbf{f})-\mathbf{d}_r||_2^2\right)\Bigg|_{\boldsymbol{\delta}\widetilde{\mathbf{m}}=\mathbf{0}}=\widetilde{\mathbf{L}}^\dagger\mathbf{R}^\dagger(\mathbf{R}\boldsymbol{\mathcal{L}}\mathbf{f}-\mathbf{d}_r)\ . \end{aligned} \tag{24}$$

If we take a step along the negative gradient and use it as an estimate of $\boldsymbol{\delta}\widetilde{\mathbf{m}}$ we obtain

$$\boldsymbol{\delta}\widetilde{\mathbf{m}}(\mathbf{m}_0)=-\widetilde{\mathbf{L}}^\dagger(\mathbf{m}_0)\mathbf{R}^\dagger(\mathbf{R}\boldsymbol{\mathcal{L}}(\mathbf{m}_0)\mathbf{f}-\mathbf{d}_r)\ . \tag{25}$$

This $\boldsymbol{\delta}\widetilde{\mathbf{m}}$ obtained this way will in general have components at $\tau\neq0$, and has an implicit dependence on $\mathbf{m}_0$. At this stage, the TFWI philosophy motivates us to find a way to change $\mathbf{m}_0$ such that $\boldsymbol{\delta}\widetilde{\mathbf{m}}$ becomes focused at $\tau=0$. To that extent, we now define a focusing operator $\boldsymbol{\mathcal{T}}(\boldsymbol{\delta}\widetilde{\mathbf{m}})$ that achieves this using Kolmogorov spectral factorization. This method proceeds by isolating the strictly positive and strictly negative time lag parts of $\boldsymbol{\delta}\widetilde{\mathbf{m}}$ at each grid point of space, and then applying minimum phase transformation to each part separately. This process is repeated at each spatial grid point to yield the final output $\boldsymbol{\mathcal{T}}(\boldsymbol{\delta}\widetilde{\mathbf{m}})$. This transformation is the first key step in how the new inversion framework differs from classical TFWI.

The next step in the proposed inversion framework involves devising an update strategy for $\mathbf{m}_0$. To achieve this, we minimize the objective function

$$J(\mathbf{m})=\sum_{i=0}^{N_x-1}\sum_{j=0}^{N_z-1}\frac{1}{2}\widetilde{W}_2^2(\boldsymbol{\delta}\widetilde{\mathbf{m}}_{ij}(\mathbf{m}),\boldsymbol{\mathcal{T}}(\boldsymbol{\delta}\widetilde{\mathbf{m}}_{ij}(\mathbf{m}_0))) \tag{26}$$

based on the quadratic Wasserstein distance that was introduced in equation (20).

The quantity $\boldsymbol{\delta}\widetilde{\mathbf{m}}_{ij}$ appearing in equation (26) represents the part of the extended perturbation model corresponding to the spatial grid location indexed by (i,j), having physical coordinates $(i\Delta x, j\Delta z)$ with respect to the origin. An important idea that the notation in equation (26) tries to emphasize is that the second argument in the Wasserstein distance $\boldsymbol{\mathcal{T}}(\boldsymbol{\delta}\widetilde{\mathbf{m}}_{ij}(\mathbf{m}_0))$ only depends on the starting model $\mathbf{m}_0$, and not on $\mathbf{m}$. This step is expected to be important in achieving a faster convergence rate than conventional TFWI, and the intuition is based on prior work done in 1D using other objective functions (Biondi et al., 2016). Of course its efficacy can only be fully judged in the future once a full suite of numerical tests have been carried out. We also note that the objective function $J(\mathbf{m})$ is a sum over all the grid points in the physical space, i.e, the square of the quadratic Wasserstein distance is first evaluated at every spatial point, and then added up to yield the complete $J(\mathbf{m})$.

The minimization proceeds by starting from $\mathbf{m} = \mathbf{m}_0$, followed by using a suitable optimization algorithm. For our initial work, we propose to look at gradient descent as a viable descent type optimization algorithm. Using the chain rule, the gradient of $J(\mathbf{m})$ can be calculated as:

$$\frac{\partial J(\mathbf{m})}{\partial \mathbf{m}} = \sum_{i=0}^{N_x-1} \sum_{j=0}^{N_z-1} \widetilde{W}_2 \frac{\partial \widetilde{W}_2}{\partial (\boldsymbol{\delta}\widetilde{\mathbf{m}}_{ij})} \frac{\partial (\boldsymbol{\delta}\widetilde{\mathbf{m}}_{ij})}{\partial \mathbf{m}} \ , \tag{27}$$

where we have dropped the explicit dependence of the quantities to make the notation concise. The quantity $\widetilde{W}_2 \frac{\partial \widetilde{W}_2}{\partial (\boldsymbol{\delta}\widetilde{\mathbf{m}}_{ij})}$ can be calculated easily using finite differences. The other quantity $\frac{\partial (\boldsymbol{\delta}\widetilde{\mathbf{m}}_{ij})}{\partial \mathbf{m}}$ are columns of the matrix $\frac{\partial (\boldsymbol{\delta}\widetilde{\mathbf{m}})}{\partial \mathbf{m}}$. Using equation (25) this operator is given by

$$\begin{aligned} \frac{\partial (\boldsymbol{\delta}\widetilde{\mathbf{m}})}{\partial \mathbf{m}} &= -\frac{\partial}{\partial \mathbf{m}} \left(\widetilde{\mathbf{L}}^\dagger \mathbf{R}^\dagger (\mathbf{R}\boldsymbol{\mathcal{L}}\mathbf{f} - \mathbf{d}_r) \right) \\ &= -\left(\frac{\partial \widetilde{\mathbf{L}}^\dagger}{\partial \mathbf{m}} \right) \mathbf{R}^\dagger (\mathbf{R}\boldsymbol{\mathcal{L}}\mathbf{f} - \mathbf{d}_r) - \widetilde{\mathbf{L}}^\dagger \mathbf{R}^\dagger \frac{\partial (\mathbf{R}\boldsymbol{\mathcal{L}}\mathbf{f})}{\partial \mathbf{m}} \\ &= -\left(\frac{\partial \widetilde{\mathbf{L}}^\dagger}{\partial \mathbf{m}} \right) \mathbf{R}^\dagger (\mathbf{R}\boldsymbol{\mathcal{L}}\mathbf{f} - \mathbf{d}_r) - \widetilde{\mathbf{L}}^\dagger \mathbf{R}^\dagger \mathbf{R}\mathbf{L} \ . \end{aligned} \tag{28}$$

We now have all the machinery in place to solve the proposed inversion problem. We summarize the steps of the algorithm below:

- Assuming that at the k^{th} iteration the model is denoted by $\mathbf{m}^k$, we calculate the quantity $\boldsymbol{\delta}\widetilde{\mathbf{m}}^k$ as follows:
$$\boldsymbol{\delta}\widetilde{\mathbf{m}}^k(\mathbf{m}^k) = -\widetilde{\mathbf{L}}^\dagger \mathbf{R}^\dagger (\mathbf{R}\boldsymbol{\mathcal{L}}\mathbf{f} - \mathbf{d}_r) \ .$$
- Using Kolmogorov factorization, we then calculate the quantity $\boldsymbol{\mathcal{T}}(\boldsymbol{\delta}\widetilde{\mathbf{m}}^k_{ij}(\mathbf{m}^k))$, for all $0 \leq i \leq N_x - 1$ and $0 \leq j \leq N_z - 1$.
- We then perform gradient descent iterations to get a new model $\mathbf{m}^{k+1}$ starting from $\mathbf{m}^k$, on the objective function
$$J(\mathbf{m}) = \sum_{i=0}^{N_x-1} \sum_{j=0}^{N_z-1} \frac{1}{2} \widetilde{W}_2^2 (\boldsymbol{\delta}\widetilde{\mathbf{m}}_{ij}(\mathbf{m}), \boldsymbol{\mathcal{T}}(\boldsymbol{\delta}\widetilde{\mathbf{m}}_{ij}(\mathbf{m}^k))) \ .$$
- Setting $k = k + 1$, we iterate till convergence / stopping criterion.

PRELIMINARY NUMERICAL RESULTS

In this section we present some preliminary numerical results with a software library under development, that can be used to solve a class of FWI problems in 2D, including the new proposed optimization algorithm. A key feature of this library is that it attempts to integrate numerical algorithms from the world of optimal transport into the TFWI scheme with modified objective functions, like the one proposed in this paper. We intend the library to also be capable of solving the standard FWI / TFWI problems, allowing us to numerically measure any gains in efficiency the optimal transport based algorithms provide.

We still have a lot of work to do with the library, and it is in its preliminary stages. As of now, we can only solve the standard FWI problem reliably. We are currently working on implementing efficient versions of standard TFWI solver along with the optimal transport variant. In the rest of this section, we present some results where we try to invert for a low-velocity Gaussian anomaly in a two layer background model. These results illustrate the failure of the FWI objective function.

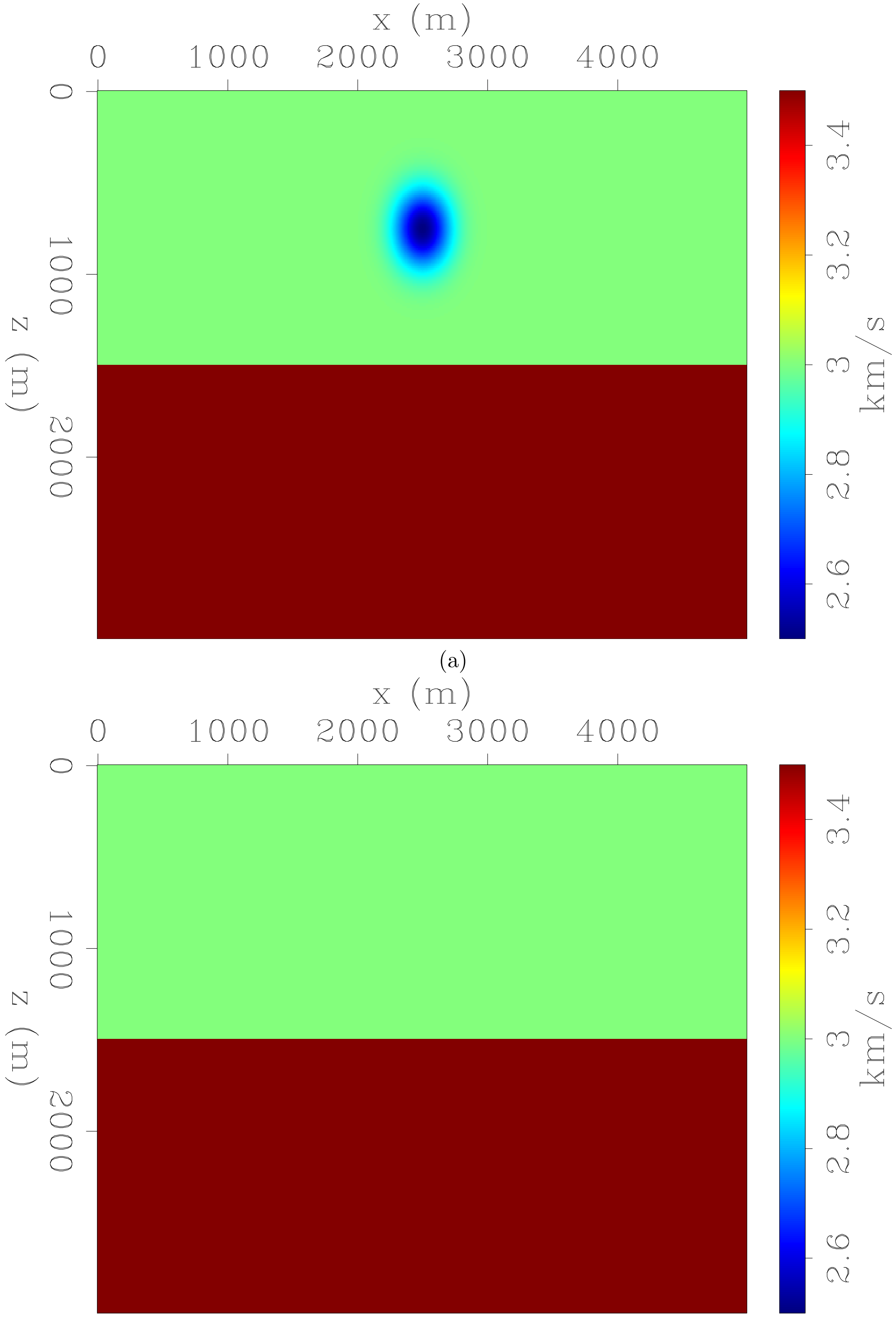
Modeling synthetic data

We start by creating a 5 km x 3 km model along X and Z directions respectively with grid spacing 10 m x 10 m along each direction. This represents our simulation world. In Figure 1a, we display the original model. The background is a two layer model with velocities 3 km/s and 3.5 km/s with the interface between the two layers located at $Z = 1.5$ km. There is a low velocity Gaussian anomaly present in the original model, centered at $X = 2.5$ km, $Z = 0.75$ km, with a peak velocity anomaly value of -0.5 km/s. We will refer to the original model as $\mathbf{m}_{true}$.

To model the simulated shots, we place receivers at every grid location on the surface of the model, i.e, at $Z = 0$ km. The source locations are placed at every 100 m intervals along X, also at $Z = 0$ km. An example shot is plotted in Figure 2 and corresponds to a shot located at $X = 2.5$ km, $Z = 0$ km. The data are modeled using a Ricker wavelet with dominant frequency 12.5 Hz, which is plotted in Figure 3. As can be clearly seen in the shot record, the low velocity Gaussian anomaly introduces triplication of raypaths that lead to the effect seen at around $X = 2.5$ km, $t = 1.2$ s.

Failure of FWI

Figure 1b represents the starting velocity model used in the FWI inversion, which is identical to the original velocity model (Figure 1a) without the Gaussian anomaly. We will refer to it as $\mathbf{m}_0$. In Figure 4a, we plot the modeled data with $\mathbf{m}_0$ for the same central shot location given by $X = 2.5$ km, $Z = 0$ km. The residual for this shot is also shown in Figure 4b. On the residual, we can clearly see cycle-skipping on the mid to far offsets. The FWI gradient with $\mathbf{m}_0$ is displayed in Figure 5.



(a)

(b)

Figure 1: a) The original 2D model $\mathbf{m}_{true}$ used to generate synthetic shots for the study. b) The starting model $\mathbf{m}_0$ used for starting FWI. [**ER**]

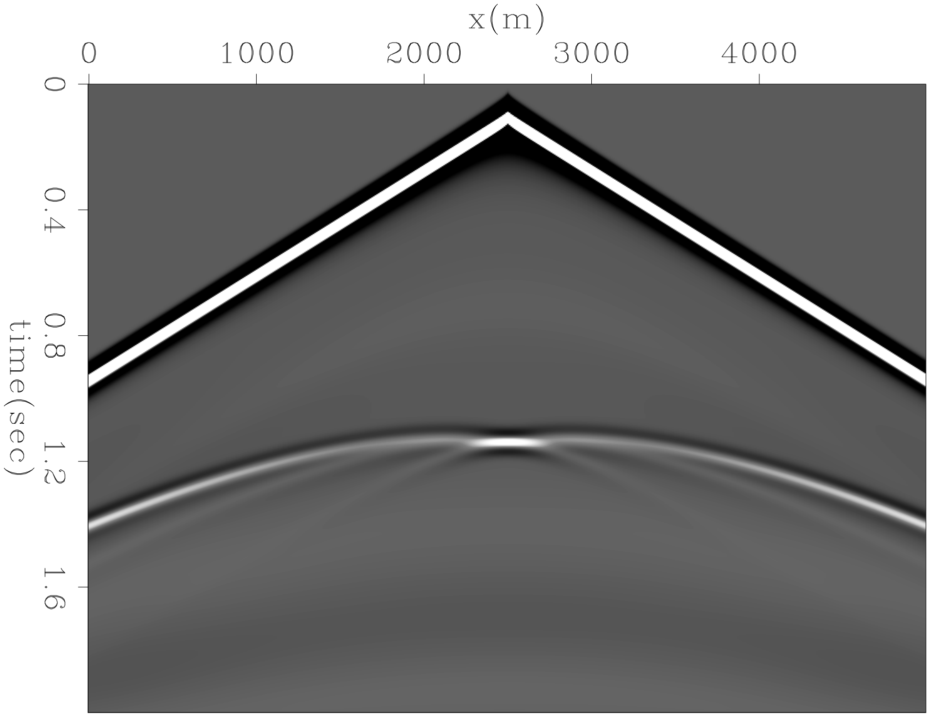


Figure 2: A shot record extracted from the center of the survey. The location of source is $X = 2.5$ km, $Z = 0$ km. [**ER**]

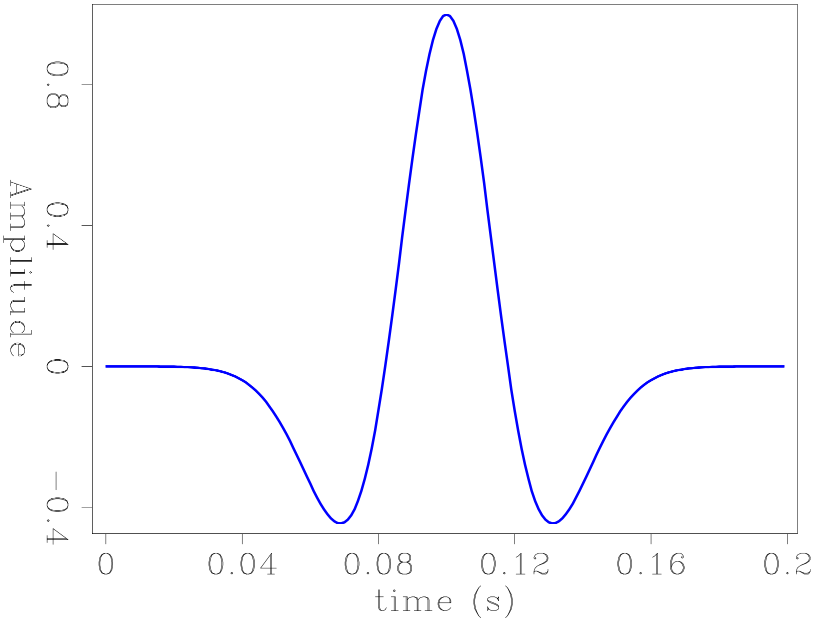
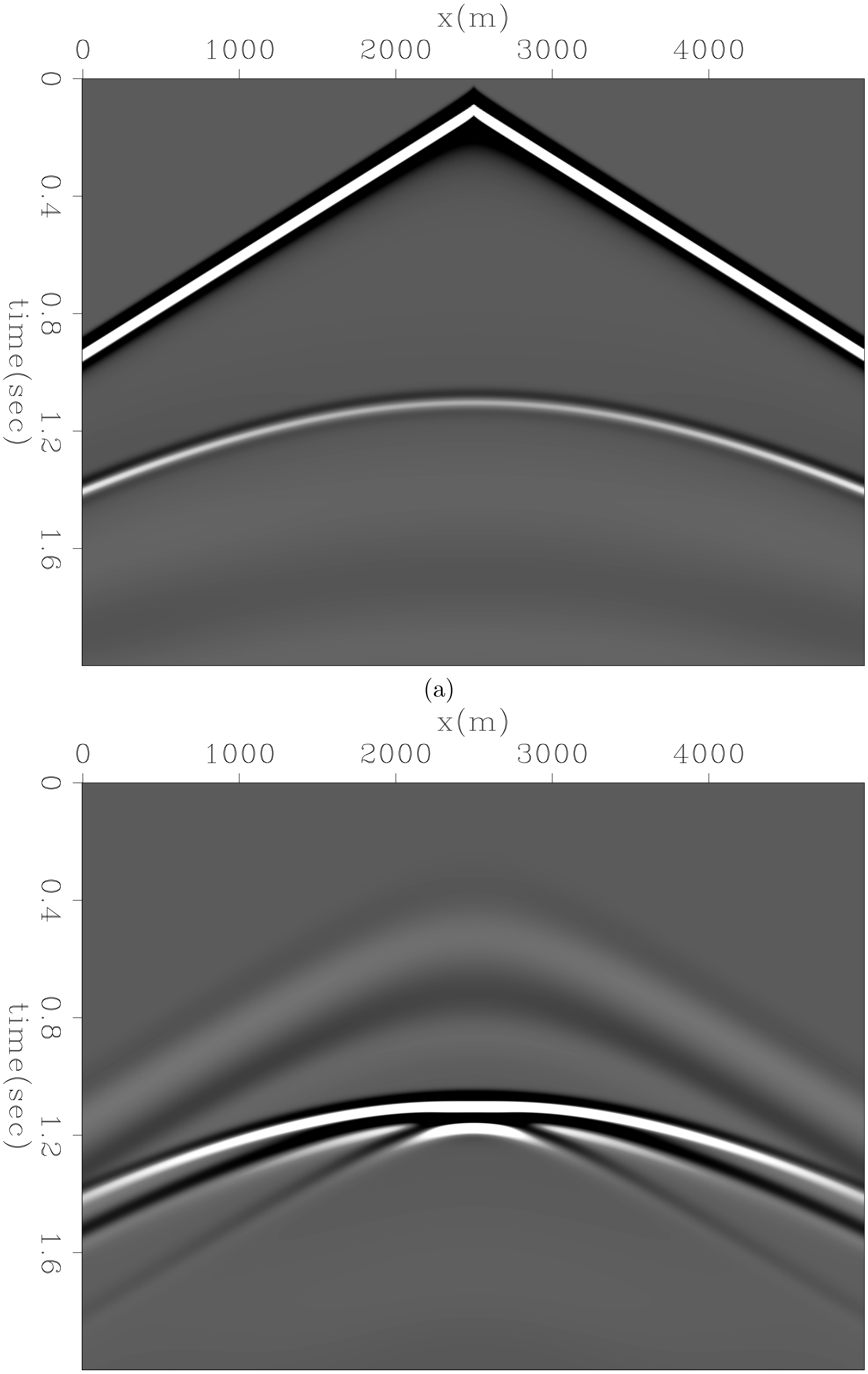


Figure 3: Ricker wavelet used in the modeling of synthetic data (12.5 Hz dominant frequency). [**ER**]



(a)

(b)

Figure 4: a) Data modeled for the same central shot as in Figure 2 with the starting velocity model $\mathbf{m}_0$. b) The residual for the same shot with gain applied where we can see cycle-skipping on mid to far offsets. [**ER**]

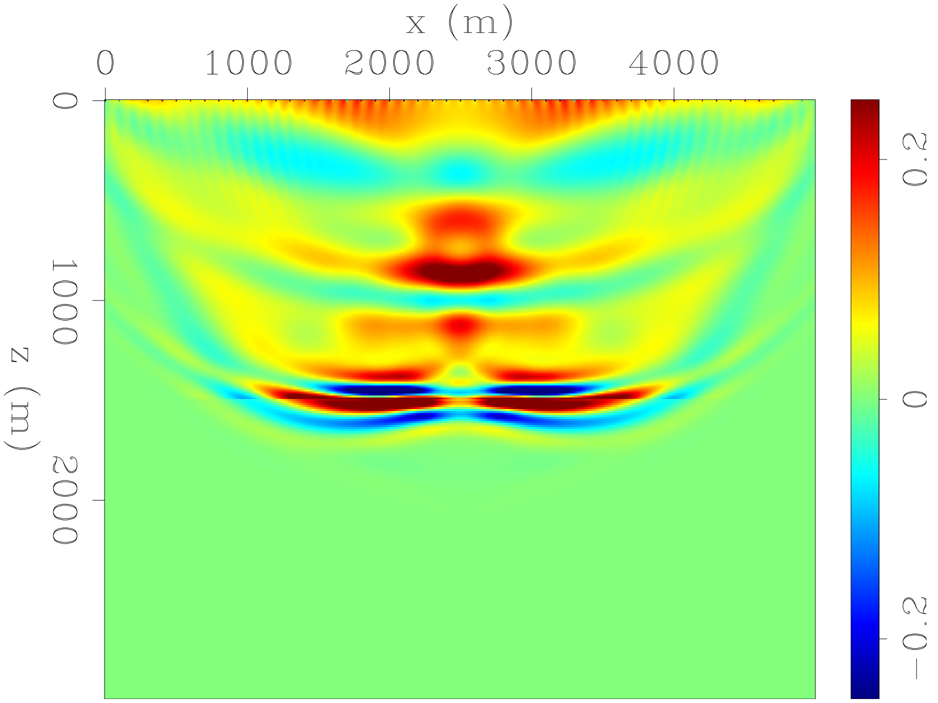
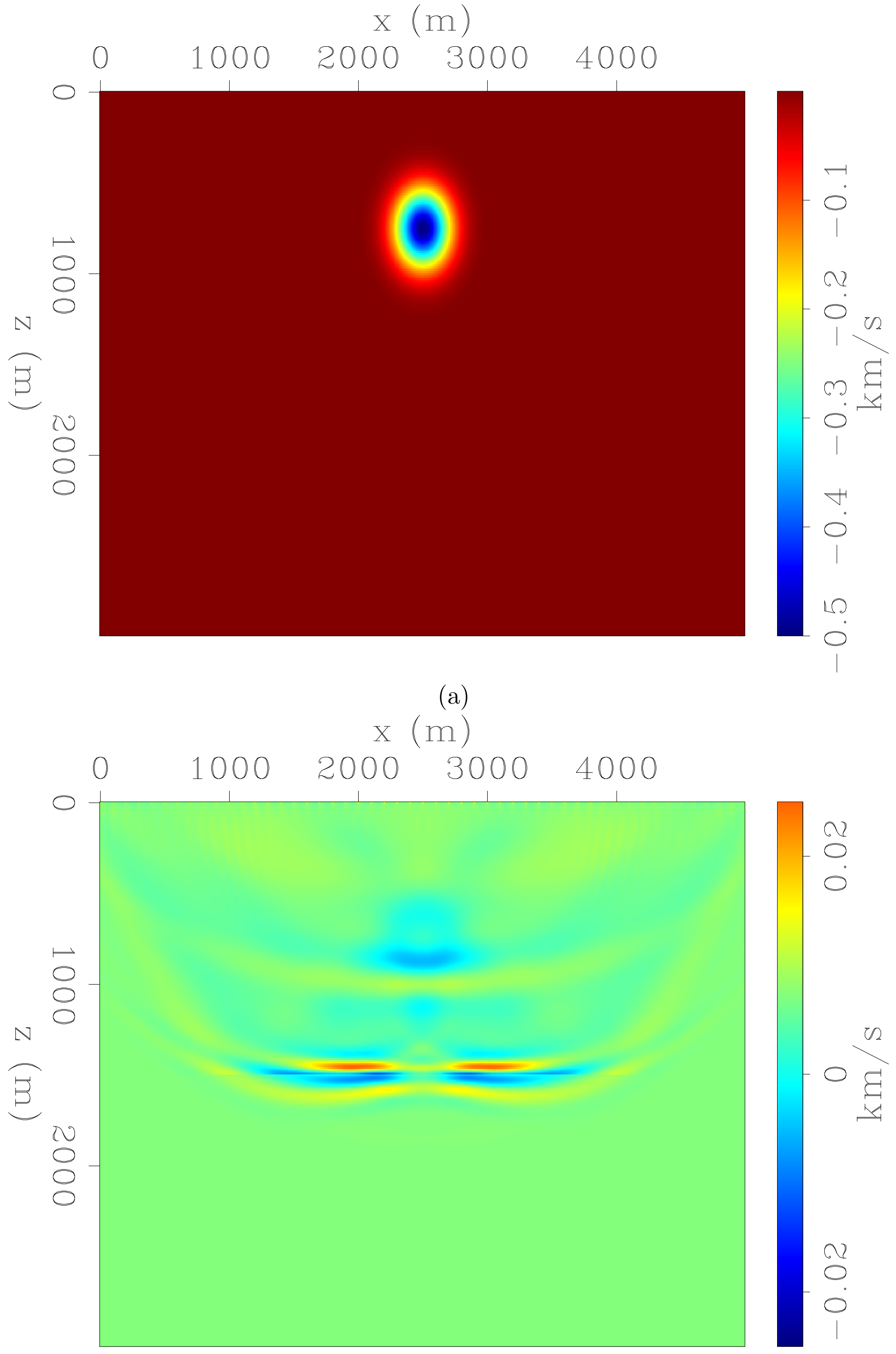


Figure 5: FWI gradient calculated using the starting model $\mathbf{m}_0$. [**CR**]

We ran 10 iterations of FWI, starting from $\mathbf{m}_0$. The result of the update in the velocity after 10 iterations is shown in Figure 6b. This can be contrasted with the true update required to recover the original model as shown in Figure 6a. The maximum update is only about 20 m/s after 10 iterations. Running more FWI iterations does not improve the model significantly. Moreover we also observed that its convergence is really slow as we obtained very small step sizes during the line search process used in the inversion. Lastly, it is to be noted that this example was run to illustrate the limitation of FWI in the presence of cycle-skipping and thus we specifically did not use a multiscale inversion scheme. The inversion was carried out in the same frequency band as the original synthetic data.

CONCLUSIONS

In this paper, we have introduced the concept of optimal transport as a tool to modify the standard TFWI inversion algorithm, and proposed a new inversion framework that can help address the issue of slow convergence in TFWI. We have also started to develop a 2D library that can be used to solve the new optimization problem. The library is intended to be capable of also solving the standard FWI and the standard TFWI problems, for efficient comparison. Some preliminary results with this library have been presented in this paper. In particular, we have demonstrated the need for devising better alternatives to the standard FWI algorithm, by showing its failure to converge, even in the case of a simple low-velocity Gaussian anomaly in a two layer



(a)

(b)

Figure 6: a) Difference between true model $\mathbf{m}_{true}$ and initial starting model $\mathbf{m}_0$. b) Difference between inverted model after 10 iterations and initial starting model $\mathbf{m}_0$. [**CR**]

model.

ACKNOWLEDGMENTS

Rahul Sarkar would like to thank Bob Clapp for his advise on several computational aspects of this work, that allowed him to write software to generate results in a reasonable amount of time. He would also like to thank Jason Chang, Ettore Biondi, Guillaume Barnier and Yinbin Ma for several discussions and their help with SEPlib. We thank Ettore Biondi and Stewart Levin for reading the manuscript and suggesting edits that greatly helped to improve the quality of the document. Finally, we thank all the sponsors of SEP for their generous support and funding.

REFERENCES

Almomin, A. and B. Biondi, 2014, Preconditioned tomographic full waveform inversion by wavelength continuation: SEG Technical Program Expanded Abstracts, **33**, 944–948.

Biondi, B. and A. Almomin, 2014, Simultaneous inversion of full data bandwidth by tomographic full waveform inversion: Geophysics, **79**, WA129–WA140.

Biondi, B., R. Sarkar, and J. Jennings, 2016, Solving nonlinear inverse problems by linearized model extension — A survey of possible methods: SEP-Report, **165**, 93–122.

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

Claerbout, J. F., 1985, Fundamentals of geophysical data processing.

Engquist, B. and B. D. Froese, 2014, Application of the Wasserstein metric to seismic signals: Communications in Mathematical Sciences, **12**, 979–988.

Engquist, B., B. D. Froese, and Y. Yang, 2016, Optimal transport for seismic full waveform inversion: Communications in Mathematical Sciences, **14**, 2309–2330.

Kolmogorov, A. N., 1939, Sur linterpolation et extrapolation des suites stationnaires: C.R. Acad. Sci, **208**, 2043–2045.

Luo, S. and P. Sava, 2011, A deconvolution-based objective function for wave-equation inversion: SEG Technical Program Expanded Abstracts, 2788–2792.

Luo, Y. and G. T. Schuster, 1991, Wave-equation traveltime inversion: Geophysics, **56**, 645–653.

Ma, Y. and D. Hale, 2013, Wave-equation reflection traveltime inversion with dynamic warping and full-waveform inversion: Geophysics, **78**, R223–R233.

Métivier, L., R. Brossier, Q. Mérigot, E. Oudet, and J. Virieux, 2016, Measuring the misfit between seismograms using an optimal transport distance: Application to full waveform inversion: Geophysical Journal International, **205**, 345–377.

Pratt, R. G., 1999, Seismic waveform inversion in the frequency domain, Part 1: Theory and verification in a physical scale model: Geophysics, **64**, 888–901.

Robinson, E. and S. Treitel, 1980, Geophysical signal analysis: Prentice-hall: Englewood Cliffs.

Sava, P. and B. Biondi, 2004a, Wave-equation migration velocity analysis—I: Theory: Geophysical Prospecting, **52**, 593–606.

———, 2004b, Wave-equation migration velocity analysis—II: Examples: Geophysical Prospecting, **52**, 607–623.

Shen, P. and W. W. Symes, 2008, Automatic velocity analysis via shot profile migration: Geophysics, **73**, VE49–VE59.

Symes, W. W., 2008, Migration velocity analysis and waveform inversion: Geophysical Prospecting, **56**, 765–790.

Symes, W. W. and J. J. Carazzone, 1991, Velocity inversion by differential semblance optimization: Geophysics, **56**, 654–663.

Villani, C., 2003, Topics in optimal transportation: American Mathematical Soc.

———, 2008, Optimal transport: Old and New, volume **338**: Springer Science & Business Media.

Yang, Y., B. Engquist, J. Sun, and B. D. Froese, 2016, Application of optimal transport and the quadratic Wasserstein metric to full-waveform inversion: arXiv preprint arXiv:1612.05075.

Zhang, Y. and B. Biondi, 2013, Moveout-based wave-equation migration velocity analysis: Geophysics, **78**, U31–U39.