Full waveform inversion with unbalanced optimal transport distance

Da Li, Michael P. Lamoureux and Wenyuan Liao

ABSTRACT

Full waveform inversion (FWI) has become a major seismic imaging technique. However, using the least-squares norm in the misfit functional possibly leads to cycle-skipping issue and increases the nonlinearity of the optimization problem. Several works of applying optimal transport distances to mitigate this problem have been proposed recently. The optimal transport distance is to compare two positive measures with equal mass. To overcome the mass equality limit, we introduce an unbalanced optimal transport (UOT) distance with Kullback–Leibler divergence to balance the mass difference. An entropy regularization and a scaling algorithm have been used to compute the distance and its gradient efficiently. Two strategies of normalization methods which transform the seismic signals into non-negative functions have been compared. Numerical examples in one and two dimension have been provided.

INTRODUCTION

Full waveform inversion (FWI) is a high resolution seismic imaging algorithm and it was proposed by (Lailly and Bednar, 1983) and (Tarantola, 1984) in the early 1980s. It is a nonlinear PDE-constrained optimization problem with physical properties such as velocity and density of underground as the control parameters, and the waveform received by the receivers as the state parameters. Depending on different physical model, the constraint PDE can be simple wave equation, acoustic wave equation or elastic wave equation. Because of the huge size of the scale, gradient based optimization methods such as gradient descent, l-BFGS and Newton method is needed. And the gradient generally can be achieved by the adjoint state method. With the improvement of the computing power, FWI has been more and more applied in the industry.

In conventional methods, the L_2 distance is used in the misfit functional during optimization to measure the difference between observed and synthetic data. As a nonlinear optimization problem, FWI algorithm suffers the existence of local minima. One of the reason causing the local minima is cycle-skipping issue, which can occurs as the phase difference between two seismic signals is larger then half wavelength. To mitigate this problem, using optimal transport (OT) distances or Wasserstein distance in FWI problem have been proposed recently. The optimal transport distance is to compare two positive measures with equal total mass. When comparing two non-negative equal mass functions, the OT distance will keep monotonically increasing as one function is shifting away from another function. This property provides convexity of OT distance as a misfit functional and it is one of the main reason to introduce OT distance to FWI problem (Engquist and Froese, 2013; Engquist et al., 2016).

However, the seismic signal is oscillating around 0 and usually the condition of equal mass is not satisfied. There are two main strategies that have been proposed to integrate the

OT distance to seismic signals. In the work of (Métivier et al., 2016b,a; Yong et al., 2019), the work is based on the connection between KR norm (Bogachev, 2007) and a special OT distance, 1-Wasserstein distance, to generalize the OT distance to signed measure. The second strategy is to normalize the signals into positive functions with equal mass, and then use the 2-Wasserstein distance to compare the difference and compute gradient (Yang and Engquist, 2017; Yang et al., 2018).

In this work we follow the second strategy to consider the FWI problem with wave equation as the constraint. We introduce the unbalanced optimal transport (UOT) distance to remove the equal mass restriction. Two normalization methods, linear normalization and exponential normalization have been used and compared. To compute the distance and gradient efficiently, an entropy regularization method and a scaling method have been used. In section 2 we give a short review of the optimal transport problem and unbalanced optimal transport problem, we give the algorithm in the end of the section. In section 3, the adjoint state method with UOT distance has been provided. In section 4, numerical examples has been provided to compare the UOT distance and L_2 distance.

BACKGROUND ON OPTIMAL TRANSPORT

In this section, we provide a short review of the optimal transport problem in discrete sense. The definition of unbalanced optimal transport problem with Kullback–Leibler divergence has been provided. A scaling algorithm is used to compute the unbalanced optimal transport distance and its gradient.

Unbalanced optimal transport review

The optimal transport problem has a long history and can date back to 18th century (Monge, 1781). The modern formulation is given by Kantorovich (Kantorovich, 2006). Please refer to (Villani, 2008; Santambrogio, 2015) for a comprehensive review.

For $X, Y \subset \mathbb{R}^d$, the cost function $c(x, y) : X \times Y \to \mathbb{R}_+$ measures the distance between $x \in X$ and $y \in Y$. Given two probability measures $\mu \in P(X)$ and $\nu \in P(Y)$, the Kantorovich formulation of optimal transport problem is defined as

$$\min_{\gamma \in \mathcal{U}(\mu,\nu)} \int_\Omega c(x, y)\, \mathrm{d}\gamma(x, y),$$

where $\mathcal{U}(\mu, \nu)$ is the joint probability measure on $X \times Y$,

$$\mathcal{U}(\mu, \nu) = \left\{\gamma \in P(X \times Y) : \pi^X_\sharp \gamma = \mu,\ \pi^Y_\sharp \gamma = \nu, \right\},$$

the $\pi^X_\sharp$ and $\pi^Y_\sharp$ are the projection operators to X and Y.

We focus on the discrete setting in this work. Let $X = Y = \{x_1, x_2, \cdots, x_N\} \subset \mathbb{R}^d$, $\mu = \sum_i f_i \delta_{x_i}$, $\nu = \sum_i g_i \delta_{x_i}$. Also, we only consider the cost function $c(x, y)$ be the squared Euclidean distance. The optimal transport problem in the discrete form is:

$$\min_{T \in \mathbb{R}^{N \times N}} \langle T, C\rangle = \sum_{i,j=1}^{N} T_{i,j} C_{i,j}, \quad \text{s.t. } T\mathbf{1}_N = f,\ T^T \mathbf{1}_N = g. \tag{1}$$

Here matrix C represents the cost distance defined by $C_{i,j} = |x_i - x_j|^2$.

One of the disadvantages of optimal transport is nonnegative measures with the same total mass are required. To overcome this limitation, the unbalanced optimal transport problem is raised in (Benamou, 2003) in a dynamic approach. Later several works have been proposed in both static and dynamic approach (Piccoli and Rossi, 2014; Chizat et al., 2015, 2018). In this paper we introduce the unbalanced optimal transport distance mainly based on the work in (Chizat et al., 2018). To relax the marginal constraints in (1), we define the unbalanced optimal transport problem as:

$$\min_{T\in\mathbb{R}^{N\times N}} \langle T, C\rangle + F_f(T\mathbf{1}_N) + F_g(T^T\mathbf{1}_N). \tag{2}$$

Both F_f and F_g are proper convex functions.

For example consider:

$$F_f(T\mathbf{1}_N) = \iota_{\{=\}}(T\mathbf{1}_N|f) = \begin{cases} 0, & \text{as } T\mathbf{1}_N = f, \\ \infty, & \text{otherwise.} \end{cases}$$

$$G_f(T^T\mathbf{1}_N) = \iota_{\{=\}}(T^T\mathbf{1}_N|g) = \begin{cases} 0, & \text{as } T^T\mathbf{1}_N = g, \\ \infty, & \text{otherwise.} \end{cases}$$

In this case the equation (2) is equivalent to the optimal transport problem (1). We set

$$F_f(T\mathbf{1}_N) = \varepsilon_m KL(T\mathbf{1}_N|f), \quad F_g(T^T\mathbf{1}_N) = \varepsilon_m KL(T^T\mathbf{1}_N|g),$$

in this paper. Here F_f and F_g are the Kullback-Leibler divergence between the projection of the transport matrix T and f, g, ε_m controls the weight of the mass balancing term in (2).

Regularized primal and dual problem

The entropy regularization is introduced to the optimal transport problem in the work (Cuturi, 2013) to increase the computation efficiency. Then the optimal transport distance with entropy regularization is used in an optimization problem named Wasserstein Barycenter which provides great improvements comparing to L_2 distance (Cuturi and Doucet, 2014; Benamou et al., 2015). In this subsection, the regularized primal and dual problem has been introduced and in next subsection we introduce a coordinate ascent algorithm to solve a regularized version of the problem (2). We define the entropy function for the matrix $T \in \mathbb{R}_+^{N\times N}$ as

$$E(T) = -\sum_{i,j=1}^{N} T_{i,j}\left(\log(T_{i,j}) - 1\right),$$

here we use the convention $0\log(0) = 0$. As we consider the mass balance term F_f and F_g be the Kullback-Leibler divergence, given the regularization parameter ε, the regularized problem (2) can be represented as

$$\min_{T\in\mathbb{R}^{N\times N}} \langle T, C\rangle - \varepsilon E(T) + \varepsilon_m KL(T\mathbf{1}_N|f) + \varepsilon_m KL(T^T\mathbf{1}_N|g) \tag{3}$$

The equation (3) can be rewritten as

$$\min_{T\in\mathbb{R}^{N\times N}} \varepsilon KL(T|K) + \varepsilon_m KL(T\mathbf{1}_N|f) + \varepsilon_m KL(T^T\mathbf{1}_N|g),$$

where

$$KL(T|K) = \sum_{i,j} T_{i,j}\left(\log\left(\frac{T_{i,j}}{K_{i,j}}\right) - 1\right), \quad K_{i,j} = e^{-\frac{C_{i,j}}{\varepsilon}}.$$

We can have the following definition for unbalanced optimal transport distance.

Definition 1 *Define the ground cost matrix C by $C_{i,j} = |x_i - x_j|^2$. Given $f, g \in \mathbb{R}_+^N$, regularization parameter ε and mass balancing parameter ε_m, the regularized unbalanced optimal transport distance with Kullback-Leibler divergence can be defined as*

$$W^2_{2,\varepsilon,\varepsilon_m}(f,g) = \min_{T\in\mathbb{R}^{N\times N}} \varepsilon KL(T|K) + \varepsilon_m KL(T\mathbf{1}_N|f) + \varepsilon_m KL(T^T\mathbf{1}_N|g). \tag{4}$$

Here $KL(\cdot|\cdot)$ is the Kullback-Leibler divergence between two matrices or vectors. And $K_{i,j} = e^{-\frac{C_{i,j}}{\varepsilon}}$.

To compute the unbalanced optimal transport distance, the dual problem of equation (4) is needed.

Theorem 1 *The dual problem of* (4) *is*

$$W^2_{2,\varepsilon,\varepsilon_m}(f,g) = \max_{\phi,\psi\in\mathbb{R}_+^N} \sum_{i,j=1}^{N} -\varepsilon_m f_i\left(e^{-\phi_i/\varepsilon_m} - 1\right) - \varepsilon_m g_i\left(e^{-\psi_i/\varepsilon_m} - 1\right) - \varepsilon K_{i,j}\left(e^{\phi_i/\varepsilon}e^{\psi_i/\varepsilon} - 1\right). \tag{5}$$

Strong duality holds. There exists a unique T^* for the primal problem (4)*. And ϕ^*, ψ^* maximize* (5) *if and only if*

$$T^*_{i,j} = e^{\phi_i^*/\varepsilon} K_{i,j} e^{\psi_j^*/\varepsilon}.$$

This theorem is a straight forward application of Theorem 3.2 in (Chizat et al., 2018).

The scaling algorithm

To compute the unbalanced optimal transport distance, a coordinate ascent method can be used.

Proposition 1 *Suppose ϕ^*, ψ^* solves the dual problem* (5)*, let $u, v \in \mathbb{R}^N$ with $u_i^* = e^{\phi_i^*/\varepsilon}$ and $v_j^* = e^{\psi_j^*/\varepsilon}$. Matirx K , coefficient ε and ε_m is defined in Theorem 1. For $i, j = 1, \cdots, N$:*

$$u_i^* = \left(\frac{f_i}{\sum_j K_{i,j} v_j^*}\right)^{\frac{\varepsilon_m}{\varepsilon_m+\varepsilon}}, \quad v_j^* = \left(\frac{g_j}{\sum_i K_{i,j} u_i^*}\right)^{\frac{\varepsilon_m}{\varepsilon_m+\varepsilon}}.$$

The above proposition can be easily checked by computing the first order optimality condition of dual problem (5). The following remark provides the algorithm to compute the unbalanced optimal transport distance with entropy regularization as Definition 1.

Remark 1 *Given $f, g \in \mathbb{R}_+^N$, cost matrix $C \in \mathbb{R}^{N\times N}$, regularization parameter $\varepsilon > 0$ and mass balancing parameter $\varepsilon_m > 0$. Matrix K is defined as $K_{i,j} = e^{-C_{i,j}/\varepsilon}$. Starting with an initial value $v^{(0)} = \mathbf{1}_N$, the dual problem can be computed through a coordinate ascent algorithm: For the n-th iteration,*

$$u_i^{(n+1)} = \left(\frac{f_i}{\sum_j K_{i,j} v_j^{(n)}}\right)^{\frac{\varepsilon_m}{\varepsilon_m+\varepsilon}}, \quad v_j^{(n+1)} = \left(\frac{g_j}{\sum_i K_{i,j} u_i^{(n+1)}}\right)^{\frac{\varepsilon_m}{\varepsilon_m+\varepsilon}}.$$

Suppose the coordinate ascent algorithm converges with u^*, v^*, the transport matrix T^* in (4) *can be computed as*

$$T_{i,j}^* = u_i^* K_{i,j} v_j^*.$$

Also, the gradient of UOT distance with entropy regularization can be achieved with following remark.

Remark 2 *Suppose T^*, ϕ^* and ψ^* solves the primal problem* (4) *and dual problem* (5)*, the gradient of unbalanced optimal transport distance with respect to f is:*

$$\nabla_{f_i} W_{2,\varepsilon,\varepsilon_m}^2(f, g) = -\varepsilon_m \left(e^{-\phi_i^*/\varepsilon_m} - 1\right).$$

FULL WAVEFORM INVERSION AND GRADIENT COMPUTATION

Since optimal transport problem was proposed for positive measures, normalization is needed before introduce UOT distance for seismic signals. Normalizations with linear and exponential transform are studied in this paper:

$$h_{\text{linear},k}(f) = f + k, \tag{6}$$

$$h_{\text{exp},k}(f) = e^{kf}, \tag{7}$$

The normalization coefficient k is chosen such that $f + k \geq 0$ in equation (6). Here the exponential is taken for each entries.

We consider the full waveform inversion in a discrete form. Suppose $x_i \in \Omega$ is the i-th discrete point of the spatial domain, and t_j for $j = 1, \cdots, N_t$ be the discrete point of time as $t = t_j$. Suppose there are N_s sources and N_r receivers in the physical model. Let $s = 1, \cdots, N_s$, $r = 1, \cdots, N_r$ be the indexes of sources and receivers. Let the observed data be $d_{\text{obs},s,r}$ and synthetic data be $d_{s,r}$.

Consider the wave equation:

$$\frac{1}{c^2(x)}\frac{\partial^2}{\partial t^2} u_s(x,t) - \Delta u_s(x,t) = f_s(x,t), \quad s = 1, \cdots, N_s$$

with suitable initial and boundary conditions, it can be represented in a discrete setting:

$$F[c]u_s(x_i,t_j) = f_s, \quad s = 1, \cdots, N_s. \tag{8}$$

Here $F[c]$ is the finite difference operator with velocity field c, f_s is the s-th source and u_s is the wavefield generated by $F[c]$ and f_s.

The full waveform inversion is a PDE-constrained optimization problem in both state (wavefield) and control (velocity) space. Since the wave equation is well posed with suitable initial and boundary condtions, the FWI problem has a reduced form. We consider the reduced full waveform inversion problem with the unbalanced optimal transport distance defined in Definition 1.

$$\min_c J[c] = \sum_{s=1}^{N_s}\sum_{r=1}^{N_r} W^2_{2,\varepsilon,\varepsilon_m}(h_k(d_{s,r}), h_k(d_{\text{obs},s,r})). \tag{9}$$

Here $J[c]$ is the misfit functional of the reduced optimization problem, $h_k(d_{s,r})$ and $h_k(d_{\text{obs},s,r})$ are seismic traces normalized by (6) or (7) with coefficient k. Also,

$$d_{s,r}(t_j) = P_r u_s[c](x_i,t_j) = u_s[c](x_r,t_j), \quad s = 1, \cdots, N_s,\ r = 1, \cdots, N_r,\ j = 1, \cdots, N_t.$$

Here P_r is the projection operator represents the wavefield as recorded at the r-th receiver location denoted as x_r. The $u_s[c]$ is the solution of the wave equation (8) with velocity c and source f_s.

The gradient of the misfit functional can be computed by the adjoint state method (Plessix, 2006). In this case the gradient $\nabla J[c]$ will be represented as an inner product by the derivative of forward modelling wavefield and adjoint wavefield.

$$\nabla J[c](x_i) = \sum_{s=1}^{N_s}\sum_{j=1}^{N_t} \frac{-2}{c(x_i)^3}\left(\frac{\partial^2}{\partial t^2}u_s(x_i,t_j)\right) v_s(x_i,t_j). \tag{10}$$

The adjoint wavefields v_s are the solutions of the adjoint equations with time reversed,

$$F[c]v_s(x_i,t_j) = \tilde{f}_s, \quad s = 1, \cdots, N_s. \tag{11}$$

The adjoint sources can be computed through Remark 2. For the linear normalization (6), the adjoint sources are

$$\tilde{f}_s = -\sum_{r=1}^{N_r} P_r^T \nabla W^2_{2,\varepsilon,\varepsilon_m}(\hat{d}_{s,r}, \hat{d}_{\text{obs},s,r}), \quad s = 1, \cdots, N_s. \tag{12}$$

And for the exponential normalization:

$$\tilde{f}_s = -\sum_{r=1}^{N_r} P_r^T \left(k e^{k d_{s,r}}\right)^T \nabla W^2_{2,\varepsilon,\varepsilon_m}(\hat{d}_{s,r}, \hat{d}_{\text{obs},s,r}), \quad s = 1, \cdots, N_s. \tag{13}$$

Here the ∇ is the gradient of the first term in UOT distance. As the gradient $\nabla J[c]$ is achieved, gradient based methods or quasi Newton methods such as l-BFGS can be used to minimize (9).

NUMERICAL EXAMPLE

We provide three numerical examples to show the different behavior of optimization with L_2 distance and unbalanced optimal transport distance with linear and exponential normalization.

Shifted Ricker example

We investigate the sensitivity to time shift of Ricker wavelets with L_2 distance and UOT distance in this example. To compare the behavior of different distance, we define a cost function:

$$J_1(s) = \mathrm{d}(f(t-s), g(t-0.5)).$$

The distance d can be L_2 distance, UOT distance with linear and exponential normalization. Here f and g are two Ricker wavelets with center at time 0s and peak frequency 10 Hz, the amplitude of f is 1.2 times of g. The sample frequency is 1000 Hz. The case $f(t-0.7)$ and $g(t-0.5)$ is shown in Figure 1.

We fix g as the reference signal, then move the center of f along time axis from 0.3s to 0.7s. The normalized $J_1(s)$ of L_2 distance, UOT distance with linear normalization and UOT distance with exponential normalization has been shown in Figure 2 (a), (c), (e) respectively. In Figure 2 (a), one global minima and two local minima has been observed which is a sign for the cycle-skipping issue. In Figure 2 (c), the cycle-skipping issue slightly reduced by using UOT distance with linear normalization comparing to L_2 distance. With smaller normalization coefficient k, the better performance can be achieved. However, k can not be less than the absolute value of minimal value of f and g. In Figure 2 (e), as $k = 0.5$, the misfit function is similar to the case of (a) and (c). One global minima has got with $k = 1, 1.5$, the misfit function will increase as the difference of center time increase and the cycle-skipping issue is avoided.

Adjoint sources of the case in Figure 1 are given in Figure 2 (b), (d), (f). The adjoint source in L_2 distance is shown in (b) which is $f(t-0.7) - g(t-0.5)$. The adjoint sources in UOT distance with linear normalization are shown in (d). Different than the L_2 case, the appearance of adjoint sources seems the difference between the envelope of f and g. Similar results can be found in the study of (Yang et al., 2018; Yang and Engquist, 2017; Yong et al., 2019). Also, as the normalization coefficient k decrease, the amplitude of difference between the envelope is increasing. In figure (f), adjoint sources in UOT distance with exponential normalization are shown. As $k = 0.5$, the adjoint source is similar to the case in (d), but great distortion happens as $k = 1.5$. This is the reason in the normalization (7), the interval where $kf(t) > 1$ plays an predominate role in the amplitude of the adjoint source according to the exponential term in (13).

Comparing to L_2 adjoint source, the UOT adjoint sources are more sensitive to the position of wavelets while providing less information of the shape of the wavelets. This behavior brings less large wavenumber components in the gradient which is achieved through adjoint state method (10) and (11). With comprehensive consideration of the behavior of misfit functions and adjoint sources, smaller k is encouraged for UOT distance with linear normalization. In the case of UOT distance with exponential normalization, dedicated k

need to be chosen to make the amplitude of $kf(t)$ close to 1. This effort can largely reduce the cycle-skipping issue and cause less distortion in the adjoint source at the same time.

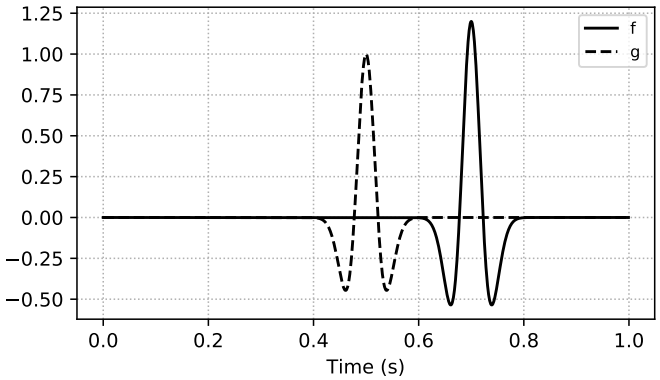

FIG. 1. Ricker wavelet $f(t-0.7)$ and $g(t-0.5)$.

Single layer model

Due to the large size and nonlinear behavior of FWI problem, a single layer model example with 2 coefficients is provided in this section for a detailed insight. We investigate a simple 2D FWI problem in a region with 1km wide and 1km deep, discretize into 101×101 grid points. We use 51 receivers at the top of the region, and a source with 10 Hz Ricker wavelet located at $x = 0.5$km, $z = 0.05$km. Consider the velocity model:

$$c(\delta c, z) = c_0(x, z) + \delta c H(z),$$

here $H(z)$ is the Heaviside step function along z direction. Background velocity $c_0(x, z) =$ 1km/s. Define the objective:

$$J_2(\delta c, z) = J[c(\delta c, z)],$$

where J is defined in (9). The true model of FWI problem is $\delta_c = 0.05, z = 0.51$. We set $\delta c \in [-0.06, 0.12]$ with grid size 0.01 and $z \in [0.45, 0.65]$ with discrete grid size 0.01. Then we evaluate J_2 for each δc and z by using L_2 distance, UOT distance with linear normalization and exponential normalization, results are shown in Figure 3. In Figure 3 (a), (b) and (c), the z axis is the normalized misfit function $J_2(\delta c, z)$.

The L_2 distance case is shown in (a), for this toy example with coefficients δc and z as bad initial values are provided, the convergence can trapped in local minima due to the wrinkles in the surface of misfit function. Comparing to (a), the surface of misfit functions in (b) and (c) have less wrinkles in the surface of objectives. Notice that, the flat areas as δc goes smaller and z goes larger in all three figures may represent the existence of local minima and may affect the convergence speed. However, in this example, the UOT distance with linear or exponential normalization still provides larger region which leads to convergence to global minima.

2D crosshole model

In this part we perform the full waveform inversion in a 2D crosshole model to investigate the behavior of the gradient during minimization of objective. The model width and

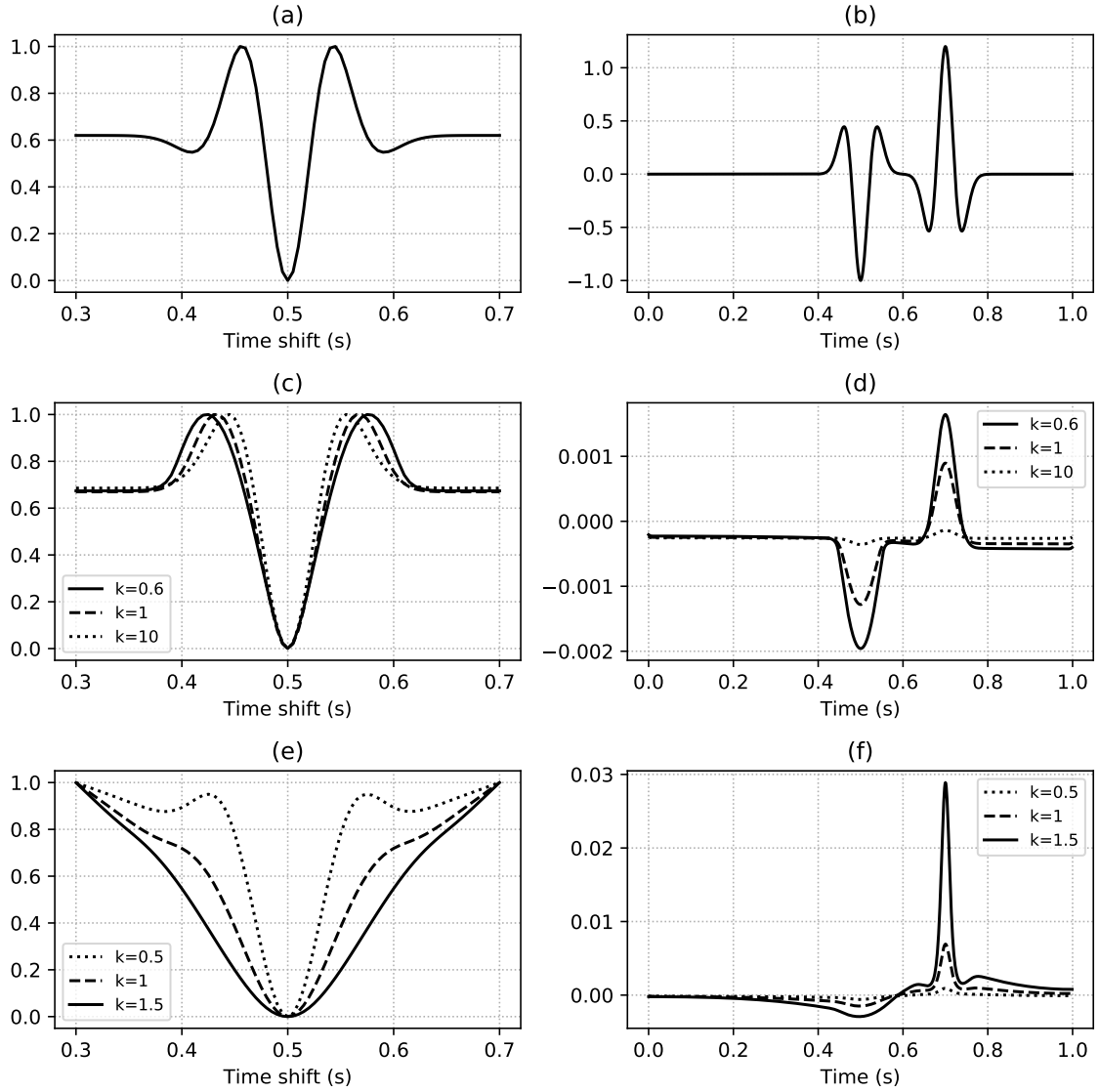

FIG. 2. (a), (b): misfit function and adjoint source using L_2 distance. (c), (d): misfit functions and adjoint sources using UOT distance with linear normalization. (e), (f): misfit functions and adjoint sources using UOT distance with exponential normalization.

depth are 2km with grid size 0.02km. In the true model, background velocity is 3km/s, a single circle anomaly is located at the center of the model, with radius 0.6km and velocity 3.6km/s as shown in Figure 5 (a). There are 11 sources are equally spaced on the left side and 101 receivers on the right side. Synthetic data is generated with the 10 Hz Ricker wavelet and an initial model with homogeneous 3km/s is used in FWI problem.

Figure 4 show the adjoint sources of L_2 distance, UOT distance with linear and exponential normalization respectively. The adjoint sources generated by the UOT distance provide smooth transitions on the positions of reflective seismic waves which will leads the gradients with less high wavenumber components due to (10) and (11). Figure 5 (b), (c), (d) displays the inverse results of L_2 distance, UOT distance with linear and exponential normalization respectively. Gradient descent method is used here and we proceed 5 iterations to show the directions of velocity model updates. All three results describe the presence

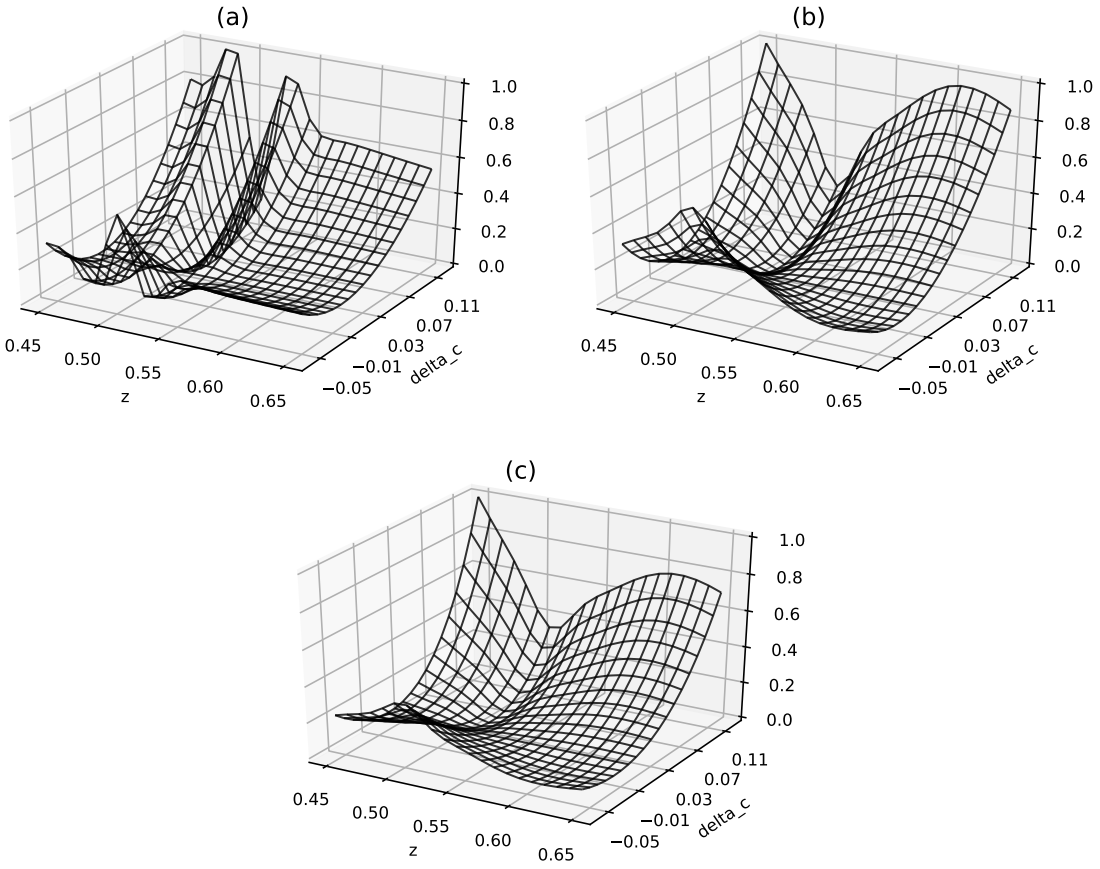

FIG. 3. (a), (b), (c): misfit function by using L_2 distance, UOT distance with linear and exponential normalization respectively.

of the circle anomaly. However, the L_2 result contains abnormal high wavenumber disturbances on the top and bottom of the center. This experiment shows the UOT distance can reduce the risk of wrong velocity updates which may cause the updates be trapped in the local minima.

CONCLUSION AND FUTURE WORKS

The optimal transport based distance and normalization strategies in full waveform inversion problem has already been studied in several works. Numerical experiments have shown the optimal transport based distance is more sensitive with respect to time shift

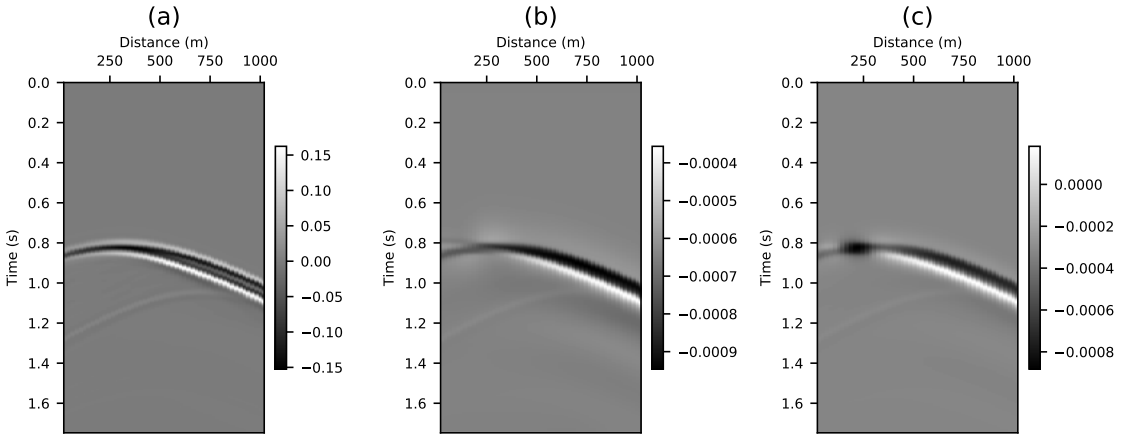

FIG. 4. (a), (b), (c): the adjoint sources generated by the first source in the model with L_2 distance, UOT distance with linear and exponential normalization respectively.

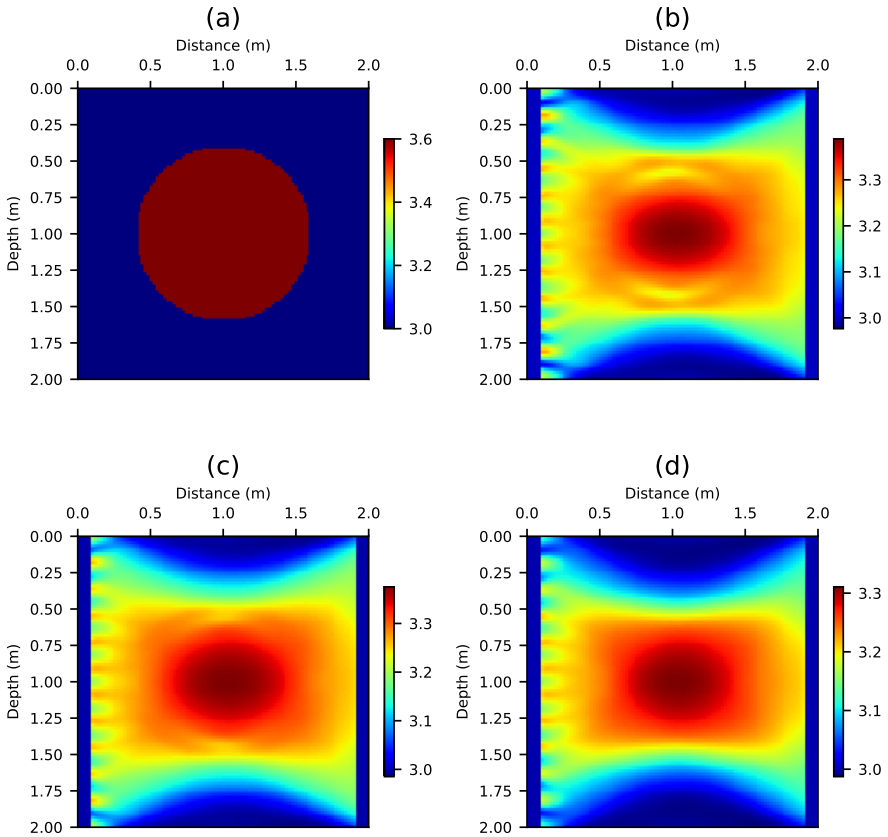

FIG. 5. (a): the true velocity model. (b), (c), (d): inverse results of gradient descent after 5 iterations with L_2 distance, UOT distance with linear and exponential normalization respectively.

comparing to L_2 distance, and that can mitigate the cycle-skipping issue in certain cases. Also, the gradient generated by optimal transport distance provides less large wavenumber

components which can reduce the risk leading to local minima.

There are still problems that need study. First, the optimal transport based distance brings a sub-problem which is controlled by several parameters into the full waveform inversion. Different parameters will have a different impact on the inversion result, therefore it is necessary to study how to set the parameters efficiently. Second, mathematical results are needed to show how the optimal transport distance can mitigate the cycle-skipping issue. Third, more realistic numerical experiments are needed such as Marmousi 2 model or SEG 2014 benchmark data.

ACKNOWLEDGEMENTS

We thank the sponsors of CREWES for continued support. This work was funded by CREWES industrial sponsors, and NSERC (Natural Science and Engineering Research Council of Canada) through the grant CRDPJ 461179-13 and the Discovery grant RGPIN-2015-06038 of the second author. The first author thanks the China Scholarship Council (CSC) for supporting the research.

REFERENCES

Benamou, J.-D., 2003, Numerical resolution of an "unbalanced" mass transport problem: ESAIM: Mathematical Modelling and Numerical Analysis, **37**, No. 5, 851–868.

Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L., and Peyré, G., 2015, Iterative Bregman projections for regularized transportation problems: SIAM Journal on Scientific Computing, **37**, No. 2, A1111–A1138.

Bogachev, V. I., 2007, Measure theory, vol. 1: Springer Science & Business Media.

Chizat, L., Peyré, G., Schmitzer, B., and Vialard, F.-X., 2015, Unbalanced Optimal Transport: Dynamic and Kantorovich Formulation: arXiv preprint arXiv:1508.05216.

Chizat, L., Peyré, G., Schmitzer, B., and Vialard, F.-X., 2018, Scaling algorithms for unbalanced optimal transport problems: Mathematics of Computation, **87**, No. 314, 2563–2609.

Cuturi, M., 2013, Sinkhorn distances: Lightspeed computation of optimal transport, *in* Advances in neural information processing systems, 2292–2300.

Cuturi, M., and Doucet, A., 2014, Fast computation of Wasserstein barycenters, *in* International Conference on Machine Learning, 685–693.

Engquist, B., and Froese, B. D., 2013, Application of the Wasserstein metric to seismic signals: arXiv preprint arXiv:1311.4581.

Engquist, B., Froese, B. D., and Yang, Y., 2016, Optimal transport for seismic full waveform inversion: arXiv preprint arXiv:1602.01540.

Kantorovich, L. V., 2006, On the translocation of masses: Journal of Mathematical Sciences, **133**, No. 4, 1381–1382.

Lailly, P., and Bednar, J., 1983, The seismic inverse problem as a sequence of before stack migrations, *in* Conference on inverse scattering: theory and application, SIAM Philadelphia, PA, 206–220.

Métivier, L., Brossier, R., Merigot, Q., Oudet, E., and Virieux, J., 2016a, An optimal transport approach for seismic tomography: Application to 3D full waveform inversion: Inverse Problems, **32**, No. 11, 115,008.

Métivier, L., Brossier, R., Mérigot, Q., Oudet, E., and Virieux, J., 2016b, Measuring the misfit between seismograms using an optimal transport distance: Application to full waveform inversion: Geophysical Supplements to the Monthly Notices of the Royal Astronomical Society, **205**, No. 1, 345–377.

Monge, G., 1781, Mémoire sur la théorie des déblais et des remblais: Histoire de l'Académie Royale des Sciences de Paris.

Piccoli, B., and Rossi, F., 2014, Generalized Wasserstein distance and its application to transport equations with source: Archive for Rational Mechanics and Analysis, **211**, No. 1, 335–358.

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**, No. 2, 495–503.

Santambrogio, F., 2015, Optimal transport for applied mathematicians: Birkäuser, NY, **55**, 58–63.

Tarantola, A., 1984, Inversion of seismic reflection data in the acoustic approximation: Geophysics, **49**, No. 8, 1259–1266.

Villani, C., 2008, Optimal transport: old and new, vol. 338: Springer Science & Business Media.

Yang, Y., and Engquist, B., 2017, Analysis of optimal transport and related misfit functions in full-waveform inversion: Geophysics, **83**, No. 1, A7–A12.

Yang, Y., Engquist, B., Sun, J., and Hamfeldt, B. F., 2018, Application of optimal transport and the quadratic Wasserstein metric to full-waveform inversion: Geophysics, **83**, No. 1, R43–R62.

Yong, P., Liao, W., Huang, J., Li, Z., and Lin, Y., 2019, Misfit function for full waveform inversion based on the Wasserstein metric with dynamic formulation: Journal of Computational Physics, **399**, 108,911.