GEOPHYSICS, VOL. 83, NO. 1 (JANUARY-FEBRUARY 2018); P. R43–R62, 29 FIGS.
10.1190/GEO2016-0663.1

Application of optimal transport and the quadratic Wasserstein metric to full-waveform inversion

Yunan Yang[1], Björn Engquist[1], Junzhe Sun[2], and Brittany F. Hamfeldt[3]

ABSTRACT

Conventional full-waveform inversion (FWI) using the least-squares norm as a misfit function is known to suffer from cycle-skipping issues that increase the risk of computing a local rather than the global minimum of the misfit. The quadratic Wasserstein metric has proven to have many ideal properties with regard to convexity and insensitivity to noise. When the observed and predicted seismic data are considered to be two density functions, the quadratic Wasserstein metric corresponds to the optimal cost of rearranging one density into the other, in which the transportation cost is quadratic in distance. Unlike the least-squares norm, the quadratic Wasserstein metric measures not only amplitude differences but also global phase shifts, which helps to avoid cycle-skipping issues. We have developed a new way of using the quadratic Wasserstein metric trace by trace in FWI and compare it with the global quadratic Wasserstein metric via the solution of the Monge-Ampère equation. We incorporate the quadratic Wasserstein metric technique into the framework of the adjoint-state method and apply it to several 2D examples. With the corresponding adjoint source, the velocity model can be updated using a quasi-Newton method. Numerical results indicate the effectiveness of the quadratic Wasserstein metric in alleviating cycle-skipping issues and sensitivity to noise. The mathematical theory and numerical examples demonstrate that the quadratic Wasserstein metric is a good candidate for a misfit function in seismic inversion.

INTRODUCTION

Full-waveform inversion (FWI) was originally proposed three decades ago in an attempt to obtain high-resolution subsurface properties based on seismic waveforms (Lailly, 1983; Tarantola, 1984). Over the past few decades, there have been many encouraging results using FWI in the seismic processing of marine and land data (Virieux and Operto, 2009; Sirgue et al., 2010). FWI iteratively updates a subsurface model and computes the corresponding synthetic data to reduce the data misfit between the synthetic and recorded seismic data.

The objective of FWI is to match the synthetic and recorded data in a comprehensive way such that all information in the waveforms is accounted for in the data misfit. If we denote the predicted data by f and the observed data by g, then the unknown velocities are determined by minimizing the mismatch $d(f, g)$.

FWI has the potential to generate high-resolution subsurface models but suffers from the ill posedness of the inverse problem. This issue can be handled by considering multiple data components ranging from low to high frequency (Bunks et al., 1995) or by adding regularization terms (Gholami and Siahkoohi, 2010; Esser et al., 2015; Qiu et al., 2016).

The least-squares norm L^2 is the most widely used misfit function in FWI but suffers from cycle skipping and sensitivity to noise. Other norms have been proposed in the literature including the L^1 norm, the Huber norm (Ha et al., 2009), and hybrid L^1/L^2 norms (Brossier et al., 2010). These misfit functions follow the same path of dealing with the predicted and observed data independently.

Differences between the predicted velocity model and true model produce a misfit in the data, which is the information FWI uses to update the velocity model. This motivates us to take a different view of the predicted and observed data by considering a "map" connect-

Manuscript received by the Editor 8 December 2016; revised manuscript received 18 August 2017; published ahead of production 10 October 2017; published online 05 January 2018.
[1]The University of Texas at Austin, Department of Mathematics, Austin, Texas, USA. E-mail: yunanyang@math.utexas.edu; engquist@ices.utexas.edu.
[2]The University of Texas at Austin, Bureau of Economic Geology, John A. and Katherine G. Jackson School of Geosciences, Austin, Texas, USA. E-mail: junzhesun@utexas.edu.
[3]New Jersey Institute of Technology, Department of Mathematical Sciences, University Heights, Newark, New Jersey, USA. E-mail: bdfroese@njit.edu.

ing them (Ma and Hale, 2013). The idea of mapping synthetic data to observed data with stationary and nonstationary filters in the time domain has been promoted recently (Warner and Guasch, 2014; Zhu and Fomel, 2016). Although the misfits in these two approaches are not critical metrics between two objects in mathematics, they demonstrate the advantages and feasibility of map-based ideas.

Optimal transport has become a well-developed topic in mathematics since it was first proposed by Monge (1781). Due to their ability to incorporate differences in intensity and spatial information, optimal transport-based metrics for modeling and signal processing have recently been adopted in a variety of applications including image retrieval, cancer detection, and machine learning (Kolouri et al., 2016).

The idea of using optimal transport for seismic inversion was first proposed by Engquist and Froese (2014). The Wasserstein metric is a concept based on optimal transportation (Villani, 2003). Here, we treat our data sets of seismic signals as density functions of two probability distributions, which can be imagined as the distributions of two piles of sand with equal mass. Given a particular cost function, different plans of transporting one pile into the other lead to different costs. The plan with the lowest cost is the optimal map, and this lowest cost is the Wasserstein metric. In computer science, the metric is often called the "earth mover's distance." Here, we will focus on the quadratic cost functions. The corresponding misfit is the quadratic Wasserstein metric W_2.

Following the idea that changes in velocity cause a shift or "transport" in the arrival time, Engquist et al. (2016) demonstrate the advantageous mathematical properties of the quadratic Wasserstein metric W_2 and provide rigorous proofs that lay a solid theoretical foundation for this new misfit function. In this paper, we continue the study of the quadratic Wasserstein metric with more focus on its applications to FWI. We also develop a fast and robust trace-by-trace technique.

After the paper of Engquist and Froese (2014), researchers in geophysics started to work on other optimal transport-related misfit functions (Métivier et al., 2016a, 2016b, 2016c). The Kantorovich-Rubinstein (KR) norm in their papers is a relaxation of the 1-Wasserstein distance, which is another optimal transport metric with the absolute value cost function. The advantage of the KR norm is that it does not require data to satisfy nonnegativity or mass balance conditions.

The Wasserstein distance measures the difference between nonnegative measures or functions with equal mass. These are not natural constraints for seismic signals, and thus they first have to be normalized. In our earlier work, we separated the positive and negative part of the signals to achieve nonnegativity. The resulting signal was then divided by its integral. This worked well in our earlier test cases, but it is less effective for the larger scale problems with the adjoint-state method studied here. In this paper, we apply a linear transformation to the signals to satisfy the requirements of optimal transport. This, on the other hand, is effective in spite of the fact that it results in a measure that is not convex concerning simple shifts.

In one dimension, the optimal transport problem can be solved explicitly, which allows for accurate and efficient computations. However, computation becomes much more challenging in higher dimensions. Several numerical methods have been proposed, but these still have limitations for extremely large scale realistic data sets, e.g., those in seismic inversion. Numerical methods based on the Benamou-Brenier fluid formulation introduce an extra time dimension to the problem, which increases the computational cost (Benamou and Brenier, 2000). Optimal transport via entropic regularization is computationally efficient but with very low accuracy in the computed map (Benamou et al., 2015). The numerical solution may become unstable when the regularization term is small because it is close to the original optimal transport problem. Methods based on linear programming have the disadvantage of doubling the dimension of the underlying problem (Oberman and Ruan, 2015; Schmitzer, 2016). For the quadratic Wasserstein distance, the optimal map can be computed via the solution of a Monge-Ampère partial differential equation (PDE) (Benamou et al., 2014). This approach has the advantage of drawing on the more well-developed field of numerical PDEs. The drawback to the PDE approach is that data must be sufficiently regular for solutions to be well defined. To remain robust on realistic examples, these methods effectively smooth the seismic data, which can lead to a loss of high-frequency information. For illustration in this paper, we will perform computations using a Monge-Ampère solver for synthetic examples. Even in 2D, some limitations are apparent. This is expected to become even more of a problem in higher dimensions and motivates our introduction of a trace-by-trace technique that relies on the exact 1D solution. The trace-by-trace technique is currently more promising for practical problems, as is evidenced in our computational examples.

In this paper, we briefly review the theory of optimal transport and revisit the mathematical properties of W_2 that were proved by Engquist et al. (2016), including the convexity and insensitivity to noise. Next, we apply the quadratic Wasserstein metric W_2 as misfit function in two different ways: trace-by-trace comparison and entire data set comparison. The trace-by-trace strategy and global strategy lead to different formulations of the misfit computation and the adjoint source (Plessix, 2006). The trace-by-trace technique is new, and the results for inversion are very encouraging. The computational cost is low and similar to that of the classic L^2 method. Finally, after introducing the adjoint source formulas, we show the application of FWI using the W_2 metric on three synthetic models: the Camembert, the Marmousi, and the 2004 BP models. Discussions and comparisons between the FWI results using W_2 and L^2 metrics illustrate that the W_2 metric is very promising for overcoming the cycle-skipping issue in FWI.

THEORY

Formulation

Conventional FWI defines a least-squares waveform misfit as

$$d(f,g) = J_0(m) = \frac{1}{2}\sum_r \int |f(x_r,t;m) - g(x_r,t)|^2 dt, \quad (1)$$

where g is the observed data, f is the simulated data, x_r are the receiver locations, and m is the model parameter. This formulation can also be extended to the case with multiple shots. We get the modeled data $f(x,t;m)$ by solving a wave equation with a finite-difference method (FDM) in the space and time domain.

In this paper, we propose using the quadratic Wasserstein metric W_2 as an alternative misfit function to measure the difference

between the synthetic data f and observed data g. There are two ways to apply this idea: trace-by-trace W_2 and global W_2.

We can compare the data trace by trace and use the quadratic Wasserstein metric W_2 in 1D to measure the misfit. The overall misfit is then

$$J_1(m) = \sum_{r=1}^{R} W_2^2(f(x_r, t; m), g(x_r, t)), \tag{2}$$

where R is the total number of traces.

In the global case, we compare the full data sets and consider the whole synthetic data f and observed data g as objects with the general quadratic Wasserstein metric W_2:

$$J_2(m) = W_2^2(f(x_r, t; m), g(x_r, t)). \tag{3}$$

We treat the misfit $J(m)$ as a function of the model parameter m. Our aim is to find the model parameter m^* that minimizes the objective function, i.e., $m^* = \operatorname{argmin} J(m)$. This is a PDE-constrained optimization problem, and we use a gradient-based iterative scheme to update the model m.

Background

Optimal transport originated in 1781 with the French mathematician Monge. This problem seeks the minimum cost required to transport the mass of one distribution into another given a cost function. More specifically, we consider two probability measures μ and ν defined on spaces X and Y, respectively. For simplicity, we regard X and Y as subsets of $\mathbb{R}^d$. Measures μ and ν have density functions f and g: $d\mu = f(x)dx$ and $d\nu = g(y)dy$. In applications, $f(x)$ can represent the height of a pile of sand at location x, the gray scale of one pixel x for an image, or as here the amplitude of a seismic waveform at mesh grid point x.

Although they must share the same total mass, measures μ and ν are not the same; i.e., $f \neq g$. We want to redistribute "sand" from μ into ν, and it requires effort. The cost function $c(x, y)$ maps pairs $(x, y) \in X \times Y$ to $\mathbb{R} \cup \{+\infty\}$, which denotes the cost of transporting one unit mass from location x to y. The most common choices of $c(x, y)$ include $|x - y|$ and $|x - y|^2$. Once we find a transport plan $T: X \to Y$ such that for any measurable set $B \subset Y$, $\nu[B] = \mu[T^{-1}(B)]$, the cost corresponding to this plan T is

$$I(T, f, g, c) = \int_X c(x, T(x)) f(x) dx. \tag{4}$$

Although there are many maps T that can perform the relocation, we are interested in finding the optimal map that minimizes the total cost

$$I(f, g, c) = \inf_{T \in \mathcal{M}} \int_X c(x, T(x)) f(x) dx, \tag{5}$$

where $\mathcal{M}$ is the set of all maps that rearrange f into g.

Thus, we have informally defined the optimal transport problem, the optimal map as well as the optimal cost, which is also called the Wasserstein distance:

Definition 1 (The Wasserstein distance). *We denote by $\mathcal{P}_p(X)$ the set of probability measures with finite moments of order p. For all $p \in [1, \infty)$,*

$$W_p(\mu, \nu) = \left(\inf_{T \in \mathcal{M}} \int_{\mathbb{R}^n} |x - T(x)|^p d\mu(x) \right)^{\frac{1}{p}}, \quad \mu, \nu \in P_p(X). \tag{6}$$

$\mathcal{M}$ is the set of all maps that rearrange the distribution μ into ν.

In this paper, we focus on the case of a quadratic cost function: $c(x, y) = |x - y|^2$. The mathematical definition of the distance between the distributions $f: X \to \mathbb{R}^+$ and $g: Y \to \mathbb{R}^+$ can then be formulated as

$$W_2^2(f, g) = \inf_{T \in \mathcal{M}} \int_X |x - T(x)|^2 f(x) dx, \tag{7}$$

where $\mathcal{M}$ is the set of all maps that rearrange the distribution f into g (for details, see Villani, 2003). The optimal transport formulation requires nonnegative distributions and equal total masses that are not natural for seismic signals. We will discuss this in the section on data normalization below.

Optimal transport on the real line

For f and g in one dimension, it is possible to exactly solve the optimal transportation problem (Villani, 2003) in terms of the cumulative distribution functions

$$F(x) = \int_{-\infty}^{x} f(t) dt, G(y) = \int_{-\infty}^{y} g(t) dt. \tag{8}$$

In fact, the optimal map is just the unique monotone rearrangement of the density f into g (Figure 1a). To compute the quadratic Wasserstein metric W_2, we need the cumulative distribution functions F and G (Figure 1b) and their inverses F^{-1} and G^{-1} (Figure 1c) as the following theorem states.

Theorem 1 (Optimal transportation for a quadratic cost on $\mathbb{R}$). *Let $0 < f, g < \infty$ be two probability density functions, each supported on a connected subset of $\mathbb{R}$. Then the optimal map from f to g is $T = G^{-1} \circ F$.*

For the synthetic data f and the observed data g from one trace, we assume that they are continuous in time without loss of generality. After proper normalization signals f and g can be rescaled to be positive, supported on [0,1], and have a total mass of one. From the theorem above, we derive another formulation for the 1D quadratic Wasserstein metric:

$$W_2^2(f, g) = \int_0^1 |x - G^{-1}(F(x))|^2 f(x) dx. \tag{9}$$

Optimal transport in higher dimensions

The simple exact formula for 1D optimal transportation does not extend to optimal transportation in higher dimensions. Nevertheless, it can be computed by relying on two important properties

of the optimal mapping $T(x)$: conservation of mass and cyclical monotonicity. From the definition of the problem, $T(x)$ maps f into g. The change of variables formula formally leads to the requirement

$$f(x) = g(T(x))\det(\nabla T(x)). \tag{10}$$

The optimal map takes on additional structure in the special case of a quadratic cost function: It is cyclically monotone (Knott and Smith, 1984; Brenier, 1991).

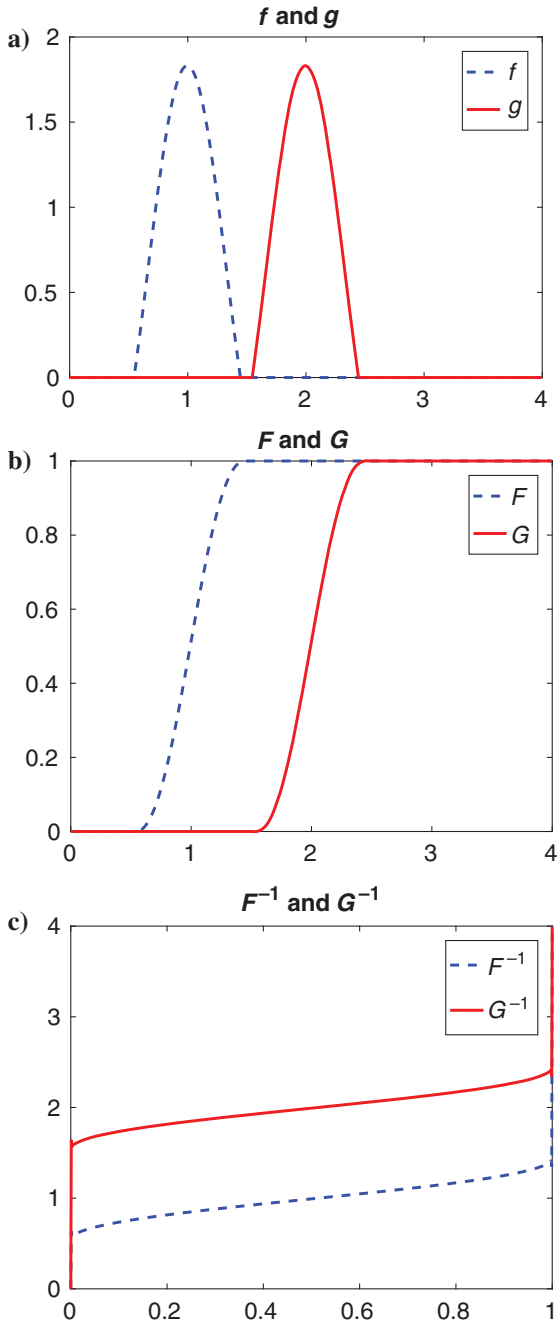

Figure 1. (a) One-dimensional densities f and g. (b) Cumulative distribution functions F and G and (c) inverse distribution function F^{-1} and G^{-1} for densities f and g.

Definition 2 (cyclical monotonicity). *We say that* $\boldsymbol{T}:\boldsymbol{X}\rightarrow\boldsymbol{Y}$ *is cyclically monotonic if for any* $\boldsymbol{m}\in\mathbb{N}^{+}$, $\boldsymbol{x_i}\in\boldsymbol{X}, 1\leq\boldsymbol{i}\leq\boldsymbol{m}$,

$$\sum_{i=1}^{m}|\boldsymbol{x_i}-\boldsymbol{T}(\boldsymbol{x_i})|^2\leq\sum_{i=1}^{m}|\boldsymbol{x_i}-\boldsymbol{T}(\boldsymbol{x}_{i-1})|^2 \tag{11}$$

or equivalently

$$\sum_{i=1}^{m}\langle\boldsymbol{T}(\boldsymbol{x_i}),\boldsymbol{x_i}-\boldsymbol{x}_{i-1}\rangle\geq 0 \tag{12}$$

where $\boldsymbol{x}_0\equiv\boldsymbol{x_m}$.

In addition, a cyclically monotone mapping is formally equivalent to the gradient of a convex function (Knott and Smith, 1984; Brenier, 1991). Making the substitution $T(x)=\nabla u(x)$ into the constraint (equation 10) leads to the Monge-Ampère equation

$$\det(D^2u(x))=\frac{f(x)}{g(\nabla u(x))},\quad u\text{ is convex.} \tag{13}$$

To compute the misfit between distributions f and g, we first compute the optimal map $T(x)=\nabla u(x)$ via the solution of this Monge-Ampère equation coupled to the nonhomogeneous Neumann boundary condition

$$\nabla u(x)\cdot n=x\cdot n,\quad x\in\partial X. \tag{14}$$

The squared Wasserstein metric is then given by

$$W_2^2(f,g)=\int_X f(x)|x-\nabla u(x)|^2dx. \tag{15}$$

Convexity

As demonstrated by Engquist et al. (2016), the squared Wasserstein metric has several properties that make it attractive as a choice of misfit function. One highly desirable feature is its convexity for data shifts, dilation, and partial amplitude change, which occur naturally in seismic waveform inversion.

We recall the overall setup for FWI, in which we have a fixed observation g and a simulation $f(m)$ that depends on unknown model parameters m. The model parameters are recovered via the minimization

$$m^*=\underset{m}{\operatorname{argmin}}\{W_2^2(f(m),g)\}. \tag{16}$$

To perform this minimization effectively and efficiently, we desire the distance $W_2^2(f(m),g)$ to be convex in the model parameter m.

This is certainly not the case for all possible functions $f(m)$, but it is true for many settings that occur naturally in seismic inversion. For example, variations in the wave velocity lead to simulations $f(m)$ that are derived from shifts,

$$f(x;s)=g(x+s\eta),\quad \eta\in\mathbb{R}^n, \tag{17}$$

or dilations,

$$f(x;A) = g(Ax), A^T = A, \quad A > 0, \tag{18}$$

applied to the observation g. Variations in the strength of a reflecting surface or the focusing of seismic waves can also lead to local rescalings of the form

$$f(x;\beta) = \begin{cases} \beta g(x), & x \in E \\ g(x), & x \in \mathbb{R}^n \setminus E. \end{cases} \tag{19}$$

Proving the convexity of W_2^2 follows nicely from the interpretation of the misfit as a transportation cost, with the underlying transportation cost exhibiting a great deal of structure. In particular, the cyclical monotonicity of the transport map $T(x)$ leads readily to estimates of

$$W_2^2(f(\lambda m_1 + (1-\lambda)m_2), g), \quad 0 < \lambda < 1, \tag{20}$$

which in turn yields the desired convexity results. The convexity was studied in detail by Engquist et al. (2016), where the following theorem was proved.

Theorem 2 (convexity of the squared Wasserstein metric [Engquist et al., 2016]). *The squared Wasserstein metric $\boldsymbol{W_2^2(f(m), g)}$ is convex with respect to the model parameters $\boldsymbol{m}$ corresponding to a shift $\boldsymbol{s}$ in (17), the eigenvalues of a dilation matrix $\boldsymbol{A}$ in (18), or the local rescaling parameter $\boldsymbol{\beta}$ in (19).*

Insensitivity to noise

When performing FWI with real data, it is natural to experience noise in the measured signal. Consequently, it is imperative that a misfit function is robust regarding noise. As demonstrated by Engquist et al. (2016), the Wasserstein metric is substantially less sensitive to noise than the traditional L^2 norm.

The property again follows from the interpretation of W_2^2 as a transportation cost. Intuitively, noise added to the data will increase the distance $|T(x) - x|$ that mass moves at some points x, but it will also decrease this distance at other points. Thus, the overall effect of noise on the total transportation cost

$$\int_X f(x)|T(x) - x|^2 dx \tag{21}$$

will be negligible.

This is simplest to calculate in one dimension. For example, we can consider the setting from Engquist et al. (2016). Here, the data f and g are given on a grid with a total of N data points along each dimension. At each grid point, the difference $f - g$ is given by a random variable drawn from a uniform distribution of the form $U[-c, c]$ for some constant c. Regardless of the number of data points, noise of this type is expected to have a large effect on the L^2 distance,

$$\mathbb{E}\|f - g\|_{L^2} = \mathcal{O}(1). \tag{22}$$

Using the exact formula for the 1D optimal transport plan, we can also directly compute the expected value of the squared Wasserstein metric:

$$\mathbb{E}W_2^2(f, g) = \mathcal{O}\left(\frac{1}{N}\right). \tag{23}$$

Thus, even if the noise is very strong (with order-one amplitude), its effect on the misfit is negligible if there are a large number of data points.

Although there is no exact formula to exploit in higher dimensions, we can place a bound on the expected effects of noise by considering a sequence of 1D optimal transport problems. That is, we can produce a sequence of mappings $T_j(x)$, $j = 1, \ldots, n$ that optimally rearrange the mass along the jth dimension (see Figure 2). These 1D maps can again be expressed exactly. The resulting composite map

$$\tilde{T}(x) = T_n \circ T_{n-1} \circ \cdots \circ T_1(x) \tag{24}$$

will be mass preserving, but not optimal. As described by Engquist et al. (2016), this leads to the estimate

$$\begin{aligned} \mathbb{E}\tilde{W}(f, g) &= \mathbb{E}\int f(x)|x - T(x)|^2 dx \\ &\le \mathbb{E}\int f(x)|x - \tilde{T}(x)|^2 = \mathcal{O}\left(\frac{1}{N}\right). \end{aligned} \tag{25}$$

Thus, for typical seismic data, the effect of noise is expected to have a negligible effect on the behavior of the squared Wasserstein metric.

NUMERICAL SCHEME

In this section, we describe the numerical schemes we use to compute the W_2 misfit. We also explain the adjoint source that is needed for efficient inversion on geophysical data.

Data normalization

In optimal transport theory, there are two main requirements for signals f and g: positivity and mass balance. Because these are not expected for seismic signals, some data preprocessing is needed before we can implement Wasserstein-based FWI. In Engquist and Froese (2014) and Engquist et al. (2016), the signals were separated into positive and negative parts $f^+ = \max\{f, 0\}$, $f^- = \max\{-f, 0\}$ and scaled by the total mass $\langle f \rangle = \int_X f(x)dx$. Inversion was accomplished using the modified misfit function:

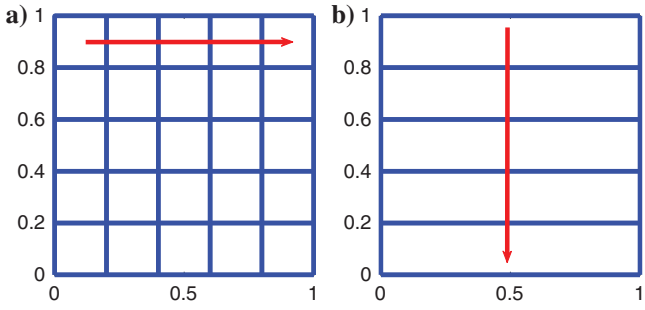

Figure 2. (a) The optimal map for each row: $T_x = T_i$ for $x_i < x \le x_{i+1}$ and (b) the optimal map in the y direction: T_y.

$$W_2^2\left(\frac{f^+}{\langle f^+\rangle},\frac{g^+}{\langle g^+\rangle}\right)+W_2^2\left(\frac{f^-}{\langle f^-\rangle},\frac{g^-}{\langle g^-\rangle}\right). \tag{26}$$

Although this approach preserves the desirable theoretical properties of convexity to shifts and noise insensitivity, it is not easy to combine with the adjoint-state method and more realistic examples. We require the scaling function to be differentiable so that it is easy to apply the chain rule when calculating the Fréchet derivative and it is also better suited for the Monge-Ampère solver.

There are other different ways to rescale the data sets so that they become positive. For example, we can square the data as $\tilde{f}=f^2$ or extract the envelope of the data. The convexity concerning shifts are preserved by these methods, but we have lost some information in the gradient. In the squaring case, the gradient of W_2 with respect to f is zero when f is zero, which can cause severe difficulties in recovering reflections. The envelope approach, on the other hand, loses important phase information.

In this paper, we propose normalization via a linear transformation and rescaling. We begin by selecting a constant c such that $f+c>0$ and $g+c>0$. In the experiments, c is chosen approximately 1.1 times $|g_{\min}|$. This constant is fixed in inversion. After shifting the signals to ensure positivity, we rescale so all signals share a common total mass. Thus, we obtain the modified data $\tilde{f}=P(f)$ and $\tilde{g}=P(g)$ where

$$P(f)\equiv\frac{f+c}{\langle f+c\rangle}. \tag{27}$$

This normalization has several advantages. First, the number and location of local maximum and minimum are maintained. In addition, it has high regularity, which is important for the adjoint-state method. The normalization function $P(f)$ does not change significantly from iteration to iteration because of the mean zero property of the data, which aids in convergence. There is, however, a serious concern in that this normalization results in a misfit function that is not convex for simple shifts (Figure 3a) even if the W_2 misfit is slightly better than that of L^2.

We use an example from Métivier et al. (2016c) to empirically demonstrate a convexity result in a higher dimensional model domain with the linear normalization proposed in this paper. The model velocity is increasing linearly in depth as $v(x,z)=v_0+\alpha z$, where v_0 is the starting velocity on the surface, α is the vertical gradient, and z is the depth. The model is 17 km in width and 3.5 km in depth. We place 681 receivers on the top with a 25 m fixed acquisition and one source in the top middle with a Ricker wavelet centered at 5 Hz.

The reference for (v_0,α) is (2 km/s, 0.7 s^{-1}), and we plot the misfit curves with $\alpha\in[0.4,1]$ and $v_0\in[1.75,2.25]$ on 41×45 grid in Figure 3b. It is globally convex with respect to two model variables. We compare the convexity of L^2 and W_2 in one variable when the value of the other variable is wrong (Figure 4). The L^2 results have local minima, whereas the curves for W_2 are convex.

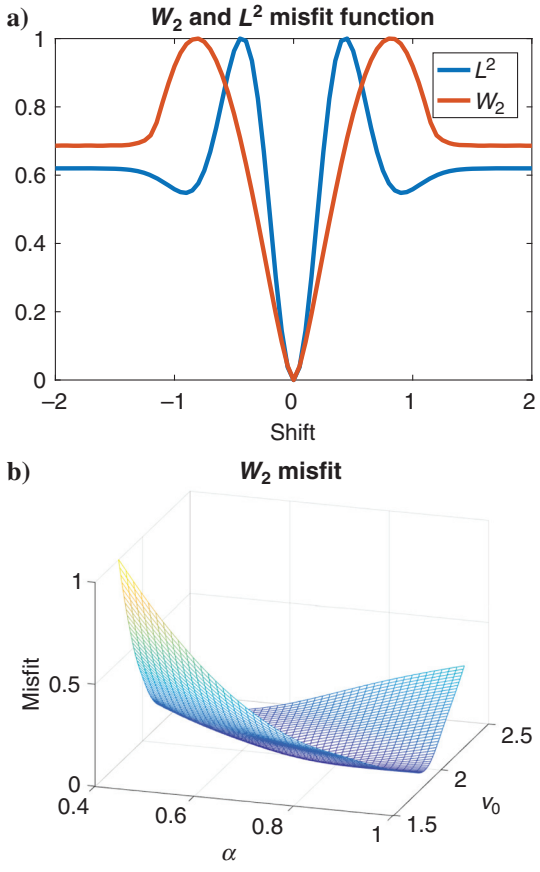

Figure 3. (a) The L^2 and W_2 misfits between a Ricker wavelet f and its shift $f(x-s)$ and (b) misfit sensitivity with respect to model parameters v_0 and α.

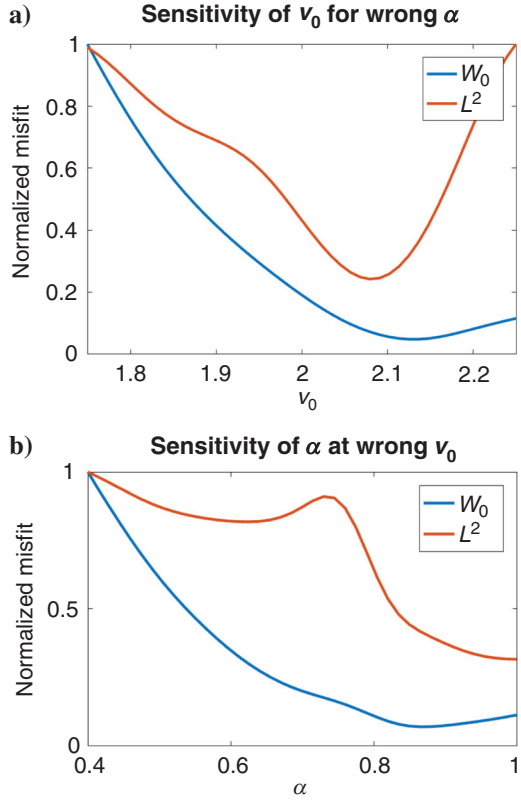

Figure 4. (a) Misfit sensitivity with respect to v_0 at $\alpha=0.535$ and (b) misfit sensitivity with respect to α at $v_0=1.8523$ km/s.

Our empirical experience is that this linear normalization works remarkably well on realistic examples, but we believe further research is desirable in solving this problem and increasing the understanding of convergence properties.

To simplify notation, we will hereafter use f and g denoting their normalized version $\tilde{f}$ and $\tilde{g}$ in equation 27.

Compare trace by trace: $W_2^2(f,g)$ in 1D

We first describe the scheme used for the 1D Wasserstein metric, which we use to compare the data trace by trace for an overall misfit:

$$d(f,g) = \sum_{r=1}^{R} W_2^2(f(x_r,t), g(x_r,t)), \tag{28}$$

where x_r denotes the receiver location.

Computation of the objective function

In this setting, if the last time record for a receiver is at T_0, we can use the exact formula (equation 9) to express the 1D quadratic Wasserstein metric as

$$W_2^2(f,g) = \int_0^{T_0} |t - G^{-1}(F(t))|^2 f(t)dt, \tag{29}$$

where F and G are the cumulative distribution functions for f and g, respectively, $F(t) = \int_0^t f, G(t) = \int_0^t g$.

This will be approximated in a discrete setting, that is, assuming that f and g are given at a discrete set of points $t = (t_0, t_1, \ldots, t_n)^T$ in the time domain. We compute F and G using numerical integration. For each value y, because G is monotone increasing, we can find t_n and t_{n+1} such that $G(t_n) < y \le G(t_{n+1})$ in $\mathcal{O}(\log(N))$ complexity by binary search and N is the number of data samples in each trace. For y in this range we can estimate $G^{-1}(y) = t_{n+1}$. Here, we will also do numerical interpolation between t_n and t_{n+1} for better accuracy.

Using FD matrices, we can express the discrete 1D quadratic Wasserstein metric as

$$d_1(f,g) = (t - G^{-1} \circ F(t))^T \mathrm{diag}(f)(t - G^{-1} \circ F(t))dt, \tag{30}$$

where $G^{-1} \circ F$ is the optimal map that transports f onto g.

After summing over all the traces, we obtain the final misfit between the synthetic data and observed data: $d(f,g) = \sum_{r=1}^{R} d_1(f_r, g_r)$. By exploiting the explicit solution for optimal transport on the real line, we can compute the misfit in $\mathcal{O}(N)$ complexity.

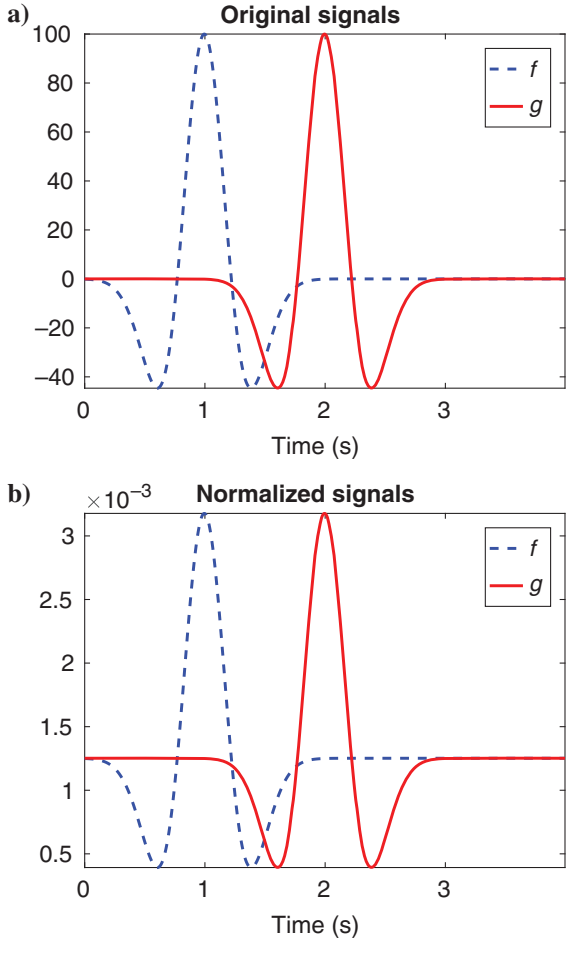

Figure 5. (a) Original synthetic signal f and observed signal g and (b) normalized synthetic signal f and observed signal g that satisfy the requirements of optimal transport.

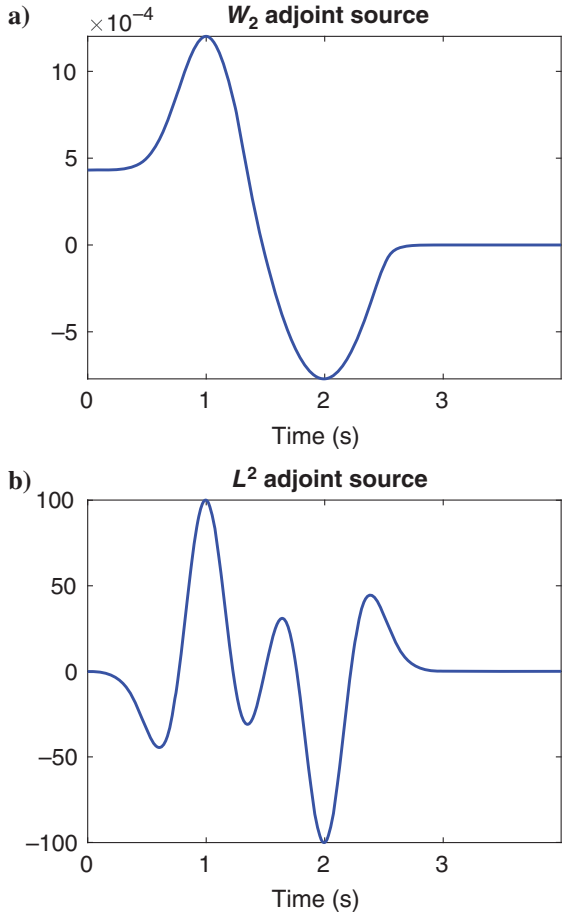

Figure 6. (a) Adjoint source of $W_2^2(f,g)$ with respect to f and (b) adjoint source of $L^2(f,g)$ with respect to f.

Computation of adjoint source

We also derive the Fréchet derivative of the misfit, which acts as the adjoint source in the adjoint-state method.

The first variation of the squared Wasserstein metric for the 1D case is

$$\delta d_1 = \left[U\mathrm{diag}\left(-2f(t)\left.\frac{dG^{-1}(y)}{dy}\right|_{F(t)} dt \right)(t - G^{-1}\circ F(t)) + \mathrm{diag}(t - G^{-1}\circ F(t))(t - G^{-1}\circ F(t)) \right]^T \delta f dt, \quad (31)$$

where U is the upper triangular matrix whose nonzero components are 1.

By the inverse function theorem, we have

$$\left.\frac{dG^{-1}(y)}{dy}\right|_{y=F(t)} = \frac{1}{\left.\frac{dG(s)}{ds}\right|_{s=G^{-1}\circ F(t)}} = \frac{1}{g(G^{-1}\circ F(t))}. \quad (32)$$

The adjoint source term for the discrete 1D quadratic Wasserstein metric can be computed as

$$\nabla d_1(t) = \left[U\mathrm{diag}\left(\frac{-2f(t)dt}{g(G^{-1}\circ F(t))} \right) + \mathrm{diag}(t - G^{-1}\circ F(t)) \right] (t - G^{-1}\circ F(t))dt. \quad (33)$$

One can refer to the Appendix for a step-by-step derivation of the continuous Frechet derivative (A-7).

Compare globally: $W_2^2(f, g)$ in higher dimensions

Second, we wish to examine the effects of comparing the data f and g globally via a single, higher dimensional optimal transportation computation.

Computation of the objective function

In this case, there is no simple exact formula for the Wasserstein metric. Instead, we will compute it via the solution of the Monge-Ampère equation:

$$\begin{cases} \det(D^2u(x)) = f(x)/g(\nabla u(x)) + \langle u \rangle, & x \in X \\ \nabla u(x)\cdot n = x\cdot n, & x \in \partial X \\ u \text{ is convex.} \end{cases} \quad (34)$$

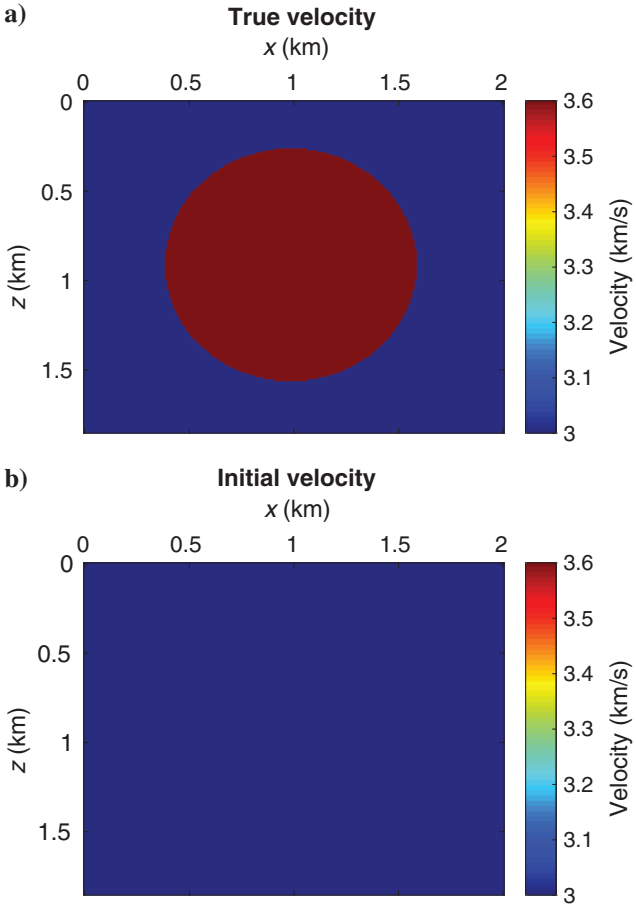

Figure 7. (a) True velocity and (b) initial velocity for the Camembert model.

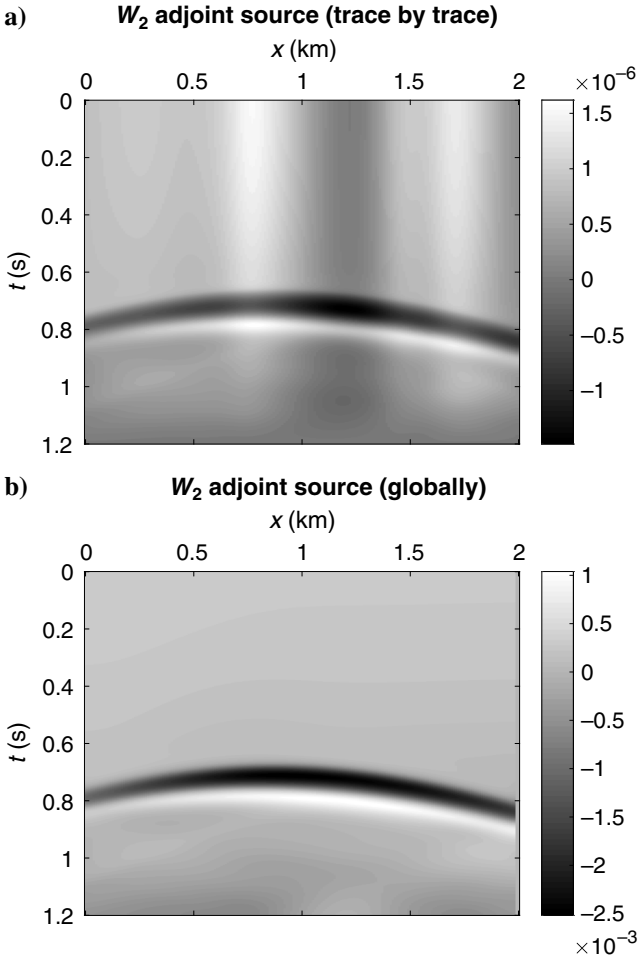

Figure 8. (a) Adjoint source of W_2 processed trace by trace and (b) adjoint source for global W_2 for the Camembert model.

The squared quadratic Wasserstein metric is then given by

$$W_2^2(f,g) = \int_X f(x)|x - \nabla u(x)|^2 dx. \tag{35}$$

We solve the Monge-Ampère equation numerically using an almost-monotone FDM relying on the following reformulation of the Monge-Ampère operator, which automatically enforces the convexity constraint (Froese, 2012).

$$\det{}^+(D^2u) = \min_{\{v_1,v_2\}\in V} \{\max\{u_{v_1,v_1},0\}\max\{u_{v_2,v_2},0\} + \min\{u_{v_1,v_1},0\} + \min\{u_{v_2,v_2},0\}\}, \tag{36}$$

where V is the set of all orthonormal bases for $\mathbb{R}^2$.

Equation 36 can be discretized by computing the minimum over finitely many directions $\{\nu_1,\nu_2\}$, which may require the use of a wide stencil. For simplicity and brevity, we describe a low-order version of the scheme and refer to Froese (2012) and Froese and Oberman (2013) for complete details. In practice, this simplified scheme is sufficient for obtaining accurate inversion results.

The scheme relies on the finite-difference operators

$$\begin{aligned}
[\mathcal{D}_{x_1x_1}u]_{ij} &= \frac{1}{dx^2}(u_{i+1,j} + u_{i-1,j} - 2u_{i,j}),\\
[\mathcal{D}_{x_2x_2}u]_{ij} &= \frac{1}{dx^2}(u_{i,j+1} + u_{i,j-1} - 2u_{i,j}),\\
[\mathcal{D}_{x_1x_2}u]_{ij} &= \frac{1}{4dx^2}(u_{i+1,j+1} + u_{i-1,j-1} - u_{i+1,j-1} - u_{i-1,j+1}),\\
[\mathcal{D}_{x_1}u]_{ij} &= \frac{1}{2dx}(u_{i+1,j} - u_{i-1,j}),\\
[\mathcal{D}_{x_2}u]_{ij} &= \frac{1}{2dx}(u_{i,j+1} - u_{i,j-1}),\\
[\mathcal{D}_{vv}u]_{ij} &= \frac{1}{2dx^2}(u_{i+1,j+1} + u_{i-1,j-1} - 2u_{i,j}),\\
[\mathcal{D}_{v^\perp v^\perp}u]_{ij} &= \frac{1}{2dx^2}(u_{i+1,j-1} + u_{i+1,j-1} - 2u_{i,j}),\\
[\mathcal{D}_{v}u]_{ij} &= \frac{1}{2\sqrt{2}dx}(u_{i+1,j+1} - u_{i-1,j-1}),
\end{aligned} \tag{37}$$

In the low-order version of the scheme, the minimum in equation 36 is approximated using only two possible values. The first uses directions aligning with the grid axes:

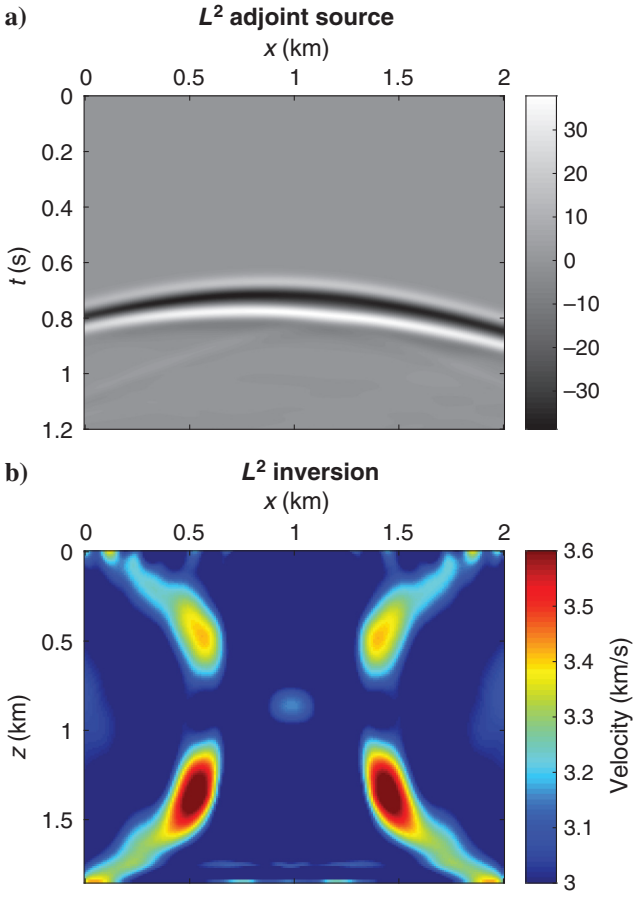

Figure 9. (a) Adjoint source for L^2 for the Camembert model and (b) inversion result using L^2 as misfit function

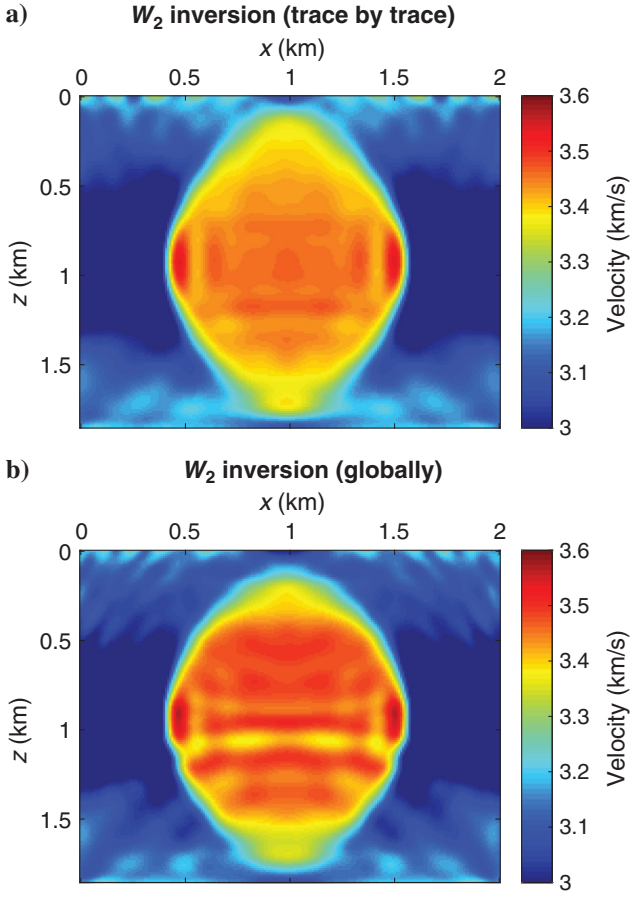

Figure 10. (a) Inversion result for W_2 processed trace by trace and (b) inversion result for global W_2 for the Camembert model.

$$MA_1[u] = \max\{\mathcal{D}_{x_1x_1}u, \delta\}\max\{\mathcal{D}_{x_2x_2}u, \delta\} + \min\{\mathcal{D}_{x_1x_1}u, \delta\} + \min\{\mathcal{D}_{x_2x_2}u, \delta\} - f/g(\mathcal{D}_{x_1}u, \mathcal{D}_{x_2}u) - u_0. \quad (38)$$

Here, dx is the resolution of the grid, δ (bounded below by $K\Delta x/2$) is a small parameter that bounds the second derivatives away from zero, u_0 is the solution value at a fixed point in the domain, and K is the Lipschitz constant in the y-variable of $f(x)/g(y)$.

For the second value, we rotate the axes to align with the corner points in the stencil, which leads to

$$MA_2[u] = \max\{\mathcal{D}_{vv}u, \delta\}\max\{\mathcal{D}_{v^\perp v^\perp}u, \delta\} + \min\{\mathcal{D}_{vv}u, \delta\} + \min\{\mathcal{D}_{v^\perp v^\perp}u, \delta\} - f/g\left(\frac{1}{\sqrt{2}}(\mathcal{D}_v u + \mathcal{D}_{v^\perp}u), \frac{1}{\sqrt{2}}(\mathcal{D}_v u - \mathcal{D}_{v^\perp}u)\right) - u_0. \quad (39)$$

Then, the monotone approximation of the Monge-Ampère equation is

$$M_M[u] \equiv -\min\{MA_1[u], MA_2[u]\} = 0. \quad (40)$$

We also define a second-order approximation, obtained from a standard centered difference discretization,

$$M_N[u] \equiv -((\mathcal{D}_{x_1x_1}u)(\mathcal{D}_{x_2x_2}u) - (\mathcal{D}_{x_1x_2}u^2)) + f/g(\mathcal{D}_{x_1}u, \mathcal{D}_{x_2}u) + u_0 = 0. \quad (41)$$

These are combined into an almost-monotone approximation of the form

$$M_F[u] \equiv M_M[u] + \varepsilon S\left(\frac{M_N[u] - M_M[u]}{\varepsilon}\right), \quad (42)$$

where ε is a small parameter and the filter S is given by

$$S(x) = \begin{cases} x & |x| \le 1 \\ 0 & |x| \ge 2 \\ -x+2 & 1 \le x \le 2 \\ -x-2 & -2 \le x \le -1. \end{cases} \quad (43)$$

The Neumann boundary condition is implemented using standard one-sided differences. As described by Froese (2012) and Engquist et al. (2016), the (formal) Jacobian $\nabla M_F[u]$ of the scheme can be obtained exactly. In particular, it is known to be sparse and diagonally dominant.

This FD approximation effectively replaces the Monge-Ampère equation with a large system of nonlinear algebraic equations,

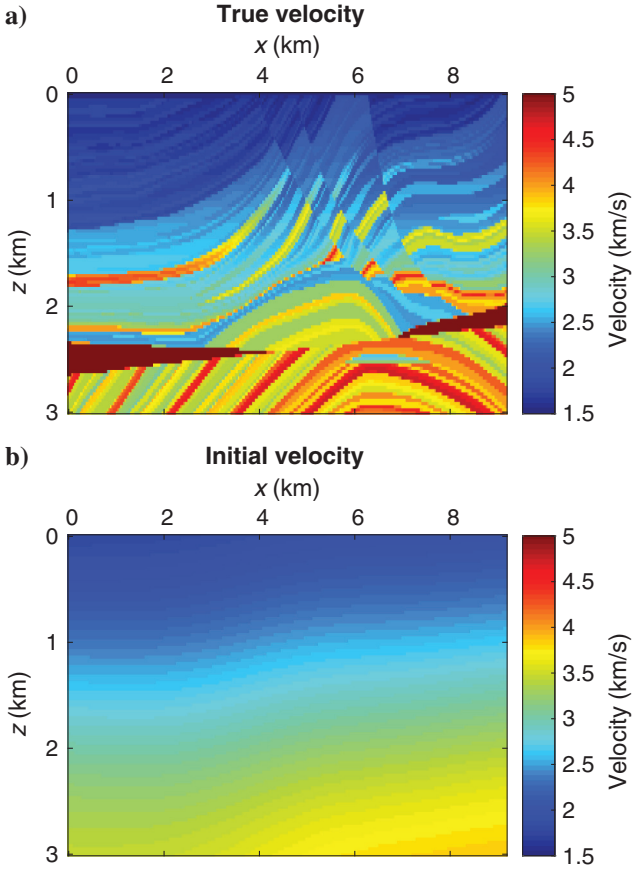

Figure 11. (a) True velocity and (b) inital velocity for the true Marmousi model.

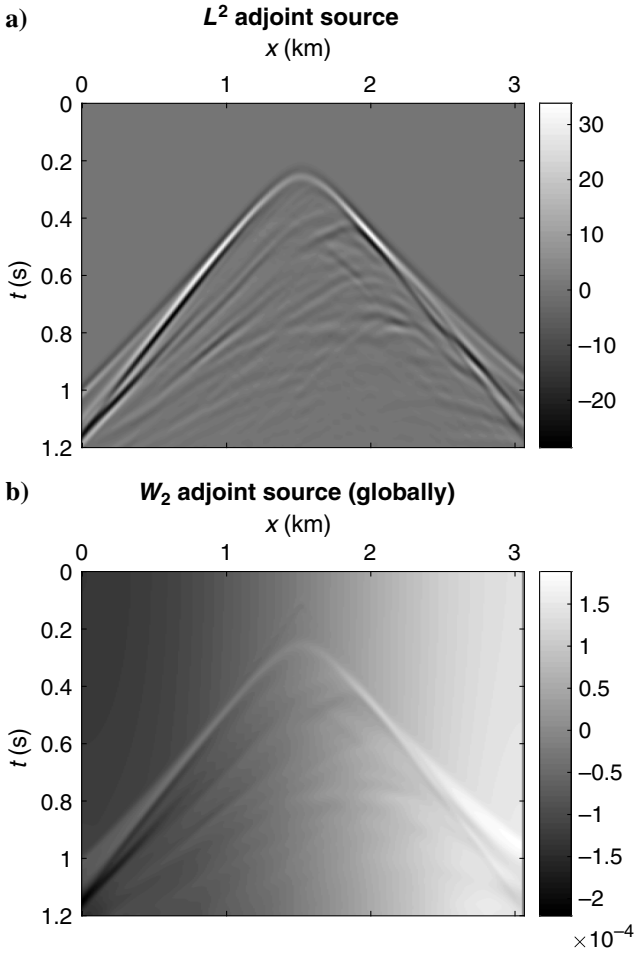

Figure 12. Adjoint source of (a) L^2 and (b) global W_2 for the scaled Marmousi model.

which can be solved using Newton's method. Computing the Newton updates requires inverting sparse M matrices, which can be done efficiently. The number of Newton iterations required depends weakly on the smoothness of the data and the resulting solution u. In numerical experiments carried out by Froese (2012), the total computational complexity required to solve the Monge-Ampère equation varied from $\mathcal{O}(N)$ to $\mathcal{O}(N^{1.3})$ where N was the total number of grid points.

Once the discrete solution u_h is computed, the squared Wasserstein metric is approximated via

$$W_2^2(f,g) \approx \sum_{j=1}^{n} (x_j - D_{x_j} u_h)^T \mathrm{diag}(f)(x_j - D_{x_j} u_h). \quad (44)$$

Computation of adjoint source

In Engquist et al. (2016), we consider the linearization of the discretized version of the Wasserstein metric. Using the FD matrices introduced, we can express the discrete Wasserstein metric as

$$d(f) = \sum_{j=1}^{n} (x_j - D_{x_j} u_f)^T \mathrm{diag}(f)(x_j - D_{x_j} u_f), \quad (45)$$

where n is the data dimension; the potential u_f satisfies the discrete Monge-Ampère equation

$$M[u_f] = 0. \quad (46)$$

The first variation of the squared Wasserstein metric is

$$\delta d = -2 \sum_{j=1}^{n} (D_{x_j} \delta u)^T \mathrm{diag}(f)(x_j - D_{x_j} u_f) + \sum_{j=1}^{n} (x_j - D_{x_j} u_f)^T \mathrm{diag}(\delta f)(x_j - D_{x_j} u_f). \quad (47)$$

Linearizing the Monge-Ampère equation, we have to the first order

$$\nabla M_F[u_f] \delta u = \delta f. \quad (48)$$

Here, ∇M_F is the (formal) Jacobian of the discrete Monge-Ampère equation, which is already being inverted in the process of solving the Monge-Ampère equation via Newton's method. Then, the gradient of the discrete squared Wasserstein metric can be expressed as

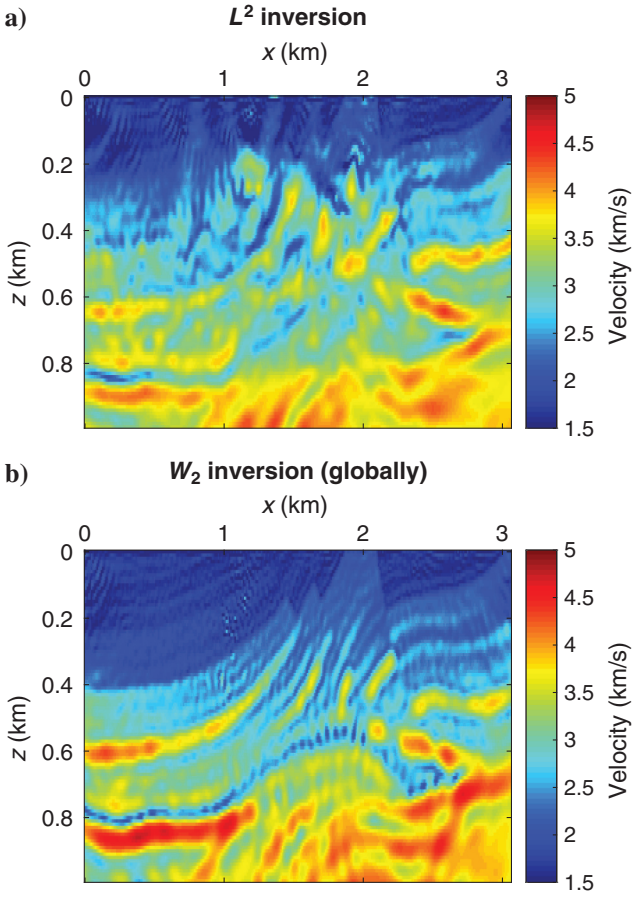

Figure 13. Inversion results of (a) L^2 and (b) global W_2 for the scaled Marmousi model.

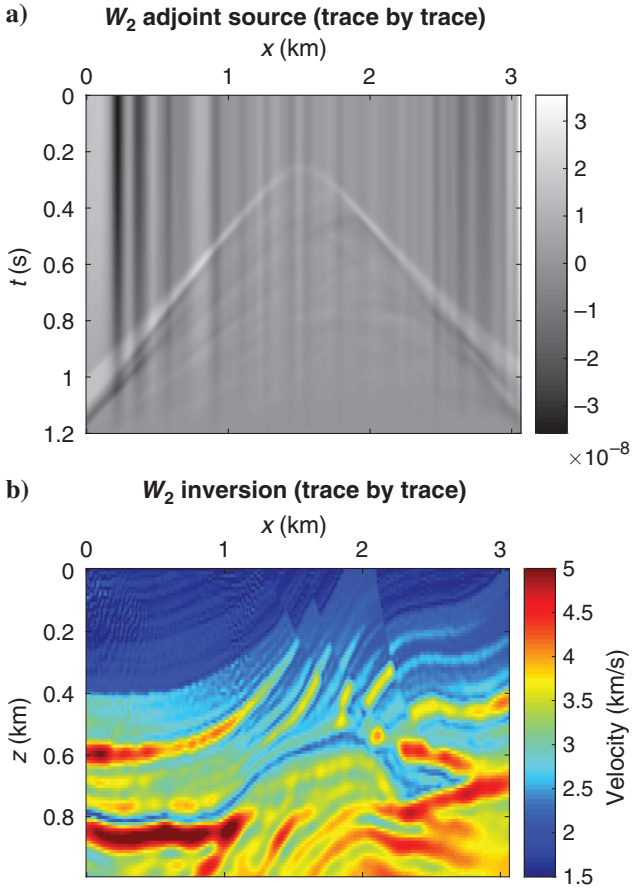

Figure 14. (a) Adjoint source of trace-by-trace W_2 and (b) the inversion result for the scaled Marmousi model.

$$\nabla d = \sum_{j=1}^{n} [-2\nabla M_F^{-1}[u_f]^T D_{x_j}^T \mathrm{diag}(f) + \mathrm{diag}(x_j - D_{x_j} u_f)](x_j - D_{x_j} u_f). \quad (49)$$

Notice that once the Monge-Ampère equation itself has been solved, this gradient is easy to compute because it only requires the inversion of a single matrix that is already being inverted as a part of the solution of the Monge-Ampère equation.

Theorem 3. *(Convergence to viscosity solution [Froese, 2012, theorem 4.4]). Let the Monge-Ampère equation (34) have a unique viscosity solution, and let $\boldsymbol{g} > 0$ be Lipschitz continuous on $\mathbb{R}^d$. Then the solutions of the scheme (42) converge to the viscosity solution of (34) with a formal discretization error of $\mathcal{O}(\boldsymbol{L}\boldsymbol{h}^2)$ where $\boldsymbol{L}$ is the Lipschitz constant of $\boldsymbol{g}$ and $\boldsymbol{h}$ is the resolution of the grid.*

We remark that the numerical error of the solver is affected by the Lipschitz constant of function g as well as the grid spacing. In the discrete setting, we achieve good accuracy if g is highly resolved data input such that h is small.

COMPUTATIONAL RESULTS

In this section, we apply the quadratic Wasserstein metric W_2 to several synthetic data models. We provide results for two approaches to using W_2: trace-by-trace comparison and using the entire data sets as objects. These are compared with results produced by using the least-squares norm L^2 to measure the misfit.

Due to limitations of current Monge-Ampère solvers, we will present global W_2-based FWI on smaller scale models with L^2 and trace-by-trace W_2 results for comparison (the third and fifth test). We also show experiments of the trace-by-trace approach on the true or larger scale benchmark to demonstrate its robustness. In the inversion process, we avoid the use of techniques such as adding regularization and smoothing the gradient to see the effectiveness of this new misfit.

1D case study

We begin with a simple test case from Engquist and Froese (2014) and focus on two Ricker wavelet signals, one a time shift of the other. We regard these two signals as observed data $g(t)$ and synthetic data $f(t;s) = g(t-s)$ as shown in Figure 5. This is a case in which the quadratic Wasserstein metric W_2 is applied to 1D signals.

The adjoint source for L^2 and W_2 misfits between these two signals is very different as shown in Figure 6. The adjoint source for W_2 is very similar to the adjoint source of the KR norm applied on this 1D case; see Figure 4 of Métivier et al. (2016c) for more details. This illustrates the character of optimal transport-based

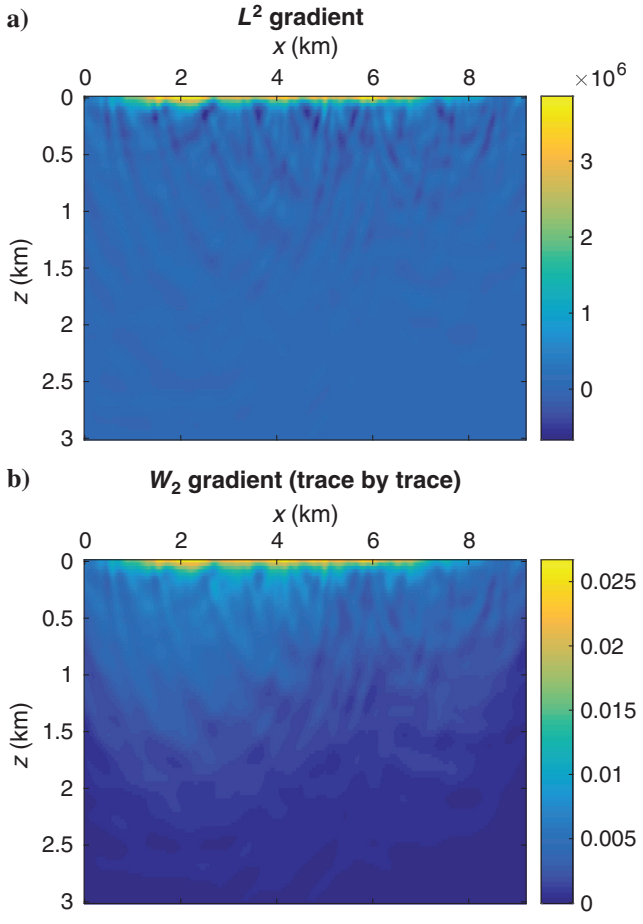

Figure 15. The gradient in the first iteration of (a) L^2 and (b) trace-by-trace W_2 inversion for the true Marmousi model.

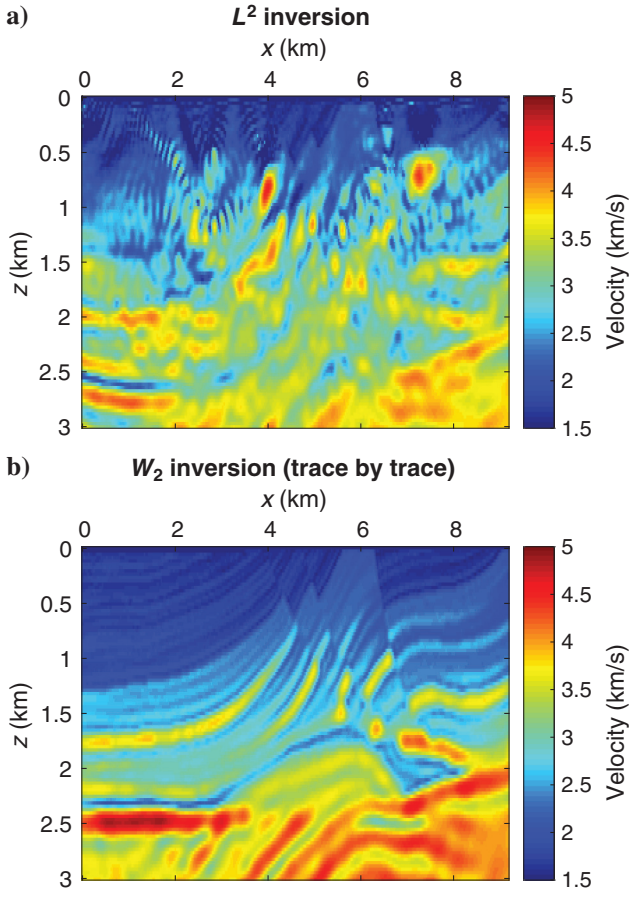

Figure 16. Inversion results of (a) L^2 and (b) trace-by-trace W_2 for the true Marmousi model.

misfit functions, which shift mass from the synthetic data to observed data in a way that corrects the phase difference between f and g. The L^2 norm, on the other hand, only seeks to correct the amplitude difference, which is the origin of the cycle skipping.

We observe that the adjoint source of W_2 is smoother than the adjoint source of the KR norm (Figure 4 in Métivier et al., 2016c) and it has no discontinuous component. The smoothness of the adjoint source is ideal for quasi-Newton methods, e.g., the L-BFGS algorithm, which is designed to minimize smooth functions. It is also numerically more stable to back propagate in time to compute the gradient updates.

Camembert model

FWI with least-squares norm L^2 minimization (Tarantola and Valette, 1982) is effective when the initial model is close to the true model. However, if the initial model is far from the true model, the L^2 misfit may suffer from local minima because it uses a point-by-point comparison that records the oscillatory and nonlinear features of the data. The difficulty of local minima in seismic inversion was demonstrated with the so-called Camembert example (Gauthier et al., 1986).

We repeat the experiments with three different misfit functions for FWI: W_2 applied trace by trace, W_2 applied globally, and the traditional L^2 least-squares norm. The comparison among these three different misfit functions illustrate the advantages of the quadratic Wasserstein metric W_2.

We set the Camembert-shaped inclusion as a circle with radius 0.6 km located in the center of the rectangular velocity model. The velocity is 3.6 km/s inside and 3 km/s outside the circle as shown in Figure 7a. The inversion starts from an initial model with homogeneous velocity 3 km/s everywhere as shown in Figure 7b. We place 11 equally spaced sources on the top at 50 m depth and 201 receivers on the bottom with 10 m fixed acquisition. The discretization of the forward wave equation is 10 m in the x- and z-directions and 10 ms in time. The source is a Ricker wavelet with a peak frequency of 10 Hz, and a high-pass filter is applied to remove the frequency components from 0 to 2 Hz.

Figures 8a, 8b, and 9a show the adjoint sources of trace-by-trace W_2, global W_2, and the L^2 misfit functions, respectively. Figure 9b shows the inversion result obtained with the traditional L^2 least-squares norm. It converges to a local minimum after 100 iterations using the L-BFGS optimization method. The inversion using the 1D optimal transport to calculate the misfit trace by trace successfully recovers the shape of the inclusion (Figure 10a). Because the data are two dimension (in the time and spatial domains), an alternative approach is to find the optimal transport map between these two data sets instead of slicing them into traces. Figure 10b shows the final inversion result respectively of comparing the two data sets via a global optimal map. Both approaches converge to reasonably good results in 10 iterations using the L-BFGS optimization method.

Although Figure 9a looks similar in shape to Figure 8 at first glance, the adjoint source of W_2-based misfit functions only have negative-positive components (the "black-white" curves in Figure 8) whereas the adjoint source for L^2 has positive-negative-positive components ("white-black-white" curves in Figure 9a). Thus, it provides L^2-based inversion with an incorrect direction updating the velocity model, which leads to local minima and cycle skipping. This is an example of the global W_2 inversion result (Figure 10b) being more accurate than the trace-by-trace results (Figure 10a) regarding the L^2 error in the computed velocity profile.

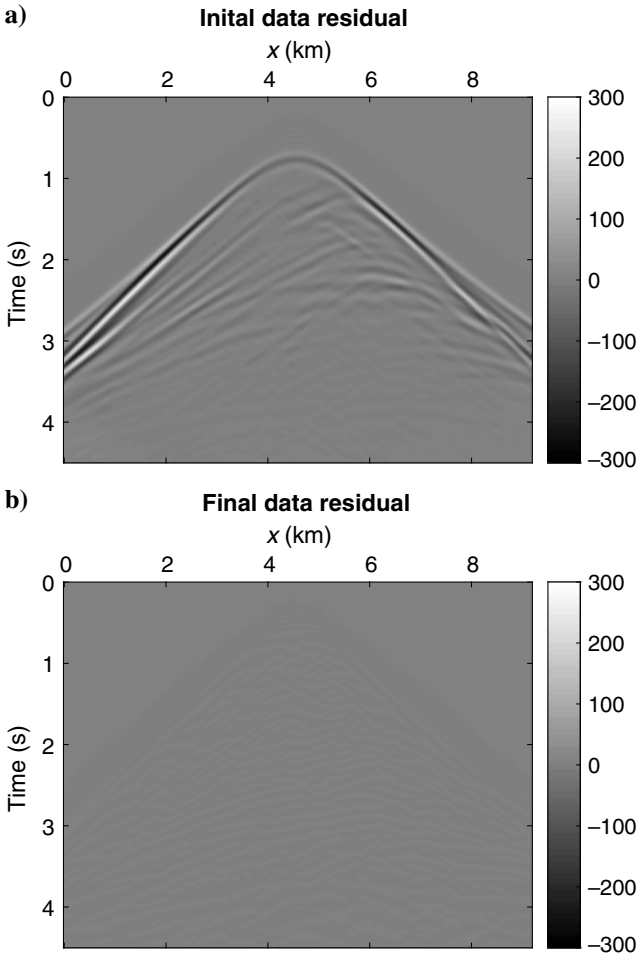

Figure 17. (a) The difference of data to be fit and the prediction with the initial model and (b) the final data residual of trace-by-trace W_2 for the true Marmousi model.

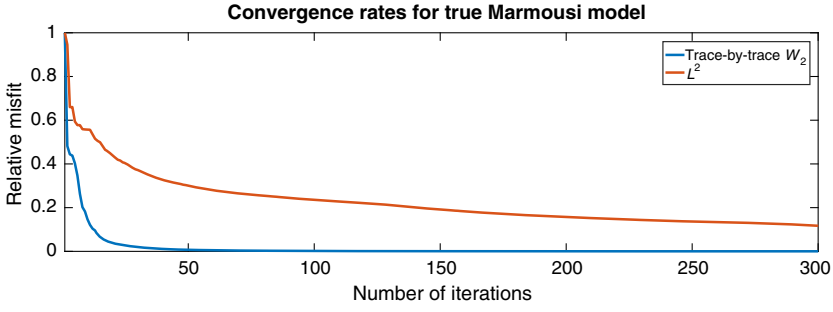

Figure 18. The convergence curves for trace-by-trace W_2- and L^2-based inversion of the true Marmousi model.

Scaled Marmousi model with global W_2 misfit computation

Our second 2D synthetic experiment is the Marmousi model. First, we use a scaled Marmousi model to compare the inversion between global W_2 and the conventional L^2 misfit function. Figure 11a shows the P-wave velocity of the true Marmousi model, but in this experiment, we use a scaled model that is 1 km deep and 3 km wide. The inversion starts from an initial model that is the true velocity smoothed by a Gaussian filter with a deviation of 40, which is highly smoothed and far from the true model (a scaled version of Figure 11b). We place 11 evenly spaced sources on top at the 50 m depth and 307 receivers on top at the same depth with a 10 m fixed acquisition. The discretization of the forward wave equation is 10 m in the x- and z-directions and 10 ms in time. The source is a Ricker wavelet with a peak frequency of 15 Hz, and a band-pass filter is applied to remove the frequency components from 0 to 2 Hz.

We compute the W_2 misfit via a global optimal map between the entire 2D data sets by solving the Monge-Ampère equation. Figures 12b and 13b show the adjoint source and final inversion results, respectively. Inversions are terminated after 200 L-BFGS iterations. Figure 13a shows the inversion result using the traditional L^2 least-squares method after 200 L-BFGS iterations. The inversion result of global W_2 avoids the problem of local minima suffered by the conventional L^2 metric, whose result demonstrates spurious high-frequency artifacts due to a point-by-point comparison of amplitude. Figure 14a and 14b shows the adjoint source and final inversion results, respectively, for the trace-by-trace W_2 metric of the scaled Marmousi model. Trace-by-trace result has better resolution and less noise than the global approach. We will further compare these two in the next section.

True Marmousi model with trace-by-trace W_2 misfit computation

The next experiment is to invert true Marmousi model by the conventional L^2 and the trace-by-trace W_2 misfit. Figure 11a shows the P-wave velocity of the true Marmousi model, which is 3 km in depth and 9 km in width. The inversion starts from an initial model that is the true velocity smoothed by a Gaussian filter with a deviation of 40 (Figure 11b). We place 11 evenly spaced sources on top at 150 m depth in the water layer and 307 receivers on top at the same depth with a 30 m fixed acquisition. The discretization of the forward wave equation is 30 m in the x- and z-directions and 30 ms in time. The source is a Ricker wavelet with a peak frequency of 5 Hz, and a high-pass filter is applied to remove the frequency components from 0 to 2 Hz.

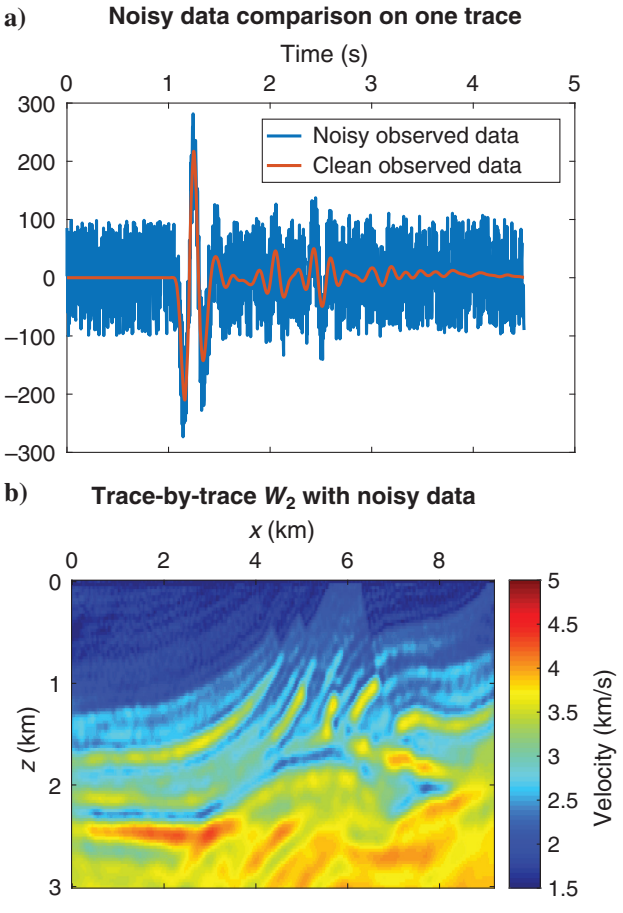

Figure 19. (a) Noisy and clean data and (b) inversion result with noisy data.

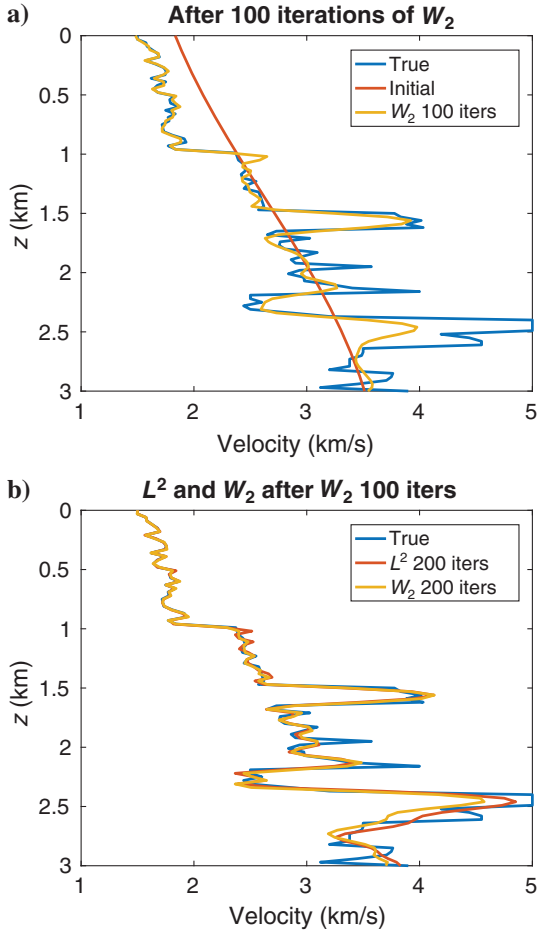

Figure 20. (a) Vertical velocity profiles after 100 iterations of trace-by-trace W_2 and (b) vertical velocity profiles of 200 iterations L^2- and W_2-based inversion starting from (a).

We compute the W_2 misfit trace by trace. For each receiver, we first normalize the data sets and then solve the optimal transport problem in 1D. With the explicit formula, the computation time is close to L^2. The final adjoint source $(dW_2^2(f,g))/df$ is a combination of the Fréchet derivative $(dW_2^2(f(x_r),g(x_r)))/(df(x_r))$ of all the receivers. Figure 15a and 15b shows the gradients in the first iteration of two misfits, respectively.

Starting from a highly smoothed initial model, in the first iteration W_2 already focuses on the "peak" of the Marmousi model as seen from Figure 15b. The darker area in the gradient matches many features in the velocity model (Figure 11a). However, the gradient of L^2 is quite uniform contrary to the model features. Inversions are terminated after 300 L-BFGS iterations. Figure 16a shows the inversion result using the traditional L^2 least-squares method, and Figure 16b shows the final result using the trace-by-trace W_2 misfit function. Again, the result of the L^2 metric has spurious high-frequency artifacts, whereas W_2 correctly inverts most details in the true model. The data residuals before and after trace-by-trace W_2-based FWI are presented in Figure 17. The convergence curves in Figure 18 show that W_2 reduces the relative misfit to 0.1 in 20 iterations, whereas L^2 converges slowly to a local minimum.

Inversion with the noisy data

One of the ideal properties of the quadratic Wasserstein metric is the insensitivity to noise (Engquist et al., 2016). We repeat the previous experiment with a noisy reference by adding a uniform random iid noise to the data from the true velocity (Figure 19a). The signal-to-noise ratio (S/N) is −3.47dB. In optimal transport, the effect of noise is, in theory, negligible due to the strong cancellation between the nearby positive and negative noise.

All the settings remain the same as the previous experiment except the observed data. After 96 iterations, the optimization converges to a velocity presented in Figure 19b. Although the result has lower resolution than Figure 16b, it still recovers most features of the Marmousi model correctly. When the noise power is much larger than the signal power, the quadratic Wasserstein metric still converges reasonably well, which again demonstrates its insensitivity to noise.

L^2-based FWI starting from W_2 enhanced initial model

Next, we perform FWI by first using the W_2 norm to overcome cycle skipping, then using the L^2-norm to increase the resolution. Starting from the initial model (Figure 11b) W_2-based FWI recovers most features greater than 2 km correctly after 100 iterations (see Figure 20a). We then start from this model and run another 200 iterations of L^2- and W_2-based inversion to check their performance

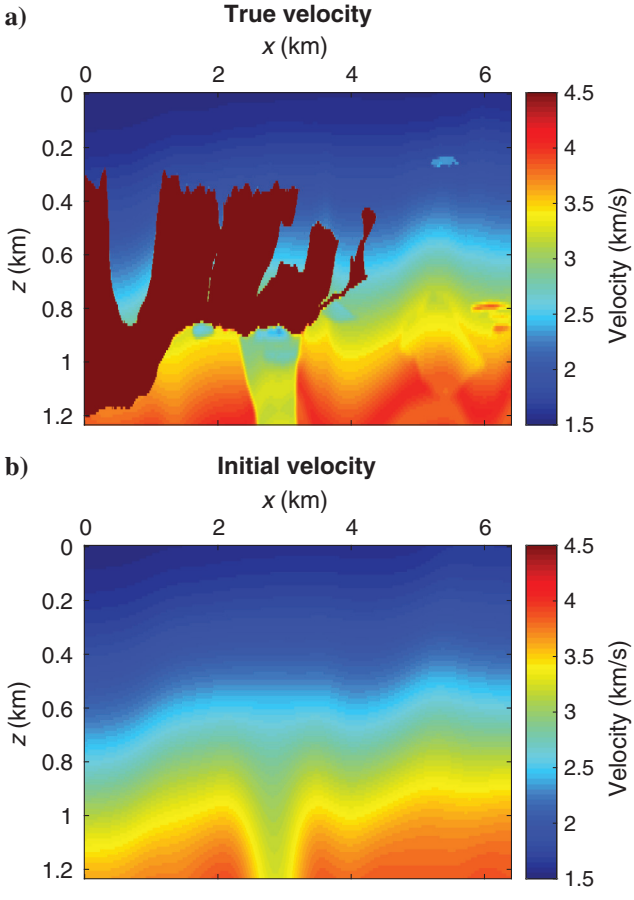

Figure 21. (a) True velocity and (b) inital velocity for the BP model.

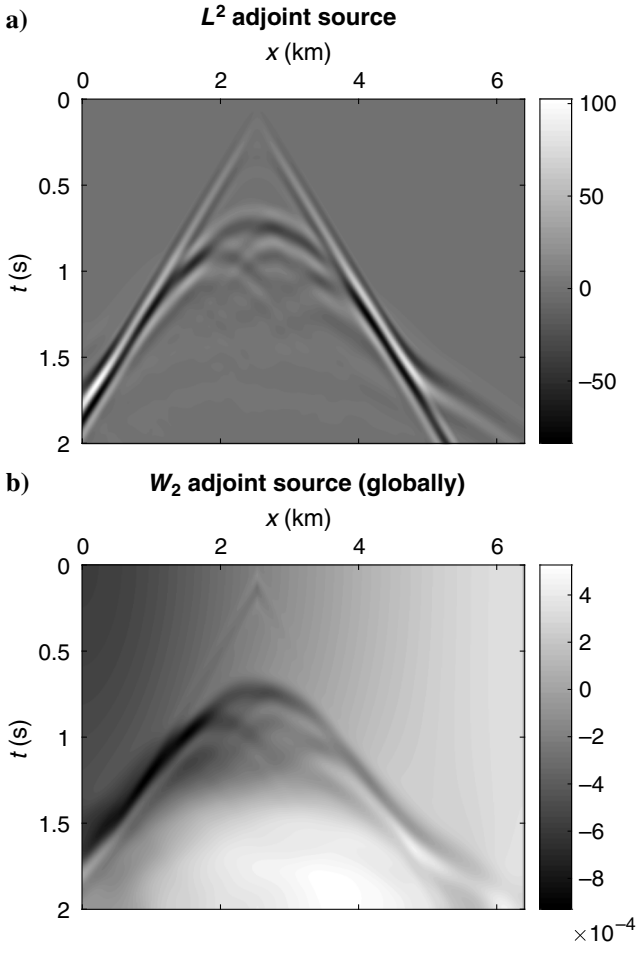

Figure 22. Adjoint sources of (a) the L^2 for the BP model and (b) the global W_2.

on increasing resolution (Figure 20b). The L^2 and W_2 iterations recover most features of the Marmousi model (Figure 11a) correctly. The L^2-norm is slightly better than W_2 in accuracy for the deeper part, but its L^2 relative error of the velocity (0.0057) is larger than the error from the W_2 misfit (0.0027). One can use the trace-by-trace W_2-norm to build a good starting model and later switch to L^2 norm to recover the high-wavenumber components.

2004 BP model with global W_2 misfit computation

For this experiment, we compare global W_2 and conventional L^2 as misfit functions for a modified BP 2004 model (Figure 21a). Part of the model is representative of the complex geology in the deep-water Gulf of Mexico. The main challenges in this area are related to obtaining a precise delineation of the salt and recover information on the subsalt velocity variations (Billette and Brandsberg-Dahl, 2005). The inversion starts from an initial model with a smoothed background without the salt (Figure 21b). We put 11 equally spaced sources on top at a 50 m depth and 641 receivers on top at a 50 m depth with a 10 m fixed acquisition. The discretization of the forward wave equation is 10 m in the x- and z-directions and 10 ms in time. The source is a Ricker wavelet with a peak frequency of 5 Hz, and a band-pass filter is applied to keep the frequency components from 3 to 9 Hz. The total acquisition time is restricted to 2 s to focus on recovering the upper portion of the salt structure.

As before, we solve the Monge-Ampère equation numerically to compute the global W_2 misfit. Figure 22a and 22b shows the adjoint source for two misfit functions. Inversions are stopped after 100 L-BFGS iterations. Figure 23b shows the inversion result for W_2, which recovered the top salt reasonably well. The L^2 metric (Figure 23a), on the other hand, converged to a model that has a low-velocity anomaly immediately beneath the top salt, which is typical of the cycle skipping commonly encountered in FWI. Figure 24a and 24b shows the adjoint source and final inversion results, respectively, for the trace-by-trace W_2 metric applied to the scaled BP model. Trace-by-trace result has less noise and more accuracy on the salt body than the global approach in Figure 23b. We will discuss this issue in detail in the next section.

2004 BP model with trace-by-trace W_2 misfit computation

In our last experiment, we compare trace-by-trace W_2 and conventional L^2 as misfit functions for another modified BP 2004 model. Different from the previous experiment, the model is much larger as 6 km deep and 16 km wide (Figure 25a), similar to BP example in Métivier et al. (2016b). The inversion starts from an initial model with smoothed background without the salt (Figure 25b). We put 11 equally spaced sources on top at 250 m depth in the water layer and 321 receivers on top at the same depth with a

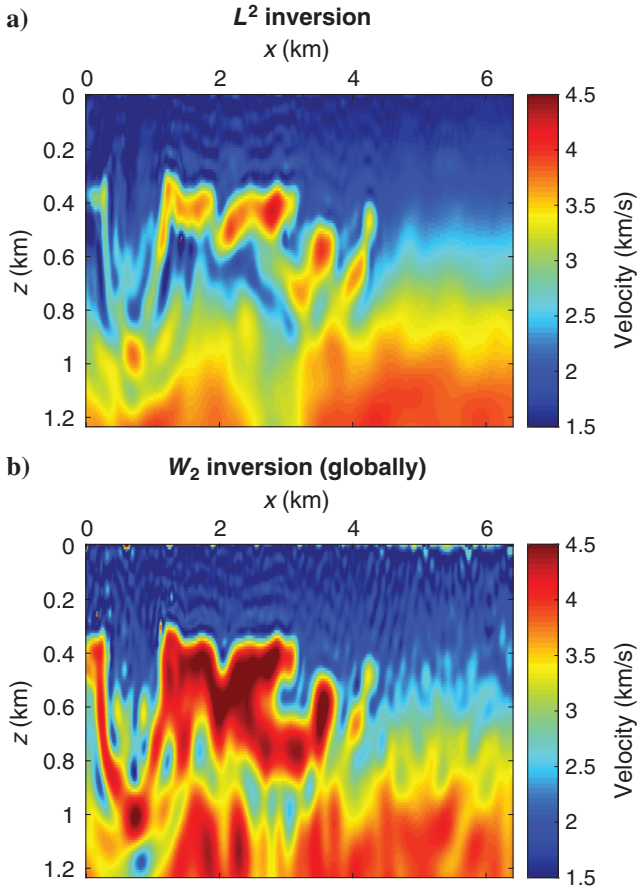

Figure 23. Inversion results (a) for L^2 and (b) for global W_2 for the BP model.

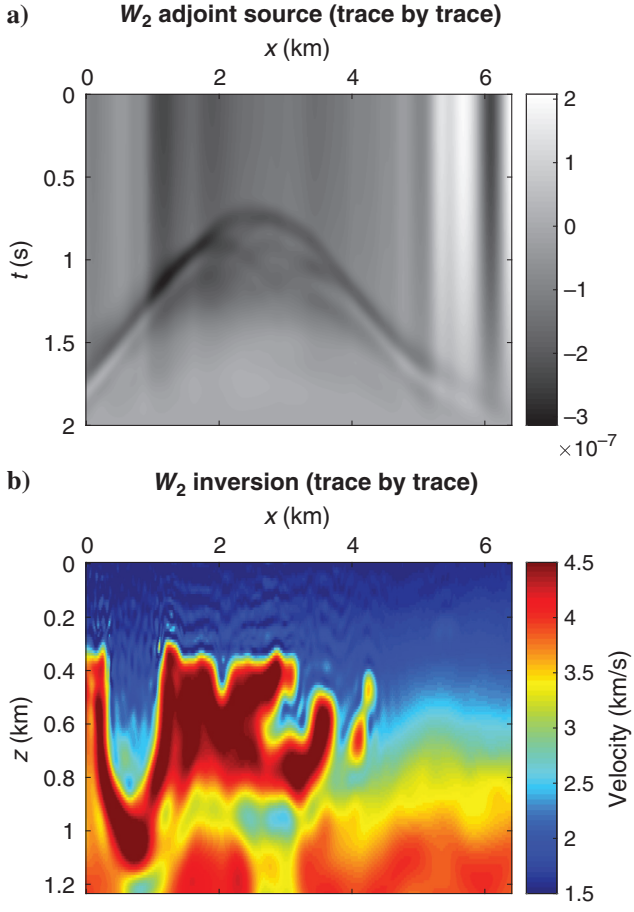

Figure 24. (a) Adjoint source of trace-by-trace W_2 and (b) the inversion result for the scaled BP model.

50 m fixed acquisition. The discretization of the forward wave equation is 50 m in the x- and z-directions and 50 ms in time. The source is a Ricker wavelet with a peak frequency of 5 Hz, and a band-pass filter is applied to keep a 3–9 Hz frequency. The total acquisition time is 10 s.

Again we compute the W_2 misfit trace by trace by solving the optimal transport problem in 1D as the Marmousi model and the Camembert model. Figure 26a and 26b shows the gradient in the first iteration of inversion using two misfit functions, respectively.

Starting from a smoothed initial model without the salt, in the first iteration W_2 inversion concentrates on the upper salt of the BP model (Figure 26b). The darker area in the gradient matches the salt part in the velocity model (Figure 25a). However, the gradient of L^2 is not very informative. Inversions are terminated after 300 L-BFGS iterations. Inversion with trace-by-trace W_2 misfit successfully constructed the shape of the salt bodies (Figure 27b), whereas FWI with the conventional L^2 failed to recover boundaries of the salt bodies as shown in Figure 27a. The data residuals before and after trace-by-trace W_2-based FWI are presented in Figure 28. Trace-by-trace W_2 reduces the relative misfit to 0.1 in 20 iterations, whereas L^2 converges slowly to a local minimum (Figure 29).

DISCUSSION ON TWO WAYS OF USING W_2

The computational complexity of performing 1D optimal transport is extremely low compared with the cost of solving the Monge-Ampère equation, which treats the synthetic and observed data as two objects and solves a 2D optimal transport problem. From observation of the running time in our experiments, inversion with the trace-by-trace W_2 misfit requires less than 1.1 times the run time required by inversion with the simple (and ineffective) L^2 misfit. Inversion using global W_2 comparison works for smaller scale models. It is more expensive because in each iteration we solve the Monge-Ampère equation numerically to compute the misfit. The total inversion takes 3–4 times the run time of the FWI with L^2 misfit in the experiments.

Figures 14a and 24a indicate one disadvantage of using W_2 trace by trace. The nonphysical variations in amplitude and nonuniform background of the adjoint source are caused by the fact that we rescale the data trace by trace to satisfy positivity and conservation of mass. On the one hand, this may lead to nonuniform contributions of data misfits to the velocity update during the inversion process. Therefore, more careful treatment of the scaling in the trace-by-trace scheme may improve the convergence result being demonstrated in this study. On the other hand, in the experiments of the true Marmousi model and the second 2004 BP model, the gradients (Figures 15b and 26b) do not have strong artifacts or nonphysical variation in the first iteration even if the corresponding adjoint sources are irregular with strong horizontal variations. It will

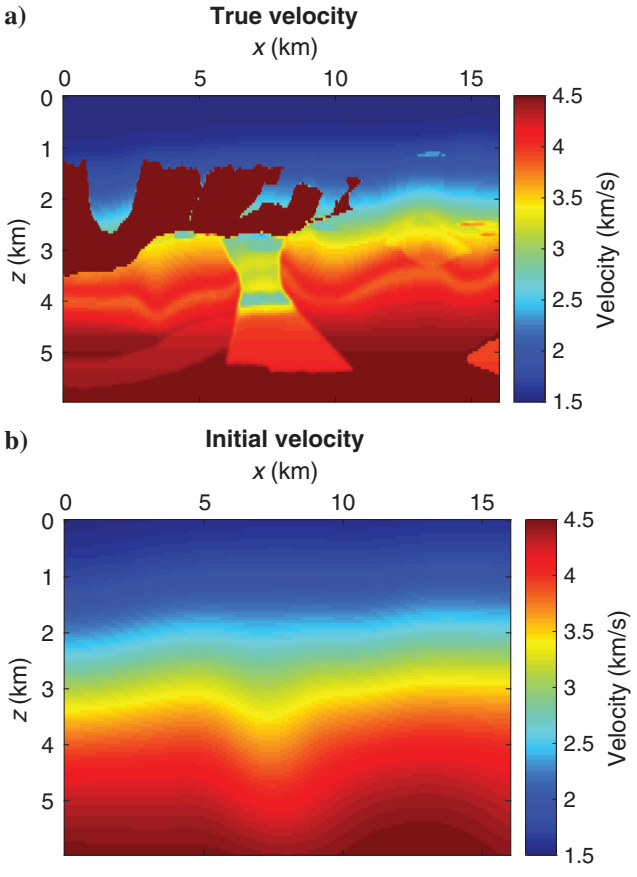

Figure 25. (a) True velocity and (b) initial velocity for the second BP model.

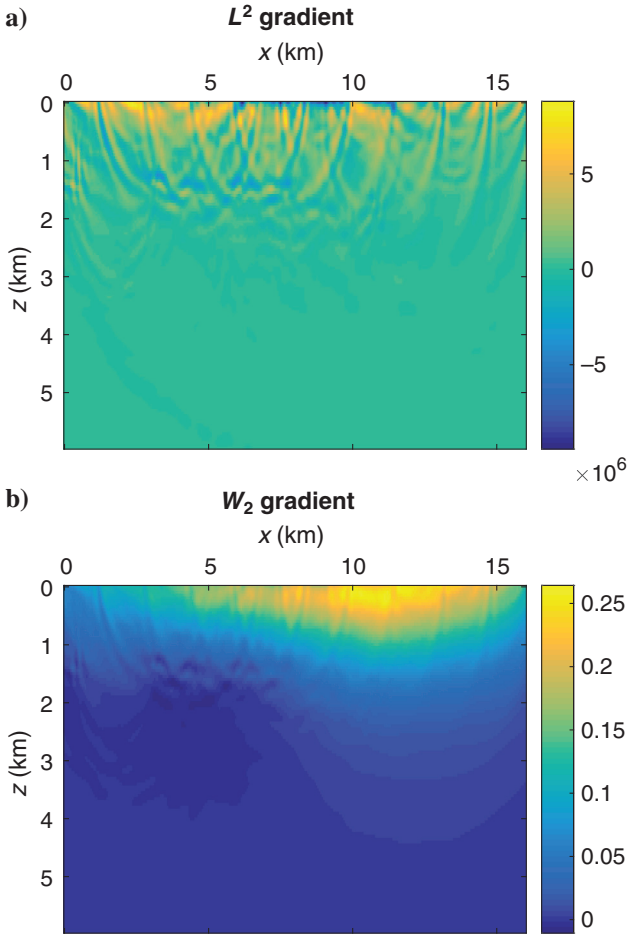

Figure 26. The gradient in the first iteration of (a) L^2 and (b) trace-by-trace W_2 inversion for the second BP model.

be interesting to study about the structure of the adjoint source in the success of FWI.

To compute the W_2 misfit globally, we solve a 2D optimal transport problem based on the Monge-Ampère equation formulation. The numerical method for the Monge-Ampère equation (Froese and Oberman, 2013) is proved to be convergent but requires the target profile g to be Lipschitz continuous, and the discretization error is proportional to the Lipschitz constant of g as theorem 3 states. A coarse discretization of the wavefield g will have a large value of Lh^2, if the Lipschitz constant is large. Because the error estimate is of the order $\mathcal{O}(Lh^2)$, this indicates a low accuracy in the numerical solution. In practice, the grid will have to be very well-refined, which means small h before we are able to achieve meaningful results. This means that for accurate results, enough data points are needed to effectively resolve steep gradients in the data; otherwise, the solver effectively regularizes the data before solving the Monge-Ampère equation. This was evident in the example of the Marmousi model: The solver was much more robust to the scaled velocity benchmarks that provided better resolution of the fronts in the data. The trace-by-trace approach, on the other hand, can make use of exact formulas for 1D optimal transportation, which allows for accurate computations even when the data are highly nonsmooth.

The oscillatory artifacts in Figures 10b, 13b, and 23b likely originate from a combination of the numerical PDE solution discussed above and the insensitivity to noise of the W_2 measure. The fact that W_2 is insensitive to noise is good for noisy measurements, but not for artifacts in the velocity model. This could, for example, be handled by total variation regularization (Rudin et al., 1992). We have chosen to present the raw results without pre or postprocessing. The trace-by-trace technique is also insensitive to noise, though to a lesser extent because there is cancellation only along one dimension. The 2D approach seems to have a slight edge for the Camembert model. The L^2 error between the converged velocity and the true model velocity with global W_2 is 25% smaller than in the trace-by-trace case.

In the sixth experiment of the computation result, we start with a rough model built by 100 iterations of trace-by-trace W_2. After the same number of iterations, the W_2 inversion result attained a lower model error than L^2 inversion, but in the vertical velocity profile L^2 has better accuracy than the trace-by-trace W_2. The W_2 norm can build a good initial model to help overcome the cycle-skipping problem. One can start with the trace-by-trace W_2 results in L^2-based FWI with higher frequency sources for better resolution and inversion of the high-wavenumber component.

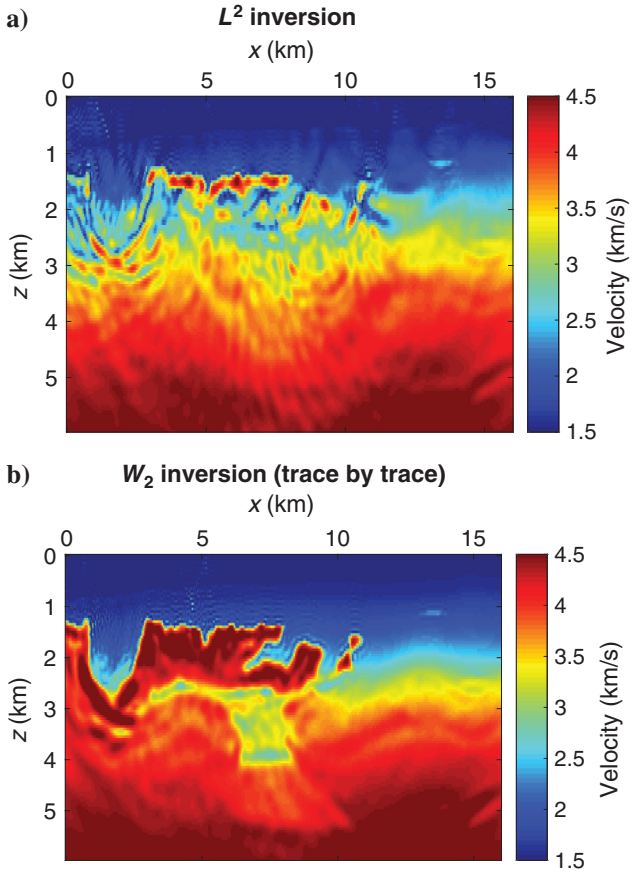

Figure 27. Inversion results of (a) L^2 and (b) trace-by-trace W_2 for the second BP model.

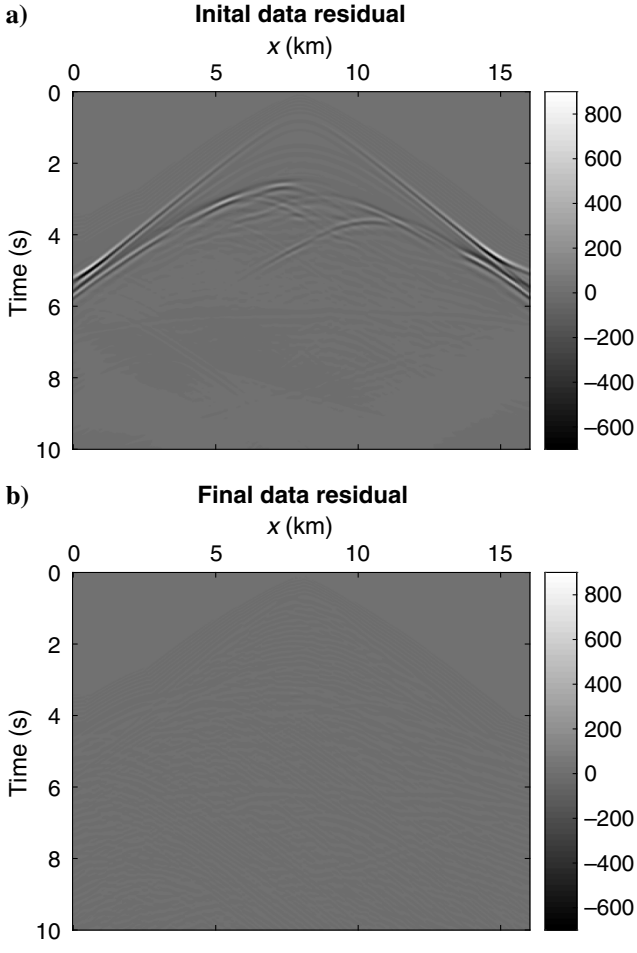

Figure 28. (a) The difference of data to be fit and the prediction with initial model and (b) the final data residual of trace-by-trace W_2 for the second BP model.

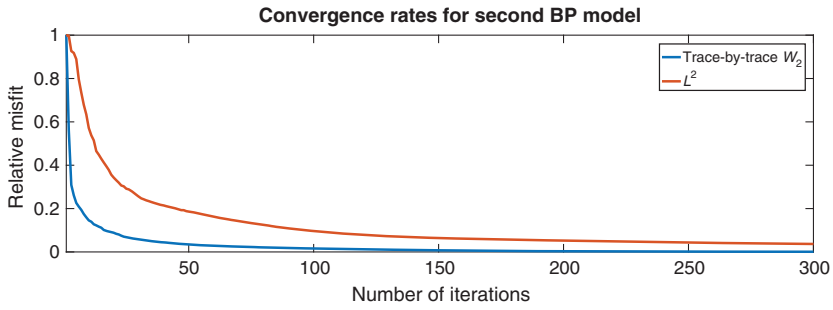

Figure 29. The convergence curves for trace-by-trace W_2- and L^2-based inversion of the second BP model.

CONCLUSION

We have developed a high-resolution FWI technique based on optimal transport and the quadratic Wasserstein metric W_2. Here, the W_2 misfit is coupled to efficient adjoint source computation for the optimization. Our earlier work with W_2 was limited to a few degrees of freedom, but here we have presented successful inversion of the Marmousi, the 2004 BP, and the so-called Camembert models. This novel technique avoids cycle skipping as is demonstrated by numerical examples. The 2D W_2 misfit is calculated by solving a relevant Monge-Ampère equation and the latest version of the solver is outlined. We also show comparable results from a trace-by-trace comparison with a W_2 misfit. This is as fast as the standard L^2-based FWI in terms of computation time, but it is more accurate and converges faster for cycle-skipping cases.

Our results clearly point to the quadratic Wasserstein metric as a potentially excellent choice for a misfit function in FWI. There are many possible directions for future improvements. In the 1D and 2D studies, the scaling or normalization of the signals play an important role. The linear normalization was by far the best for the large-scale inversion but does not satisfy the requirements of the theoretical result of convexity from shifts. This should be further investigated, and even better normalizations would be ideal. Extending the 1D trace-by-trace misfit to 1D comparisons along additional directions is also possible. This has been successfully tried in other applications under the name of the sliced Wasserstein distance.

ACKNOWLEDGMENTS

We thank S. Fomel, Z. Xue, and L. Qiu for very helpful discussions, and we thank the sponsors of the Texas Consortium for Computational Seismology for financial support. J. Sun was additionally supported by the Statoil Fellows Program at the University of Texas at Austin. B. Engquist was partially supported by NSF DMS-1620396. B. Froese was partially supported by NSF DMS-1619807. Y. Yang was partially supported by a grant from the Simons Foundation (#419126, B. Froese).

APPENDIX A

DERIVATION OF EQUATION 33

We assume that $f(t)$ and $g(t)$ are continuous density functions in $[0, T_0]$, and $F(t) = \int_0^t f(\tau)d\tau$ and $G(t) = \int_0^t g(\tau)d\tau$. Now, we perturb f by an amount δf and investigate the resulting change in equation 29 as a functional of f:

$$W_2^2(f,g) + \delta W = \int_0^{T_0} |t - G^{-1}(F(t) + \delta F(t))|^2 (f(t) + \delta f(t))dt, \quad \text{(A-1)}$$

$$= \int_0^{T_0} |t - G^{-1}(F(t) + \delta F(t))|^2 f(t)dt, \quad \text{(A-2)}$$

$$+ \int_0^{T_0} |t - G^{-1}(F(t))|^2 \delta f(t)dt + O((\delta f)^2). \quad \text{(A-3)}$$

Because G is monotone increasing, so is G^{-1}. We have the following Taylor expansion of G^{-1}:

$$G^{-1}(F(t) + \delta F(t)) = G^{-1}(F(t)) + \left.\frac{dG^{-1}(y)}{dy}\right|_{y=F(t)} \delta F(t) + O((\delta f)^2). \quad \text{(A-4)}$$

Substituting equation A-4 back into equation A-3, we obtain the first variation:

$$\delta W = \int_0^{T_0} \left(\int_t^{T_0} -2(s - G^{-1}(F(s)) \left.\frac{dG^{-1}(y)}{dy}\right|_{y=F(s)} f(s)ds \right) \delta f(t)dt, \quad \text{(A-5)}$$

$$+ \int_0^{T_0} |t - G^{-1}(F(t))|^2 \delta f(t)dt. \quad \text{(A-6)}$$

Thus, the Fréchet derivative of equation 29 with respect to f is

$$\frac{dW_2^2(f,g)}{df} = \left(\int_t^{T_0} -2(s - G^{-1}(F(s)) \left.\frac{dG^{-1}(y)}{dy}\right|_{y=F(s)} f(s)ds + |t - G^{-1}(F(t))|^2 \right) dt. \quad \text{(A-7)}$$

In our numerical scheme, we discretize equation A-7 and derive equations 31 and 33.

REFERENCES

Benamou, J.-D., and Y. Brenier, 2000, A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem: Numerische Mathematik, **84**, 375–393, doi: 10.1007/s002110050002.

Benamou, J.-D., G. Carlier, M. Cuturi, L. Nenna, and G. Peyré, 2015, Iterative Bregman projections for regularized transportation problems: SIAM Journal on Scientific Computing, **37**, A1111–A1138, doi: 10.1137/141000439.

Benamou, J.-D., B. D. Froese, and A. M. Oberman, 2014, Numerical solution of the optimal transportation problem using the Monge-Ampere equation: Journal of Computational Physics, **260**, 107–126, doi: 10.1016/j.jcp.2013.12.015.

Billette, F., and S. Brandsberg-Dahl, 2005, The 2004 BP velocity benchmark: 67th Annual International Conference and Exhibition, EAGE, Extended Abstracts, B035.
Brenier, Y., 1991, Polar factorization and monotone rearrangement of vector-valued functions: Communications on Pure and Applied Mathematics, **44**, 375–417, doi: 10.1002/(ISSN)1097-0312.
Brossier, R., S. Operto, and J. Virieux, 2010, Which data residual norm for robust elastic frequency-domain full waveform inversion?: Geophysics, **75**, no. 3, R37–R46, doi: 10.1190/1.3379323.
Bunks, C., F. M. Saleck, S. Zaleski, and G. Chavent, 1995, Multiscale seismic waveform inversion: Geophysics, **60**, 1457–1473, doi: 10.1190/1.1443880.
Engquist, B., and B. D. Froese, 2014, Application of the Wasserstein metric to seismic signals: Communications in Mathematical Sciences, **12**, 979–988, doi: 10.4310/CMS.2014.v12.n5.a7.
Engquist, B., B. D. Froese, and Y. Yang, 2016, Optimal transport for seismic full waveform inversion: Communications in Mathematical Sciences, **14**, 2309–2330, doi: 10.4310/CMS.2016.v14.n8.a9.
Esser, E., L. Guasch, T. van Leeuwen, A. Y. Aravkin, and F. J. Herrmann, 2015, Total variation regularization strategies in full waveform inversion for improving robustness to noise, limited data and poor initializations: Technical Report TR-EOAS-2015–5.
Froese, B. D., 2012, A numerical method for the elliptic Monge-Ampère equation with transport boundary conditions: SIAM Journal on Scientific Computing, **34**, A1432–A1459, doi: 10.1137/110822372.
Froese, B. D., and A. M. Oberman, 2013, Convergent filtered schemes for the Monge-Ampère partial differential equation: SIAM Journal on Numerical Analysis, **51**, 423–444, doi: 10.1137/120875065.
Gauthier, O., J. Virieux, and A. Tarantola, 1986, Two-dimensional nonlinear inversion of seismic waveforms: Numerical results: Geophysics, **51**, 1387–1403, doi: 10.1190/1.1442188.
Gholami, A., and H. Siahkoohi, 2010, Regularization of linear and nonlinear geophysical ill-posed problems with joint sparsity constraints: Geophysical Journal International, **180**, 871–882, doi: 10.1111/j.1365-246X.2009.04453.x.
Ha, T., W. Chung, and C. Shin, 2009, Waveform inversion using a back-propagation algorithm and a Huber function norm: Geophysics, **74**, no. 3, R15–R24, doi: 10.1190/1.3112572.
Knott, M., and C. S. Smith, 1984, On the optimal mapping of distributions: Journal of Optimization Theory and Applications, **43**, 39–49, doi: 10.1007/BF00934745.
Kolouri, S., S. Park, M. Thorpe, D. Slepčev, and G. K. Rohde, 2016, Transport-based analysis, modeling, and learning from signal and data distributions: arXiv preprint arXiv:1609.04767.
Lailly, P., 1983, The seismic inverse problem as a sequence of before stack migrations: Proceedings of the Conference on Inverse Scattering: Theory and Application, SIAM, 206–220.
Ma, Y., and D. Hale, 2013, Wave-equation reflection traveltime inversion with dynamic warping and full-waveform inversion: Geophysics, **78**, no. 6, R223–R233, doi: 10.1190/geo2013-0004.1.
Métivier, L., R. Brossier, Q. Mérigot, E. Oudet, and J. Virieux, 2016a, Increasing the robustness and applicability of full-waveform inversion: An optimal transport distance strategy: The Leading Edge, **35**, 1060–1067, doi: 10.1190/tle35121060.1.
Métivier, L., R. Brossier, Q. Mérigot, E. Oudet, and J. Virieux, 2016b, Measuring the misfit between seismograms using an optimal transport distance: Application to full waveform inversion: Geophysical Journal International, **205**, 345–377, doi: 10.1093/gji/ggw014.
Métivier, L., R. Brossier, Q. Mrigot, E. Oudet, and J. Virieux, 2016c, An optimal transport approach for seismic tomography: Application to 3D full waveform inversion: Inverse Problems, **32**, 115008, doi: 10.1088/0266-5611/32/11/115008.
Monge, G., 1781, Mémoire sur la théorie des déblais et de remblais. Histoire de l'académie royale des sciences de paris: Avec les Mémoires de Mathématique et de Physique pour la mme année, 666–704.
Oberman, A. M., and Y. Ruan, 2015, An efficient linear programming method for optimal transportation: arXiv preprint arXiv:1509.03668.
Plessix, R.-E., 2006, A review of the adjoint-state method for computing the gradient of a functional with geophysical applications: Geophysical Journal International, **167**, 495–503, doi: 10.1111/j.1365-246X.2006.02978.x.
Qiu, L., N. Chemingui, Z. Zou, and A. Valenciano, 2016, Full-waveform inversion with steerable variation regularization: 86th Annual International Meeting, SEG, Expanded Abstracts, 1174–1178.
Rudin, L. I., S. Osher, and E. Fatemi, 1992, Nonlinear total variation based noise removal algorithms: Physica D: Nonlinear Phenomena, **60**, 259–268, doi: 10.1016/0167-2789(92)90242-F.
Schmitzer, B., 2016, A sparse multiscale algorithm for dense optimal transport: Journal of Mathematical Imaging and Vision, **56**, 238–259, doi: 10.1007/s10851-016-0653-9.
Sirgue, L., O. Barkved, J. Dellinger, J. Etgen, U. Albertin, and J. Kommedal, 2010, Thematic set: Full waveform inversion: The next leap forward in imaging at valhall: First Break, **28**, 65–70, doi: 10.3997/1365-2397.2010012.
Tarantola, A., 1984, Inversion of seismic reflection data in the acoustic approximation: Geophysics, **49**, 1259–1266, doi: 10.1190/1.1441754.
Tarantola, A., and B. Valette, 1982, Generalized nonlinear inverse problems solved using the least squares criterion: Reviews of Geophysics, **20**, 219–232, doi: 10.1029/RG020i002p00219.
Villani, C., 2003, Topics in optimal transportation, Graduate studies in mathematics: American Mathematical Society 58, 1–145.
Virieux, J., and S. Operto, 2009, An overview of full-waveform inversion in exploration geophysics: Geophysics, **74**, no. 6, WCC1–WCC26, doi: 10.1190/1.3238367.
Warner, M., and L. Guasch, 2014, Adaptive waveform inversion: Theory: 84th Annual International Meeting, SEG, Expanded Abstracts, 1089–1093.
Zhu, H., and S. Fomel, 2016, Building good starting models for full-waveform inversion using adaptive matching filtering misfit: Geophysics, **81**, no. 5, U61–U72, doi: 10.1190/geo2015-0596.1.