Methods and Applications of Analysis.
Vol. 26, No. 2, pp. 133–148, June 2019

003

SEISMIC INVERSION AND THE DATA NORMALIZATION FOR OPTIMAL TRANSPORT*

BJÖRN ENGQUIST† AND YUNAN YANG‡

Abstract. Full waveform inversion (FWI) has recently become a favorite technique for the inverse problem of finding properties in the earth from measurements of vibrations of seismic waves on the surface. Mathematically, FWI is PDE constrained optimization where model parameters in a wave equation are adjusted such that the misfit between the computed and the measured dataset is minimized. In a sequence of papers, we have shown that the quadratic Wasserstein distance from optimal transport is to prefer as misfit functional over the standard L^2 norm. Datasets need however first to be normalized since seismic signals do not satisfy the requirements of optimal transport. There has been a puzzling contradiction in the results. Normalization methods that satisfy theorems pointing to ideal properties for FWI have not performed well in practical computations, and other scaling methods that do not satisfy these theorems have performed much better in practice. In this paper, we will shed light on this issue and resolve this contradiction.

Key words. Optimal transport, seismic inversion, Monge-Ampére Equation, optimization, data normalization.

Mathematics Subject Classification. 65K10, 86A22, 49N45.

1. Introduction.

There are two major processes in exploration seismology. One is migration or reverse time migration (RTM), which determines details of the reflecting surfaces assuming an approximate model of wave velocity [1]. Seismic inversion or full waveform inversion (FWI) is a process of recovering the quantitative features of the geophysical structure. The focus is currently on the nonlinear inverse problem of building an accurate model of the wave velocity in the earth. This is done in an iterative process where a forward seismic simulation based on the unknown velocity is matched to the actual recordings [25]. There are many related techniques in seismic exploration. Wave equation travel time tomography [21] and the ray-based tomography are phase-like inversion methods [28]. Least-squares inversion is known as linearized waveform inversion [20, 30] and the least-square reverse time migration (LSRTM) [9] based on the Born approximation [17, 33] is one example, where the background model is not updated after each iteration.

FWI is a high-resolution seismic imaging technique, which recently has been getting great attention from both academia and industry [35]. The goal of FWI is to find both the small-scale and the large-scale components, which describe the geophysical properties using the entire content of seismic traces. A trace is the time history of seismic vibrations measured at a receiver. In this paper, we will consider the inverse problem of finding the wave velocity of an acoustic wave equation in the interior of a domain from knowing the Cauchy boundary data together with natural boundary conditions [8]. This is implemented by minimizing the difference or mismatch between computed and measured data on the boundary. It is thus a partial differential

*Received January 7, 2018; accepted for publication July 30, 2019. It is our honor to dedicate this paper to Professor Roland Glowinski on the occasion of his eighties birthday. We thank Junzhe Sun and Lingyun Qiu for constructive discussions and thank the sponsors of the Texas Consortium for Computational Seismology (TCCS) for financial support. The first author was partially supported by NSF DMS-1620396.

†Department of Mathematics and ICES, The University of Texas at Austin, 1 University Station C1200, Austin, TX 78712 USA (engquist@math.utexas.edu).

‡Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, NY 10012 USA (yunan.yang@nyu.edu).

equation (PDE) constrained optimization.

FWI is increasing in popularity even if it is still facing major computational challenges. Depending on the parameterization of the velocity model this inverse PDE-constrained optimization problem is often highly non-unique and non-convex in nature. The least-squares norm (L^2), which is classically used in FWI to measure the data mismatch, suffers from local minima trapping, the so-called cycle skipping issues, and sensitivity to noise [29]. We will see that optimal transport based quadratic Wasserstein metric (W_2) is capable of dealing with some of these limitations by including both amplitudes mismatches and travel time differences.

The idea of using Wasserstein metric for seismic inversion was first proposed in [11]. This metric is based on optimal transport [34]. We first transform our datasets of seismic signals into density functions of two probability distributions. Next, we find the optimal map between these two datasets and compute the corresponding transport cost as the misfit function in FWI, either by solving a Monge-Ampère equation for the entire dataset or by using the explicit 1D formula [34] measuring the misfit trace by trace [39]. Following the idea that changes in velocity cause a shift or "transport" in the arrival time of a seismic signal, we demonstrated in [12] the advantageous mathematical properties of the quadratic Wasserstein metric (W_2) and provided rigorous proofs that laid a solid theoretical foundation for this new misfit function.

There are two main requirements for signals f and g in optimal transport theory:

$$f(t) \geq 0,\ g(t) \geq 0,\ < f >= \int f(t)dt = \int g(t)dt =< g > . \tag{1}$$

Since these constraints are not expected for seismic signals, some data pre-processing is needed before we can implement the Wasserstein-based FWI. In [38, 39] we normalized the signals by adding a constant,

$$\tilde{f}(t) = \frac{f(t)+c}{< f+c >},\ \tilde{g}(t) = \frac{g(t)+c}{< g+c >},\ c = \min_t(f(t), g(t)). \tag{2}$$

This worked remarkably well in realistic large scale examples [38, 39] together with the adjoint-state method for optimization in either the 1D or the Monge-Ampère based techniques. This linear normalization does, however, not give a convex misfit functional with respect to simple shifts. Other normalizations that generate convex misfits were also tried as, for example, only using the positive part of the signals, squaring or taking the envelope or the absolute values [11, 12]. It was puzzling that these misfit functionals performed poorly with realistic datasets.

FWI will be introduced in section two, and we will present relevant parts of optimal transport theory as background in section three. The new material is in section four where data normalizations are discussed. We will see that it is desirable to require the scaling function to be differentiable so that it is easy to apply chain rule when calculating the Fréchet derivative for FWI backpropagation and also better suited for the Monge-Ampère solver. Other aspects of normalization are also discussed that explain the contradictions mentioned above and finally ending up with a new normalization that satisfies most of the essential properties:

$$\tilde{f}(t) = \begin{cases} (f(t) + \frac{1}{c})/b,\ f(t) \geq 0,\ c > 0 \\ \frac{1}{c}\exp(cf(t))/b,\ f(t) < 0 \end{cases} \tag{3}$$

where $b =< (f + \frac{1}{c})\mathbb{1}_{f\geq 0} + \frac{1}{c}\exp(cf)\mathbb{1}_{f<0} >$.

2. Full Waveform Inversion. Full Waveform Inversion (FWI) is a nonlinear inverse technique that utilizes the entire wavefield information to estimate the earth properties. The notion of FWI was first brought up three decades ago [19, 32] and has been actively studied and applied with the increase in computing power. It is now a common technique in practice.

Wave-propagation modeling is the most basic step in seismic imaging. Without loss of generality, we will explain everything in a simple acoustic setting in this paper:

$$\begin{cases} m(\mathbf{x})\frac{\partial^2 u(\mathbf{x},t)}{\partial t^2} - \triangle u(\mathbf{x},t) = s(\mathbf{x},t) \\ u(\mathbf{x},0) = 0 \\ \frac{\partial u}{\partial t}(\mathbf{x},0) = 0 \end{cases} \tag{4}$$

We assume the model $m(\mathbf{x}) = \frac{1}{c(\mathbf{x})^2}$ where $c(\mathbf{x})$ is the velocity, $u(\mathbf{x},t)$ is the wavefield, $s(\mathbf{x},t)$ is the source. It is a linear PDE but a nonlinear operator from model domain $m(\mathbf{x})$ to data domain $u(\mathbf{x},t)$. We note that there are numerous techniques for approximating (4), for example, discontinuous and continuous finite elements, spectral elements and finite difference methods. As our focus is on the inverse problem rather than on the solution of the forward problem, we will restrict our discretization to standard finite difference methods [23]. We use the absorbing boundary condition [13] in the numerical scheme to approximate the effect of an unbounded domain.

As we will see, the mathematical formulation of FWI is PDE constrained optimization. The objective function is the misfit between the synthetic data which is generated by solving certain wave equation numerically with predicted model parameters and the observed data measured from the field which is a result of natural propagation with the real physics. For example, in time domain conventional FWI defines a least-squares waveform misfit as

$$d(f,g) = J_1(m) = \frac{1}{2}\sum_r \int |f(\mathbf{x_r},t;m) - g(\mathbf{x_r},t)|^2\, dt, \tag{5}$$

where $\mathbf{x_r}$ are receiver locations, g is observed data, and f is simulated data which solves (4) with model parameter m. This formulation can also be extended to the case with multiple sources.

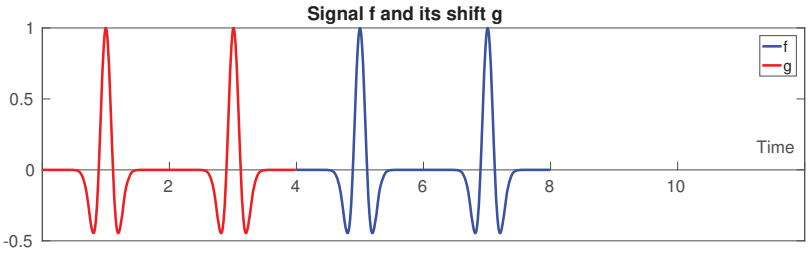
In large-scale realistic 3D FWI, there are typically millions of variables describing $m(\mathbf{x})$. It is not practical to compute the derivative of the misfit function with respect to each model variable directly. With the adjoint-state method, one only needs to solve two wave equations numerically to compute the Fréchet derivative, the forward propagation and the adjoint wavefield propagation. Different misfit functions $J(m)$ typically only affect the source term in the adjoint wave equation [26, 31]. The gradient is similar to the usual imaging condition [7]:

$$\frac{\partial J}{\partial m} = -\int_0^T \frac{\partial^2 u(\mathbf{x},t)}{\partial t^2} v(\mathbf{x},t) dt, \tag{6}$$

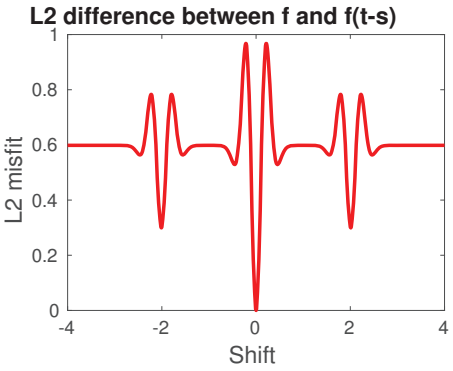
where v is the solution to the adjoint wave equation:

$$\begin{cases} m\frac{\partial^2 v(\mathbf{x},t)}{\partial t^2} - \triangle v(\mathbf{x},t) = R^T \frac{\partial J}{\partial f} \\ v(\mathbf{x},T) = 0 \\ v_t(\mathbf{x},T) = 0 \end{cases} \tag{7}$$

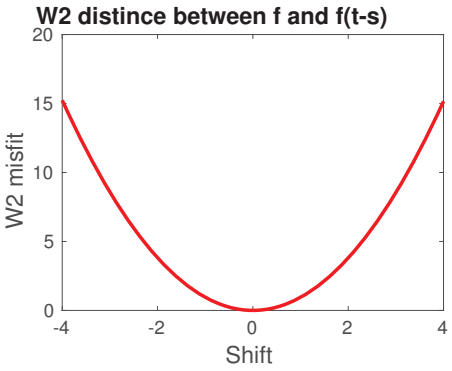
Here R is a restriction operator only at the receiver locations.



(a) Two signals



(b) L^2 sensitivity curve



(c) W_2 sensitivity curve

Fig. 1: (A) A signal consisting two Ricker wavelets (blue) and its shift (red) (B) L^2 norm between f and g which is a shift of f. (C) W_2 norm between f and g in terms of different shift s

It is well known that the accuracy of FWI with L^2 norm as misfit functional deteriorates from the lack of low frequencies, data noise, and poor starting model, which may result in local minima trapping. These limitations are on top of the potential ill-posedness of the inverse problem which we here treat as a PDE-constrained optimization. Figure 1a displays two signals, each of which contains two Ricker wavelets and f is simply a shift of g. The L^2 norm between f and g is plotted in Figure 1b as a function of the shift s. We observe many local minima and maxima in this simple two-event setting which again demonstrated the difficulty of the so-called cycle-skipping issues [37].

A recently introduced class of misfit functions to tackle the cycle-skipping issue is the quadratic Wasserstein metric [5, 11, 12, 37, 38, 39]. The L^2 misfit function measures the difference in amplitude locally. The optimal transport based methods compare the observed and simulated data globally and thus more effectively include phase information.

As a useful tool from the theory of optimal transport, the quadratic Wasserstein metric (W_2) computes the minimal cost of rearranging one distribution into another with a quadratic cost function. The squared Wasserstein metric has several properties that make it attractive as a choice for misfit function [12]. One highly desirable feature is its convexity with respect to several parameterizations that occur naturally

in seismic waveform inversion. As seen in Figure 1c, W_2 norm significantly improves the convexity of the misfit sensitivity curve. Another important property of optimal transport is the insensitivity to noise. One can find the more theoretical results in [12] and the numerical examples in [39].

3. Optimal Transport and the Wasserstein metric. The topic of optimal transport starts with the problem brought up by Gaspard Monge in 1781 [24]. Let X and Y be two metric spaces with probability measures μ and ν respectively. Assume X and Y have equal total measure:

$$\int_X d\mu = \int_Y d\nu \tag{8}$$

Without loss of generality, we will hereafter assume the total measure to be one, i.e., μ and ν are probability measures.

DEFINITION 1 (Mass-preserving map). *A map $T : X \to Y$ is mass-preserving if for any measurable set $B \in Y$,*

$$\mu(T^{-1}(B)) = \nu(B) \tag{9}$$

If this condition is satisfied, ν is said to be the push-forward of μ by T, and we write $\nu = T_{\#}\mu$

Given two nonnegative densities $f = d\mu$ and $g = d\nu$, we are interested in the mass-preserving map T such that $f = g \circ T$. The transport 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 $|x - y|^2$. We are interested in finding the optimal map that minimizes the total cost which formally defines a class of metrics: the Wasserstein distance:

DEFINITION 2 (The Wasserstein distance). *We denote by $\mathscr{P}_p(X)$ the set of probability measures with finite moments of order p. For all $p \in [1, \infty)$,*

$$W_p(\mu, \nu) = \left(\inf_{T_{\mu,\nu} \in \mathcal{M}} \int_{\mathbb{R}^n} |x - T_{\mu,\nu}(x)|^p \, d\mu(x) \right)^{\frac{1}{p}}, \quad \mu, \nu \in \mathscr{P}_p(X). \tag{10}$$

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

The optimal transport in higher dimension has no explicit solutions. It is an infinite dimensional optimization problem if we search directly in the function space for T. An alternative is to solve the relaxed dual problem by outstanding techniques in linear programming, for example, the alternating direction method of multipliers (ADMM), see the survey [16] by Glowinski. However, the optimal map takes on additional structure in the special case of a quadratic cost function (i.e. $c(x, y) = |x - y|^2$). The following Brenier's theorem [3, 10] gives an elegant result about the uniqueness of optimal transport map for the quadratic cost as well as its intrinsic connection with the Monge-Ampère equation:

THEOREM 1 (Brenier's theorem [34]). *Let μ and ν be two compactly supported probability measures on $\mathbb{R}^n$. If μ is absolutely continuous with respect to the Lebesgue measure, then*

1. *There is a unique optimal map T for the cost function $c(x, y) = |x - y|^2$.*

2. *There is a convex function* $u : \mathbb{R}^n \to \mathbb{R}$ *such that the optimal map* T *is given by* $T(x) = \nabla u(x)$ *for* μ*-a.e.* x.

Furthermore, if $\mu(dx) = f(x)dx$, $\nu(dy) = g(y)dy$, *then* T *is differential* μ*-a.e. and*

$$\det(\nabla T(x)) = \frac{f(x)}{g(T(x))}. \tag{11}$$

According to Brenier's theorem, in order to compute the misfit between distributions f and g, one can first get the optimal map $T(x) = \nabla u(x)$ via the solution of the following Monge-Ampère equation:

$$\det(D^2 u(x)) = \frac{f(x)}{g(\nabla u(x))}, \quad u \text{ is convex.} \tag{12}$$

Typically it is coupled to the non-homogeneous Neumann boundary condition

$$\nabla u(x) \cdot \nu = x \cdot \nu, \ x \in \partial X. \tag{13}$$

The squared Wasserstein metric is then given by

$$W_2^2(f, g) = \int_X f(x) \left| x - \nabla u(x) \right|^2 dx. \tag{14}$$

We have followed [2] for the numerical solution to the Monge-Ampère equation when computing the quadratic Wasserstein distance for the global comparison in FWI [39]. For a survey of recent numerical methods for nonlinear second order PDEs, see [14].

4. Data Normalization. The primary constraints for applying optimal transport to general signals are that the functions should be restricted to nonnegative measures sharing equal total mass (e.g., probability distributions). This is a crucial limitation for many applications that need to compare general signals or allow for only partial displacement of the mass.

4.1. Background. There are many proposals in the literature for dealing with the mass balance constraint. Two notions particularly stand out, which are derived rigorously as an extension based on the original optimal transport problem. One is the unbalanced optimal transport, which is formulated as another well-defined metric named the Wasserstein-Fisher-Rao distance [6, 18]. The other approach is the optimal partial transport whose mathematical properties are discussed in detail by [4, 15]. As a comparison, there are very few papers discussing the positivity constraint. In [22], a proposal is made to recombine the data using the decomposition in positive and negative part to compare positive measures with mass conservation. It is based on the following special dual form of the W_1 metric, i.e., $p = 1$ in (10), between density functions $f = d\mu$ and $g = d\nu$:

$$W_1(f, g) = \max_{\varphi \in \mathrm{Lip}_1} \int_X \varphi(x)(f(x) - g(x))dx, \tag{15}$$

where Lip_1 is the space of all 1-Lipschitz functions.

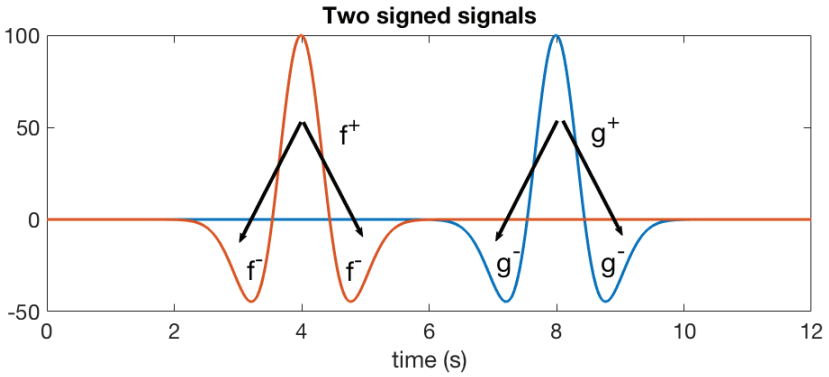


Fig. 2: The optimal transport may map f^+ to f^- and g^+ to g^- if formulated as (20) (arrows indicate transport)

Based on the dual formulation above, one can easily extend it to signed measures f and g by defining

$$\widetilde{W}_1(f,g) = \widetilde{W}_1(f^+ - f^-,\ g^+ - g^-) \tag{16}$$
$$= \max_{\varphi\in \mathrm{Lip}_1} \int_X \varphi(x)(f^+ - f^- - g^+ + g^-)dx \tag{17}$$
$$= \max_{\varphi\in \mathrm{Lip}_1} \int_X \varphi(x)(f^+ + g^- - (f^- + g^+))dx \tag{18}$$
$$= \widetilde{W}_1(f^+ + g^-,\ f^- + g^+) \tag{19}$$
$$= W_1(\rho_1, \rho_2), \tag{20}$$

where $\rho_1 = f^+ + g^-$, the sum of the positive part of f and the negative part of g, and $\rho_2 = f^- + g^+$, the sum of the negative part of f and the positive part of g. The W_1 in (20) is same as the standard 1-Wasserstein distance in (15).

The formulation above defines a cost for transporting signed measures. However, it is not a canonical optimal transport distance. There is a risk that the true optimal transport represented in (20) matches f^+ to f^- and g^+ to g^- under certain circumstances (Figure 2). Especially in FWI, we want to map one signal to the other instead of compensating within one signal itself.

4.2. Early ideas. In this section, we will introduce some normalization ideas which we proposed in the past to transform seismic data into probability signals such that the standard optimal transport theory will apply. We will analyze their properties and in particular why they often have problems with realistic large-scale FWI.

In [11, 12], the signals were separated into positive and negative parts $f^+ = \max\{f, 0\}$, $f^- = \max\{-f, 0\}$ and scaled by the total mass $\langle f\rangle = \int_X f(x)\,dx$ (Figure 3). Inversion was accomplished using the modified misfit function

$$J_2(m) = W_2^2\left(\frac{f^+}{\langle f^+\rangle}, \frac{g^+}{\langle g^+\rangle}\right) + W_2^2\left(\frac{f^-}{\langle f^-\rangle}, \frac{g^-}{\langle g^-\rangle}\right). \tag{21}$$

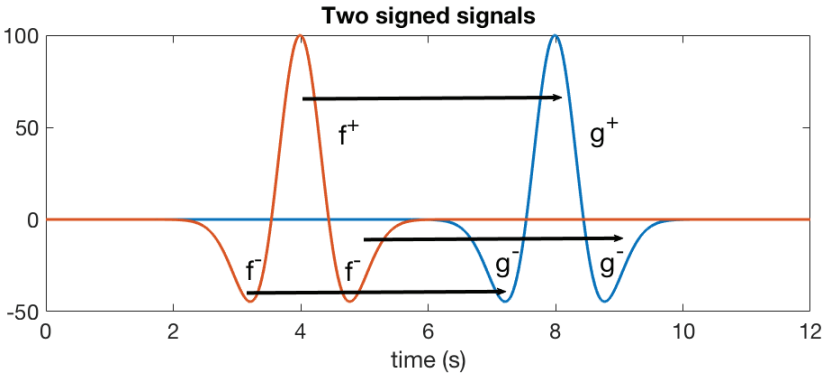


Fig. 3: The optimal transport plan maps f^+ to g^+ and f^- to g^- if formulated as (21) (arrows indicate transport)

Recall the adjoint-state equation introduced earlier (7). In order to compute the gradient of the misfit function with respect to the model parameters for FWI, we simply need the Fréchet derivative of the misfit function with respect to the synthetic data f. Therefore, the critical element in the backpropagation is $\frac{\partial J}{\partial f}$. Once we separate the signals as in (21), discontinuities are introduced in derivatives of f which causes problems in the optimization process and for the wave equation solvers. The same principle applies to the absolute-value scaling $W_2^2(|f|, |g|)$ since absolute-value function is not differentiable at zero.

The linear scaling we used in our earlier papers, i.e., Equation (2), on the other hand, works very well even if the related misfit lacks strict convexity with respect to shifts (see Figure 4). Here are several beneficial properties about the linear scaling. First, it has a wider basin of attraction than L^2 norm when it comes to simple shifts [38]. The two-variable example described in [37] is based on the linear scaling. It gives the convexity with respect to a subset of model variables in velocity compared to the result of L^2. Second, it provides a smooth bijection between the original data and the normalized data, which is favorable when combining with the adjoint-state method. Third, realistic seismic data always has the mean-zero property after a standard data processing. This indicates that $< f + c >$ is equal to $< g + c >$. This means that if two short seismic signals or, so-called events, are well matched between f and g they will stay so even after the normalization process and not be influenced by other events further away. The property is essential in the early iteration steps when the simulated signals do not include all details that are in the measured signal. On the other hand, if the individual events are void of zero frequencies the transport defining W_2 may be local as is seen in Figure 4, which can cause trapping in local minima.

One scaling method which theoretically should work well with the adjoint-state method is to square the signals first and normalize it to be mass balanced:

$$J_3(m) = W_2^2\left(\frac{f^2}{< f^2 >}, \frac{g^2}{< g^2 >}\right). \tag{22}$$

As seen in Figure 5, the two curves are the squares of the two functions in Figure 3. This particular normalization keeps the convexity of the quadratic Wasserstein metric concerning simple shifts like the setting in Figure 1c. In [5], squaring the data was used

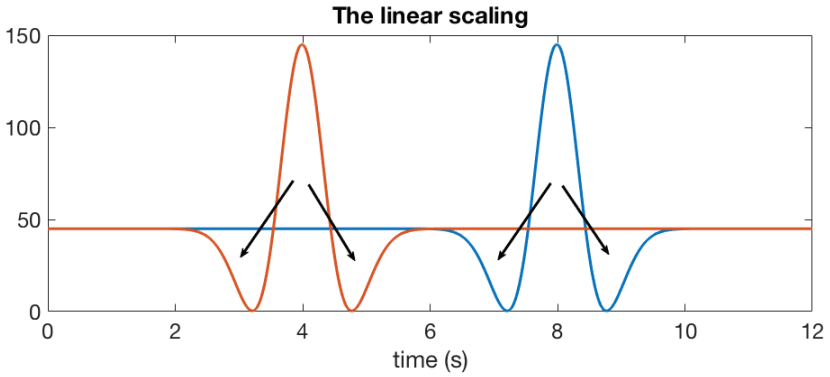


Fig. 4: The linear scaling: $f \to f + c$ and $g \to g + c$; there is chance of having local transport (arrows indicate transport)

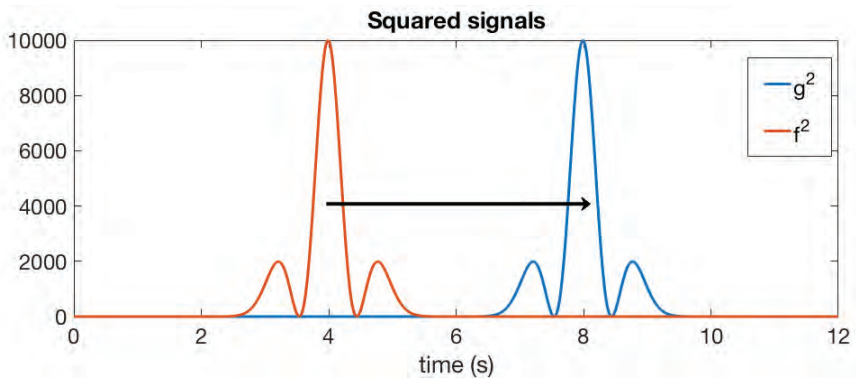
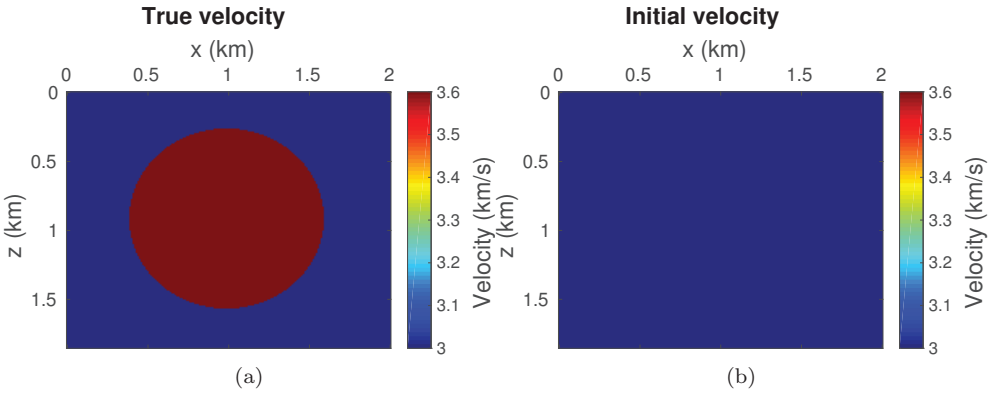


Fig. 5: Square of the data: $f \to f^2$ and $g \to g^2$ (arrows indicate transport)

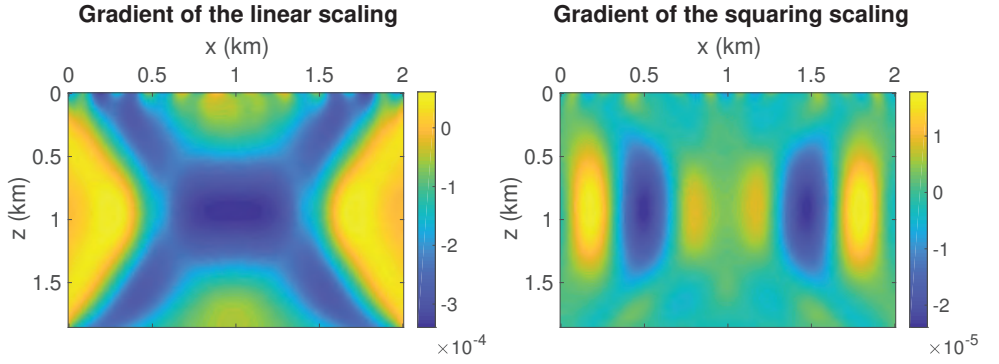
as the normalization to recover a four-variable linear source inversion. However, it has been puzzling since this normalization rarely works well in large-scale inversions with thousands of variables, such as the Camembert example and the standard Marmousi benchmark which we will show later.

It has been a dilemma until recently we can point to three potential factors that may lead to the difficulties. First of all, taking the squares boosts the higher frequency of the signal. It is well known that FWI becomes more difficult as the frequency increases. The robust convergence range is typically within half wavelength [36]. Just consider a simple oscillartory pulse $\sin(t)^2 = (1-\cos(2t))/2$. Second, the refracted, or so-called, diving wave and the reflection wave may reach the receiver at the same time with a similar amplitude but entirely different polarity. The positive and negative parts of the signal are interchanged. Squaring the signals may lose the important phase information here. Third, when we are dealing with one event in f and multiple events in g, $< g^2 >$ can be significantly larger than $< f^2 >$. This is often the



(a) (b)

Fig. 6: (a) True velocity and (b) inital velocity for the Camembert model

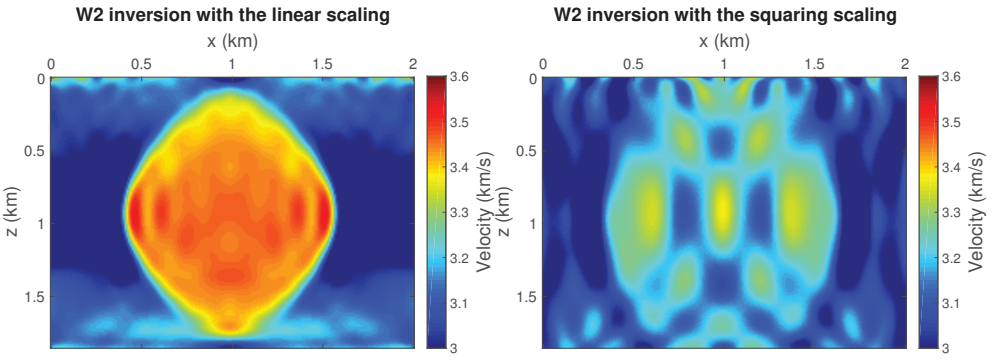


(a) Gradient by $\tilde{f} = af + b$ (b) Gradient by $\tilde{f} = f^2$

Fig. 7: The gradient in the first iteration of the inversion by using (a) the linear scaling as the data normalization, and (b) the squaring scaling as the data normalization.

case in the initial state of inversion when only one or a few reflections interfaces are known. The measured data g naturally contains the effect of all reflections. The mass normalization step can distort the correct parts of the signals that both f and g share, and consequently leads to a wrong update in the correct model variables.

We want to demonstrate the issues above with a Camembert model. The true velocity is shown in Figure 6a and the initial velocity we use in the inversion is Figure 6b. Figure 7 illustrates a comparison in gradients of the first iteration between the linear scaling and the squaring scaling as different normalizations in optimal transport FWI. There are wrong features in Figure 7b even in the first iteration. The final inversion results are shown in Figure 8. The linear scaling converges to a reasonably well model (Figure 8a) while squaring the data in normalization leads the inversion to a local minimum (Figure 8b). In both of these two experiments, we use the trace-by-trace techqniue (1D optimal transport) to compute the W_2 distance between the synthetic



(a) Inversion by $\tilde{f} = af + b$ (b) Inversion by $\tilde{f} = f^2$

Fig. 8: (a) The inversion result of using (2) as the data normalization (b) The inversion result of using (22) as the data normalization.

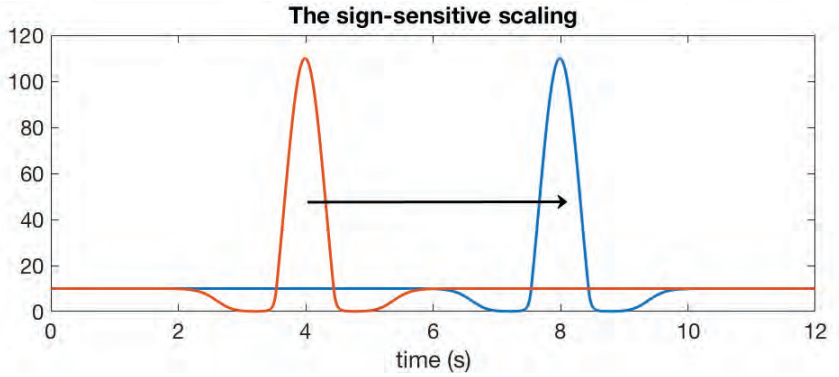


Fig. 9: One normalization combining both the linear and the exponential methods (Equation (3))

data and the observed data.

4.3. A sign-sensitive normalization. Based on the analysis of the squaring scaling in the previous section, a bijection between the original data and the normalized data is essential not to deteriorate the ill-posedness of the inverse problem. An exponential based normalization was proposed in [27] to transform seismic signals to probability distributions:

$$\tilde{f}(t) = \frac{\exp(cf(t))}{< \exp(cf) >}, \ \tilde{g}(t) = \frac{\exp(cg(t))}{< \exp(cg) >}, \ c > 0. \tag{23}$$

Here we propose a new normalization (Equation (3)) that satisfies most of the essential properties. This normalization (Figure 9) can be seen as a compromise between the linear scaling (2) and the positive part scaling (21). For the limit of small $c > 0$, it is a linear scaling, which directly follows from Taylor expansion of

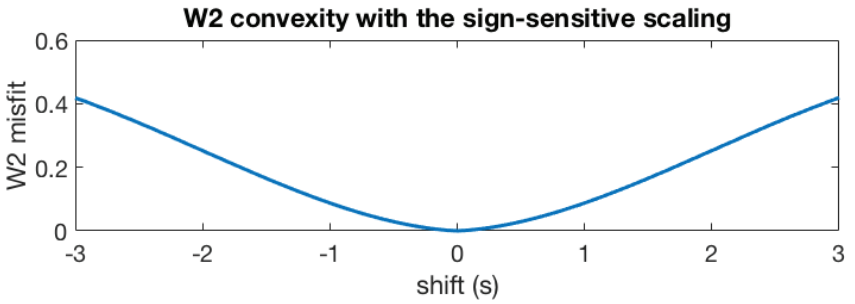


Fig. 10: The W_2 misfit regarding signal shift s by using the sign-sensitive scaling, i.e., $W_2^2(f(t-s), f(t))$.

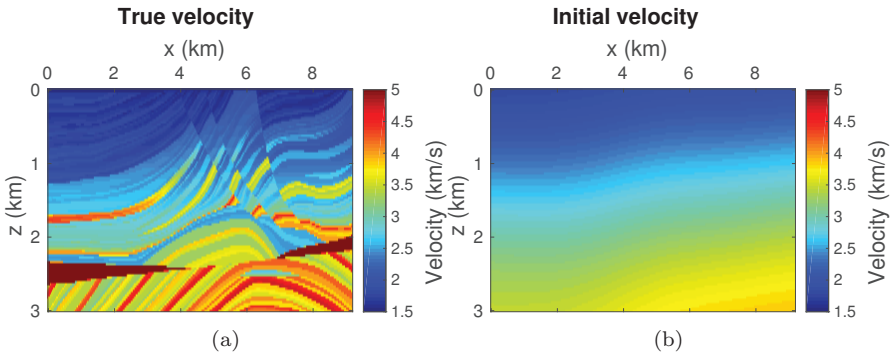


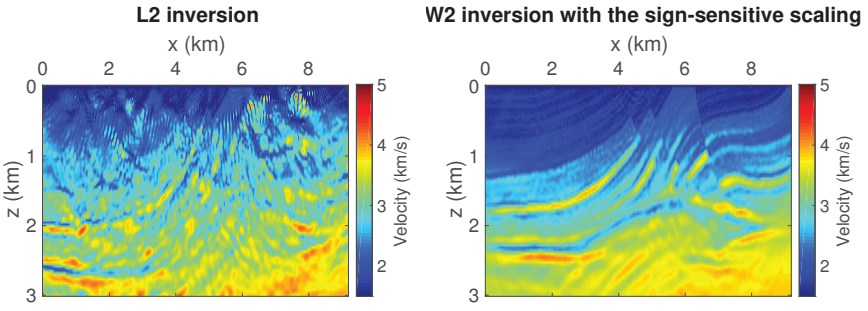
Fig. 11: (a) True velocity and (b) inital velocity for full Marmousi model

the exponential part. In the limit of large c values, f obviously converges to f^+. It is a C^1 function which is compatible with the adjoint-state method. It keeps the convexity of the quadratic Wasserstein distance with respect to signal shifts as shown in Figure 10. One needs to select the coefficient c based on the data range. The sign-sensitive scaling is similar to the exponential scaling (23) by suppressing the negative part of the signal, but it does not have the risk of exaggerating large f and g values in the exponential normalization. The computational cost for all the data normalization methods are the same and negligible in each iteration since the major cost of the inversion is the forward and backward wave propagation.

If we denote the normalization function in (3) as an operator P. One idea to use all the information of the signal (instead of just f^+) is to consider the following objective function:

$$J_4(m) = W_2^2(P(f), P(g)) + W_2^2(P(-f), P(-g)). \tag{24}$$

The final experiment is to invert full Marmousi model by conventional L^2 and trace-by-trace W_2 misfit with $J_4(m)$ as the actual objective function. Figure 11a is the P-wave velocity of the full Marmousi model, which is 3km in depth and 9km in



(a) (b)

Fig. 12: Inversion results of (a) L^2 and (b) trace-by-trace W_2 with the sign-sensitive scaling (24)

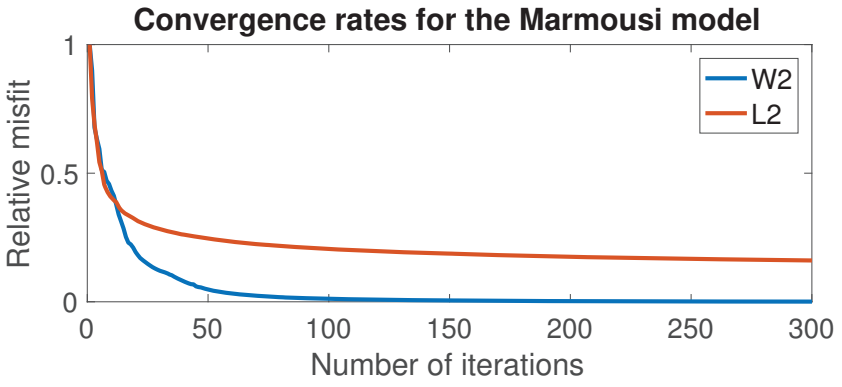


Fig. 13: The convergence rates for the Marmousi model inversion

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 150m depth in the water layer and 307 receivers on top at the same depth with a 30m fixed acquisition. The discretization of the forward wave equation is 30m in the x and z directions and 30ms in time. The source is a Ricker wavelet with a peak frequency of 15Hz, and a high-pass filter is applied to remove the frequency components from 0 to 2Hz. Inversions are terminated after 300 l-BFGS iterations which take about 3 hours on a normal workstation. Figure 12a shows the inversion result using the traditional L^2 least-squares method and Figure 12b shows the final result using trace-by-trace W_2 misfit function. Again, the result of L^2 metric has spurious high-frequency artifacts while W_2 using the sign-sensitive scaling (24) correctly inverts most details in the true model. The convergence rates in Figure 13 illustrates that L^2-based inversion ends up at a local minimum due to the nonzero data misfit after 300 iterations while W_2-based inversion fits the data and reduces the relative data mismatch from 1 to 0.

5. Conclusion. Full waveform inversion for seismic imaging and the application of optimal transport for computing the misfit between simulated and measured data are summarized. Seismic signals need to be transformed by some normalization to satisfy the requirements from optimal transport. Advantages and disadvantages of different normalization techniques are discussed. The dilemma that methods, which have provable desirable properties for simple model problems do not work well in practical large-scale settings and other methods that theoretically fail for simple examples perform very well in practice is illuminated. Quadratic scaling belongs to the first class, and linear scaling belongs to the second class of normalizations, which do very well in realistic tests. A new sign-sensitive normalization aiming at bridging these two classes is introduced, and numerical examples are presented.

REFERENCES

[1] E. Baysal, D. D. Kosloff, and J. W. C. Sherwood, *Reverse time migration*, Geophysics, 48:11 (1983), pp. 1514–1524.

[2] J.-D. Benamou, B. D. Froese, and A. M. Oberman, *Numerical solution of the optimal transportation problem using the Monge-Ampère equation*, Journal of Computational Physics, 260 (2014), pp. 107–126.

[3] Y. Brenier, *Polar factorization and monotone rearrangement of vector-valued functions*, Comm. Pure Appl. Math., 44 (1991), pp. 375–417.

[4] L. A. Caffarelli and R. J. McCann, *Free boundaries in optimal transport and monge-ampere obstacle problems*, Annals of mathematics, pp. 673–730.

[5] J. Chen, Y. Chen, H. Wu, and D. Yang, *The quadratic Wasserstein metric for earthquake location*, arXiv preprint arXiv:1710.10447, 2017.

[6] L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard, *An interpolating distance between optimal transport and Fisher–Rao metrics*, Foundations of Computational Mathematics, pp. 1–44.

[7] J. F. Claerbout, *Toward a unified theory of reflector mapping*, Geophysics, 36:3 (1971), pp. 467–481.

[8] R. Clayton and B. Engquist, *Absorbing boundary conditions for acoustic and elastic wave equations*, Bulletin of the Seismological Society of America, 67:6 (1977), pp. 1529–1540.

[9] W. Dai, P. Fowler, and G. T. Schuster, *Multi-source least-squares reverse time migration*, Geophysical Prospecting, 60: (2012), pp. 681–695.

[10] G. De Philippis and A. Figalli, *The Monge-Ampère equation and its link to optimal transportation*, Oct 2013.

[11] B. Engquist and B. D. Froese, *Application of the Wasserstein metric to seismic signals*, Communications in Mathematical Sciences, 12:5 (2014), pp. 979–988.

[12] B. Engquist, B. D. Froese, and Y. Yang, *Optimal transport for seismic full waveform inversion*, Communications in Mathematical Sciences, 14:8 (2016), pp. 2309–2330.

[13] B. Engquist and A. Majda, *Absorbing boundary conditions for numerical simulation of waves*, Proceedings of the National Academy of Sciences, 74:5 (1977), pp. 1765–1766.

[14] X. Feng, R. Glowinski, and M. Neilan, *Recent developments in numerical methods for fully nonlinear second order partial differential equations*, SIAM Review, 55:2 (2013), pp. 205–267.

[15] A. Figalli, *The optimal partial transport problem*, Archive for rational mechanics and analysis, 195:2 (2010), pp. 533–560.

[16] R. Glowinski, *On alternating direction methods of multipliers: a historical perspective*, in "Modeling, simulation and optimization for science and technology", pp. 59–82. Springer, 2014.

[17] J. A. Hudson and J. R. Heritage, *The use of the Born approximation in seismic scattering problems*, Geophysical Journal International, 66:1 (1981), pp. 221–240.

[18] S. Kondratyev, L. Monsaingeon, D. Vorotnikov, et al., *A new optimal transport distance on the space of finite radon measures*, Advances in Differential Equations, 21:11/12 (2016), pp. 1117–1164.

[19] P. Lailly, *The seismic inverse problem as a sequence of before stack migrations*, in "Conference on inverse scattering: theory and application", pp. 206–220. Society for Industrial and Applied Mathematics, Philadelphia, PA, 1983.

[20] P. Lailly, *Migration methods: partial but efficient solutions to the seismic inverse problem*,

Inverse problems of acoustic and elastic waves, 51 (1984), pp. 1387–1403.
[21] Y. Luo and G. T. Schuster, *Wave-equation traveltime inversion*, Geophysics, 56:5 (1991), pp. 645–653.
[22] E. Mainini, *A description of transport cost for signed measures*, Journal of Mathematical Sciences, 181:6 (2012), pp. 837–855.
[23] P. Moczo, J. OA Robertsson, and L. Eisner, *The finite-difference time-domain method for modeling of seismic wave propagation*, Advances in Geophysics, 48 (2007), pp. 421–516.
[24] G. Monge, *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, pp. 666–704, 1781.
[25] P. Mora, *Inversion = migration + tomography*, Geophysics, 54:12 (1989), pp. 1575–1586.
[26] R.-E. Plessix, *A review of the adjoint-state method for computing the gradient of a functional with geophysical applications*, Geophysical Journal International, 167:2 (2006), pp. 495–503.
[27] L. Qiu, J. Ramos-Martínez, A. Valenciano, Y. Yang, and B. Engquist, *Full-waveform inversion with an exponentially encoded optimal-transport norm*, in "SEG Technical Program Expanded Abstracts 2017", pp. 1286–1290. Society of Exploration Geophysicists, 2017.
[28] G. T. Schuster, *Seismic Imaging, Overview*, pp. 1121–1134. Springer, Dordrecht, 2011.
[29] G. Seismology, E. Bozda, J. Trampert, and J. Tromp, *Misfit functions for full waveform inversion based on instantaneous phase and envelope measurements*, Geophys. J. Int, 185 (2011), pp. 845–870.
[30] A. Tarantola, *Linearized inversion of seismic reflection data*, Geophysical prospecting, 32:6 (1984), pp. 998–1015.
[31] A. Tarantola, *Inverse problem theory and methods for model parameter estimation*, SIAM, 2005.
[32] A. Tarantola and B. Valette, *Generalized nonlinear inverse problems solved using the least squares criterion*, Reviews of Geophysics, 20:2 (1982), pp. 219–232.
[33] L. Van Hove, *Correlations in space and time and Born approximation scattering in systems of interacting particles*, Physical Review, 95:1 (1954), pp. 249.
[34] C. Villani, *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*, American Mathematical Society, Providence, RI, 2003.
[35] J. Virieux, A. Asnaashari, R. Brossier, L. Métivier, A. Ribodetti, and W. Zhou, *6. An introduction to full waveform inversion*. in "Encyclopedia of Exploration Geophysics", pp. R1–1–R1–40. Society of Exploration Geophysicists, Jan 2014.
[36] J. Virieux and S. Operto, *An overview of full-waveform inversion in exploration geophysics*, Geophysics, 74:6 (2009), pp. WCC1–WCC26.
[37] Y. Yang and B. Engquist, *Analysis of optimal transport related misfit functions in seismic imaging*, in "International Conference on Geometric Science of Information", pp. 109–116. Springer, 2017.
[38] Y. Yang and B. Engquist, *Analysis of optimal transport and related misfit functions in full-waveform inversion*, Geophysics, 83:1 (2018), pp. A7–A12.
[39] Y. Yang, B. Engquist, J. Sun, and B. D. Froese, *Application of optimal transport and the quadratic Wasserstein metric to full-waveform inversion*, Geophysics, 83:1 (2017), pp. 1–103.