Commun. Comput. Phys.
doi: 10.4208/cicp.OA-2024-0230

Vol. **40**, No. 1, pp. 129-152
July 2026

Full-Waveform Inversion with Unbalanced Optimal Transport Metric for Seismic Imaging

Liangrong Wang[1], Xingpeng Dong[2] and Dinghui Yang[1,*]

[1] *Department of Mathematical Sciences, Tsinghua University, Beijing 100084, P.R. China.*
[2] *Institute of Geology, China Earthquake Administration, Beijing 100029, P.R. China.*

Received 27 September 2024; Accepted (in revised version) 30 March 2025

Abstract. Full-waveform inversion (FWI) is an effective method for obtaining high-resolution images of subsurface structures. Conventional full-waveform inversion, which uses the least-square norm L_2 to measure the mismatch between observed and synthetic seismograms, frequently suffers from cycle-skipping and local minimum problems. Derived from optimal transport theory, the Wasserstein metric has been proposed to mitigate cycle-skipping issue. However, due to the requirement for mass conservation, the classical quadratic Wasserstein metric is not ideally suited for FWI applications. In this study, we introduce two unbalanced optimal transport (UOT) distances for use in FWI: the regularized UOT and the unbalanced Sinkhorn divergence. An entropy regularization approach and a truncation approximation are employed to guarantee the efficiency of calculating distances and gradients. These unbalanced optimal transport distances preserve the desirable properties of the quadratic Wasserstein metric, particularly its convexity and insensitivity to noise, while overcoming issues related to mass conservation. We compare the unbalanced optimal transport distances with the L_2 distance and the classical quadratic Wasserstein metric using the Camembert model and the crustal root model. Our numerical experiments demonstrate the superiority of the unbalanced optimal transport distances over traditional methods.

AMS subject classifications: 65K10, 86-08, 86A15, 86A22

Key words: Full-waveform inversion, Wasserstein metric, unbalanced optimal transport, Sinkhorn divergence.

1 Introduction

Full-waveform inversion (FWI), first proposed by Lailly [21] and Tarantola [41], is a data fitting procedure to obtain high-resolution information of subsurface structure. Different

*Corresponding author. *Email addresses:* ydh@tsinghua.edu.cn (D. Yang), wlr20@mails.tsinghua.edu.cn (L. Wang), dongxp@ies.ac.cn (X. Dong)

from ray-based methods [1], FWI estimates subsurface properties by solving the wave equations, thus it can acquire comprehensive information of seismic waves. It is now widely used in seismic tomography [3, 24, 32, 44, 47, 51].

Mathematically, FWI is a nonlinear optimization problem constrained with partial differential equations. The model is iteratively updated by minimizing the misfit function of synthetic data and observed data. Because of the high-dimensionality of model space, FWI is typically solved using gradient-based optimization methods in practice. The gradient of the misfit function is updated using the adjoint-state method [34, 42]. In conventional FWI, the least-square objective function L_2 is used for measuring the data misfit [5, 14, 40, 44, 47]. However, the inverse problem is ill-posed and the L_2 norm is only a point-by-point measurement of the amplitude difference. When the initial model is far from the true model, i.e., the travel time error between synthetic and observed data is more than a half period, an incorrect velocity model will be generated [31]. In other words, the optimization falls into a local minimum. This phenomenon is called cycle-skipping [15, 44], resulting in not to converge to the global minimum. Therefore, FWI employing the L_2 distance requires an accurate enough initial model, which is generally impossible in actual situations.

In the last decades, several approaches have been developed to modify the misfit function to overcome the defect of the L_2 distance. Luo and Schuster [27] first proposed to use the cross-correlation function of synthesis and observations to estimate the traveltime residual as the misfit function. This method is more robust and widely utilized in finite-frequency tomography [42]. However, the cross-correlation measurement requires that the signals have approximately the same shape. To address this issue, a misfit function based on deconvolution has been developed [26, 46]. In this approach, the traveltime misfit is computed by deconvolving the synthetic data with the observed data, rather than relying on cross-correlation. Another strategy is the envelope misfit [25, 50], which measures the instantaneous phase and amplitude envelope using the Hilbert transform. More recently, Dong and Yang [8] have redefined the traditional L_2 norm by incorporating a time shift determined by cross-correlation within the synthetic waveform, resulting in phase-sensitive full-waveform tomography. These strategies aim to improve the posedness of the FWI problem by constructing more convex misfit functions. However, these misfit functions essentially measure the L_2 distance between preprocessed data. Although these methods relax in some manner the dependency on the accuracy of the initial model, cycle-skipping may still occur [28]. Recently, Engquist and Froese [11] first applied optimal transport (OT) mapping to seismic tomography. The Wasserstein metric, derived from OT [43], is defined as the misfit function for FWI. The optimal transport problem was proposed by Monge in 1781, in order to search for the optimal way of transporting sand. Kantorovitch [16] relaxes the original nonlinear problem into a linear optimization problem with convex constraints. The objective function of the problem is known as the Wasserstein metric. Because OT-based techniques can incorporate differences in spatial information, OT-based measures have recently gained widespread use in various applications, including image retrieval [23], signal and image

representation [18,19], inverse problems [11,37], and machine learning [10,18,39]. The Wasserstein metric exhibits inherent convexity with respect to time-shift and dilation, and it is insensitive to noise [12]. Therefore, OT-based distances become appealing in FWI [9,29–31,36,47,48].

Three primary categories of OT-based distances are applied to FWI. The first category is balanced OT [11], i.e., the classical quadratic Wasserstein metric. This method is the most commonly utilized due to the closed-form solution for one-dimensional OT. However, the balanced OT-based distances can only be applied to probability measures, which indicates that the predicted and observed data should be positive and have the same mass equal to one. Consequently, a nonlinear transform and a normalization should be applied to convert raw seismic data to a probability distribution. In this context, several encoding methods are discussed, including linear [48], quadratic [4], exponential transforms [35], softplus encoding [13,36], and splitting the signals into a positive and a negative part [11]. The conversion may disrupt the convexity with respect to large time-shift and alter the phase and amplitude information. The other two categories are proposed to avoid signal transformation and normalization. The first of these two techniques is the Kantorovich-Rubinstein (KR) metric [31,49], which is derived from the dual formulation of the optimal transport problem with the L_1 cost function. The KR metric is defined for signed measures, allowing its direct use in FWI. Métivier et al. [31] proposed a simultaneous direction method of multipliers iterative algorithm to compute the KR metric in a multi-dimensional case. Thus, the KR metric can compare the seismic data source gather by source gather, which can enhance the convexity of the FWI, rather than trace by trace. The main drawback of the KR metric approach is the lack of guarantee on the convexity with respect to large time-shift. Graph-space OT [28,30] is another category. The discrete graphs of the data are compared instead of the raw data itself. Each one-dimensional time signal is transformed as a point cloud in a two-dimensional time-amplitude space, ensuring that the compared data is a probability measure. This transformation preserves the phase and amplitude information of the data, and the resultant misfit function exhibits promising convexity with respect to time-shift. Nevertheless, this method extends the dimensionality and significantly raises the computational cost.

To overcome the mass balance limitation of the classical quadratic Wasserstein metric, Benamou [2] proposed an unbalanced optimal transport (UOT) problem. Chizat et al. [6] designed an efficient numerical solution for it. The data compared using the UOT-based distances do not need to adhere to mass conservation. Furthermore, the UOT-based distances retain convexity with respect to time-shift and insensitivity to noise. In this study, we introduce two UOT-based metrics to FWI. Entropy regularization promises the computational efficiency of the regularized UOT distance. Nonetheless, the solution to the entropy UOT problem is only an approximation of the true solution, and the regularized UOT distance is typically non-zero for two identical signals. This could result in a slow or even impossible inversion in certain unique cases where the relative error decreases by a small amount. Accordingly, we consider an unbalanced Sinkhorn divergence [7,38], which is essentially a modification of the regularized UOT metric. While its minimum is

zero, the unbalanced Sinkhorn divergence inherits the convexity and smoothness of the regularized UOT distance. Therefore, we also employ the unbalanced Sinkhorn divergence in FWI.

In this study, we firstly recall the basic concept of FWI and the OT theory. Then, we introduce the regularized UOT distance and the unbalanced Sinkhorn divergence. The distances and their gradients are evaluated under the entropy regularization method. To decrease the computational costs, we create a truncation approximation method. Subsequently, we use the Camembert model and the crustal root model to demonstrate the benefits of the UOT-based metrics. Finally, we discuss the differences between the OT and UOT in terms of mass transport, as well as the parameter choices for the UOT-based metrics.

2 Method

2.1 Formulation of full-waveform inversion

The goal of FWI is to minimize the misfit function of synthetic seismograms and observed data so that we can reconstruct the subsurface geologic structures. The FWI problem can be written as [14]

$$\min_{\boldsymbol{m}\in\mathfrak{M}} \mathcal{J}(\boldsymbol{m}) = \mathcal{D}(d_{cal}(\boldsymbol{m}), d_{obs}), \tag{2.1}$$

where $\boldsymbol{m}$ denotes the model parameters, such as density and velocity. d_{cal} and d_{obs} are the synthetic data and observed data, respectively. The function $\mathcal{D}$ is the misfit function that measures the difference between d_{cal} and d_{obs}. The optimization problem (2.1) is constraint by

$$L(\boldsymbol{m})u = s, \quad d_{cal}(\boldsymbol{m}) = Ru(\boldsymbol{m}), \tag{2.2}$$

where $L(\boldsymbol{m})$ is a wave propagation operator, s is the source term, and $u(m)$ is the solution to the wave equation. The extraction operator R maps the incident wavefield $u(\boldsymbol{m})$ to the receiver locations.

The gradient of an objective functional with respect to the model parameters is crucial for updating the model. However, computing the gradient of $\mathcal{J}(\boldsymbol{m})$ by definition is computationally expensive and is practically impossible for the large number of model parameters. This is where the adjoint-state method comes into play [34,42]. The gradient can be given by

$$\nabla\mathcal{J}(\boldsymbol{m}) = \left\langle \frac{\partial L}{\partial \boldsymbol{m}} u, v \right\rangle, \tag{2.3}$$

where v is the solution of the adjoint equation

$$L^{\dagger}(\boldsymbol{m})v = s^{\dagger}. \tag{2.4}$$

The adjoint source $s^{\dagger}$ satisfies

$$s^{\dagger} = -R^{\dagger}\left(\frac{\partial \mathcal{D}}{\partial d_{cal}}\right). \tag{2.5}$$

Through the above formulation, one observes that only the adjoint wavefield v depends on the specific form of the misfit function $\mathcal{J}$. Furthermore, when the misfit function is altered, only the adjoint source $s^\dagger$ needs to be modified in the framework of the FWI.

In the conventional FWI with L_2 misfit function, we can obtain its adjoint source

$$s^\dagger = -R^\dagger(d_{cal} - d_{obs}). \tag{2.6}$$

Once the gradient $\nabla\mathcal{J}$ is determined, Eq. (2.1) can be solved by gradient-based optimization methods.

2.2 OT theory

The optimal transport problem seeks the minimum cost of rearranging one distribution into the other. For two probability distribution functions f and g, the OT problem can be formulated as [43]

$$\inf_{T\in\mathcal{M}} \int_\Omega c(x,T(x))f(x)\mathrm{d}x, \tag{2.7}$$

where $\mathcal{M}$ indicates all maps that rearrange f into g, and $c(x,T(x))$ denotes the cost of transporting one unit mass from x to $T(x)$. The solution of Eq. (2.7) is called optimal transport map, and the infimum of Eq. (2.7) is called the Wasserstein distance, which can measure the distance between two distributions. Because f and g are probability distributions, they should satisfy mass conservation

$$\int_X f(x)\mathrm{d}x = \int_Y g(y)\mathrm{d}y = 1. \tag{2.8}$$

When the cost function $c(x,y) = \|x-y\|_2^2$, the infimum $W_2^2(f,g)$ is so-called the quadratic Wasserstein distance.

$$W_2^2(f,g) = \inf_{T\in\mathcal{M}} \int_\Omega \|x-T(x)\|_2^2 f(x)\mathrm{d}x. \tag{2.9}$$

Especially, Eq. (2.9) has a closed form in one dimension case [43]. Denote the cumulative distribution function of f and g as follows

$$F(x) = \int_{-\infty}^{x} f(t)\mathrm{d}t, G(y) = \int_{-\infty}^{y} g(t)\mathrm{d}t. \tag{2.10}$$

Then the optimal map from f to g is

$$T(x) = G^{-1}(F(x)). \tag{2.11}$$

Thus, if f and g are supported on $[0,T_0]$, the Wasserstein distance can be written as

$$W_2^2(f,g) = \int_0^{T_0} |x-G^{-1}(F(x))|^2 f(x)\mathrm{d}x, \tag{2.12}$$

which is the most commonly used OT-based distance in FWI [9, 12, 48]. The Fréchet gradient of $W_2^2(f,g)$ with respect to f is [4]

$$\frac{\partial W_2^2(f,g)}{\partial f} = 2\int_0^x (t-T(t))\mathrm{d}t. \tag{2.13}$$

Because $F(x)$ and $G(y)$ are monotone increasing, Eq. (2.11) can be efficiently calculated in $\mathcal{O}(n)$ complexity by binary search, and n is the number of data samples [48].

In high dimensions, the simple formulation does not work. We should solve the Monge-Ampère equation numerically to compute the optimal map, which is more difficult than solving the wave equation (2.2) sometimes. Therefore, one-dimensional OT has a more extensive application than high-dimensional form in FWI.

In seismology, the Wasserstein metric cannot be applied directly. The seismic signal $f(t)$ and $g(t)$ should be positive and satisfy the mass conservation. Thus, a preprocessing of the seismic signals is needed to assure that the signals are strictly positive and have unit mass. The signals should be normalized as

$$\tilde{f} = \frac{P(f)}{<P(f)>}, \quad \tilde{g} = \frac{P(g)}{<P(g)>}, \tag{2.14}$$

where P is a function that transform f into positive and $<f>=\int_X f(x)\mathrm{d}x$ denotes the total mass of f. Several approaches have been proposed to convert signals into distributions [13]. However, the mass of f and g are generally unequal, resulting in the signals being normalized on different scales, which leads to a loss of amplitude information.

2.3 Unbalanced optimal transport

Traditional optimal transport requires prescaling seismic data that satisfy the mass conservation, which changes the amplitude information of the data. Different from the OT, the unbalanced optimal transport avoids the rigorous constraint. This section will briefly introduce the UOT problem and the scaling algorithm for the numerical evaluation, which is mainly based on the work of Chizat et al. [6].

Assume that $X=\{x_1,\cdots,x_n\}\in(\mathbb{R}^d)^n$ and $Y=\{y_1,\cdots,y_n\}\in(\mathbb{R}^d)^n$ are discrete finite spaces. Two discrete probability measures f and g are defined as the sum of Dirac masses

$$f(x)=\sum_{i=1}^n f_i\delta(x-x_i), g(y)=\sum_{j=1}^n g_j\delta(y-y_j). \tag{2.15}$$

Define the cost matrix $C\in\mathbb{R}^{n\times n}$ as

$$C_{ij}=c(x_i,y_j)=\|x_i-y_j\|_2^2. \tag{2.16}$$

The unbalanced optimal transport between f and g is [6]

$$\min_{P\in\mathbb{R}_+^{n\times n}} \langle C,P\rangle + \lambda \mathrm{KL}(P1_n|f) + \lambda \mathrm{KL}(P^{\mathrm{T}}1_n|g), \tag{2.17}$$

where $\langle C,P\rangle=\sum_{i,j}C_{ij}P_{ij}$, λ is the parameter controlling the cost of unbalanced mass relaxation and KL is the Kullback-Leibler divergence defined by

$$\mathrm{KL}(r|s)=\sum_{i=1}^{n}r_i\log(\frac{r_i}{s_i})-r_i+s_i,\quad r,s\in\mathbb{R}_+^n. \tag{2.18}$$

Let $\lambda\to+\infty$, the UOT problem degenerates to classical OT. The value of Eq. (2.17) is the quadratic unbalanced optimal transport distance W_λ. The entropy regularization is implemented to evaluate the UOT distance, which yields

$$\min_{P\in\mathbb{R}_+^{n\times n}}\langle C,P\rangle+\varepsilon\mathrm{Ent}(P)+\lambda\mathrm{KL}(P1_n|f)+\lambda\mathrm{KL}(P^{\mathrm{T}}1_n|g), \tag{2.19}$$

where $\varepsilon>0$ is the regularization parameter and the entropy function is defined as

$$\mathrm{Ent}(P)=\sum_{i,j}P_{ij}(\log P_{ij}-1). \tag{2.20}$$

A scaling algorithm [6], which is similar to the Sinkhorn algorithm, is proposed to solve (2.19):

$$u_i^{(k+1)}=\left(\frac{f_i}{(Kv^{(k)})_i}\right)^{\lambda/(\lambda+\varepsilon)},\quad v_j^{(k+1)}=\left(\frac{g_j}{(K^{\mathrm{T}}u^{(k+1)})_j}\right)^{\lambda/(\lambda+\varepsilon)}. \tag{2.21}$$

The components of the kernel matrix K are $K_{ij}=\mathrm{e}^{-C_{ij}/\varepsilon}$. The initializations are $u^{(0)}=v^{(0)}=1_n$. When the iteration is terminated, we can obtain optimal transport plan

$$P_\varepsilon^*=\mathrm{diag}(u^*)K\mathrm{diag}(v^*). \tag{2.22}$$

The regularized UOT distance is computed as

$$W_{\varepsilon,\lambda}(f,g)=\langle C,P_\varepsilon^*\rangle+\lambda\mathrm{KL}(P_\varepsilon^*1_n|f)+\lambda\mathrm{KL}(P_\varepsilon^{*\mathrm{T}}1_n|g). \tag{2.23}$$

The Fréchet gradient of $W_{\varepsilon,\lambda}(f,g)$ with respect to f is

$$\left(\nabla_f W_{\varepsilon,\lambda}(f,g)\right)_i=\lambda\left(1-\frac{(P_\varepsilon^*1_n)_i}{f_i}\right)=\lambda\left(1-\frac{u_i^*(Kv^*)_i}{f_i}\right)=\lambda\left(1-(u_i^*)^{-\varepsilon/\lambda}\right). \tag{2.24}$$

The convergence of the iteration (2.21) becomes slow as $\varepsilon\to0$ [6]. Besides, storing the dense kernel K and computing matrix multiplications during the iterations (2.21) requires a lot of memory and time, especially since we need to repeat this process for each receiver in FWI. Therefore, some remedies are needed to reduce computational complexity, otherwise it will be almost impossible to use for practical calculations. In actuality, the sets of sample points X and Y are time series with a fixed time step in FWI. When ε is small, most of the elements in the kernel matrix K have extremely small values, indicating that K is approaching a sparse matrix. Since sparse matrices are more advantageous in terms

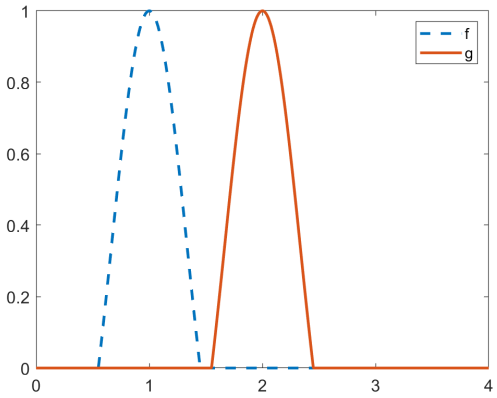


Figure 1: One-dimensional densities f and g.

of storage and computational efficiency, we propose a truncation approximation of the Sinkhorn algorithm based on sparsity.

The cost matrix C is defined as a symmetric Toeplitz matrix in FWI. Thus, only one column of C needs to be stored, as does K. For a threshold parameter $\eta>0$, we introduce

$$\overline{K}_{ij}=\begin{cases}K_{ij}, & K_{ij}\geqslant\eta,\\ 0, & K_{ij}<\eta.\end{cases} \tag{2.25}$$

The first column of $\overline{K}$ is $(k_1,k_2,\cdots,k_m,0,\cdots,0)^{\mathrm{T}}$, indicating that only an m-dimensional vector needs to be stored. The matrix-vector operation in the iteration (2.21) is now transformed into an nth vector-vector operation, reducing its computational complexity from $\mathcal{O}(n^2)$ to $\mathcal{O}(mn)$. η can be assigned a very small value, such as $\eta=1/n^2$, and ε is also small. Therefore, $m\ll n$ always holds, meaning that the computational complexity of the iteration (2.21) is close to $\mathcal{O}(n)$ in this scenario. The error inflicted by the truncation is negligible, which is easy to prove.

Consider two density functions as shown in Fig. 1, the CPU time required for calculating $W_{\varepsilon,\lambda}(f,g)$ at different numbers of sampling points is listed in Table 1. The approximation algorithm reduces the CPU time to approximately $1/8$ of the original. Moreover, the scaling iteration (2.21) can also be accelerated with GPUs. Thus, although calculating the UOT distance involves solving an optimization problem, only a small additional computational burden occurs compared to L_2 and W_2 after truncation approximation.

Similar to the balanced OT, preprocessing is needed to extend the UOT distances to seismic signals. In this case, the normalization is unnecessary, so the amplitude information of the data is maintained. The signals are transformed as

$$\tilde{f}=P(f),\quad \tilde{g}=P(g). \tag{2.26}$$

Table 1: The CPU time required for calculating $W_{\varepsilon,\lambda}(f,g)$ before and after applying the truncation approximation algorithm.

Sampling points	1000	2000	3000	5000
Origin(s)	0.41	2.5	5.6	17.6
After truncation(s)	0.21	0.36	0.70	2.15

2.4 Unbalanced Sinkhorn divergence

The entropy regularization introduces a bias increasing with ε. The regularized UOT distance $W_{\varepsilon,\lambda}$ is an approximation of the UOT distance W_λ. It is worth noting that $W_{\varepsilon,\lambda}(f,f)\neq 0$ for $\varepsilon>0$ in general. When $W_{\varepsilon,\lambda}$ is selected as the objective function of the optimization, the convergence may become slow or even terminate early in certain cases where the relative error decreases only by a small amount. To handle it, we consider the unbalanced Sinkhorn divergence [7,38]

$$S_{\varepsilon,\lambda}(f,g)=W_{\varepsilon,\lambda}(f,g)-\frac{1}{2}W_{\varepsilon,\lambda}(f,f)-\frac{1}{2}W_{\varepsilon,\lambda}(g,g). \tag{2.27}$$

The unbalanced Sinkhorn divergence can be viewed as a modification of the regularized UOT distance. It is obvious that $S_{\varepsilon,\lambda}(f,f)=0$ for any $\varepsilon>0$. Apart from it, $S_{\varepsilon,\lambda}$ is a better approximation of W_λ than $W_{\varepsilon,\lambda}$, while preserving both convexity and smoothness [38]. Different from the unbalanced Sinkhorn divergence presented in [38], we eliminate the term that calculates the mass difference between signals to avoid destroying convexity.

The Fréchet gradient of $S_{\varepsilon,\lambda}(f,g)$ with respect to f is formulated as

$$(\nabla_f S_{\varepsilon,\lambda}(f,g))_i=(\nabla_f W_{\varepsilon,\lambda}(f,g))_i-\frac{1}{2}(\nabla_f W_{\varepsilon,\lambda}(f,f))_i, \tag{2.28}$$

where $\nabla_f W_{\varepsilon,\lambda}(f,g)$ is the same as (2.24) and

$$\big(\nabla_f W_{\varepsilon,\lambda}(f,f)\big)_i=2\lambda\left(1-(u_i^*)^{-\varepsilon/\lambda}\right). \tag{2.29}$$

u^* in Eq. (2.29) is the optimal vector generated when computing $W_{\varepsilon,\lambda}(f,f)$, which is different from u^* in (2.24).

Each calculation of the unbalanced Sinkhorn divergence requires two calculations of the regularized UOT since g is observed data and $W_{\varepsilon,\lambda}(g,g)$ only needs to be computed once in the inversion.

3 Numerical experiments

Three numerical experiments are provided in this section. We compare the numerical results generated by L_2 distance, classical quadratic Wasserstein metric W_2^2, regularized

UOT distance $W_{\varepsilon,\lambda}$, and unbalanced Sinkhorn divergence $S_{\varepsilon,\lambda}$ to demonstrate the effectiveness of the UOT-based distances in FWI. The softplus encoding method [36] is used to preprocess signals for both W_2^2 and UOT-based distances. In the calculation, we consider the isotropic medium. We use SPECFEM2D [20, 32] to obtain the numerical solutions of the wave equation and the associated adjoint-state equation. In the following, we will use OT, RUOT, and USD to indicate W_2^2, $W_{\varepsilon,\lambda}$ and $S_{\varepsilon,\lambda}$, respectively.

3.1 Shifted Ricker example

First, we consider a simple 1D test case similar to that proposed by Engquist and Froese [11]. In Fig. 2a, two Ricker wavelet signals with a peak frequency of 10 Hz are presented. We regard these two signals as observed data $g(t)$ and synthetic data $f(t;s)=g(t-s)$. To evaluate the UOT-based distances, we set $\lambda=1$ and $\varepsilon=10^{-3}$. The misfit functions between two signals as functions of time shift s in different metrics are shown in Fig. 2b. For better comparison, we normalize the misfit using its maximum value. The result illustrates that the L_2 distance has some local minima and is constant when the time shift is large. On the contrary, the other three distances are uniformly convex with respect to the time shift. The OT-based distances transfer the mass from the synthetic data to the observed data to correct the phase difference between two signals [48]. However, the L_2 norm is only a point-by-point measurement of the amplitude difference, which leads to cycle-skipping. We observe that the minimum of $W_{\varepsilon,\lambda}$ is non-zero and $S_{\varepsilon,\lambda}$ is more convex than $W_{\varepsilon,\lambda}$, which is mentioned in the previous section. In this example, two signals have the same mass, so the behavior of W_2^2 is similar to two UOT-based distances.

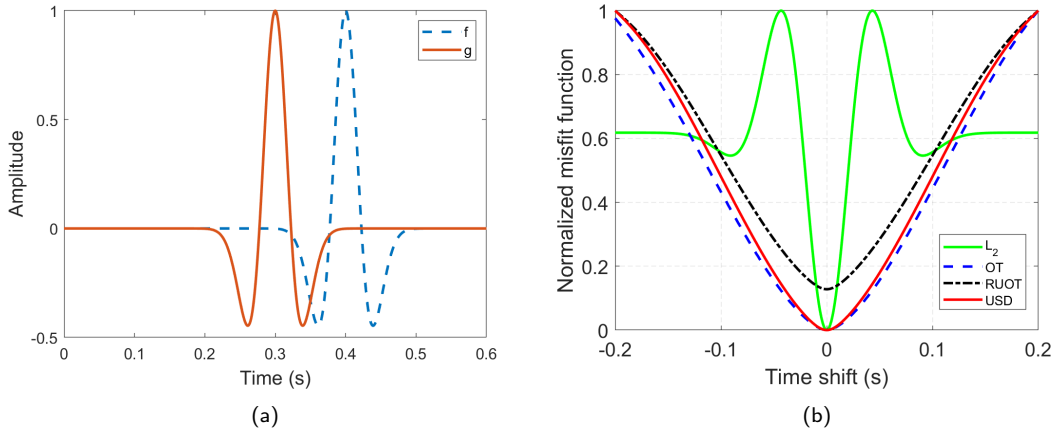


Figure 2: (a) Synthetic signal f and observed signal g. (b) The normalized misfit function with the L_2 distance, the classical quadratic Wasserstein metric, the regularized UOT and the unbalanced Sinkhorn divergence.

3.2 The Camembert model

We perform the FWI with direct waves in a Camembert model [15] to investigate the behavior of four misfit functions. The computational domain is 2km$\times$2km. The background velocity is 3km/s and an abnormal velocity of 4km/s inside the circle with a radius of 0.4km. Fig. 3a shows the true model. A homogeneous initial model is built with the background velocity. In Fig. 3a, the triangles and stars indicate the receivers and sources, respectively. 21 equally spaced sources on the bottom at a depth of 1900m and 101 equally spaced receivers on the top at 100m depth are fixed. A Ricker wavelet with a peak frequency of 10 Hz is used as the source time function. We set $\lambda=0.2$ and $\varepsilon=10^{-4}$ for the RUOT and the USD. The PML boundary conditions are used outside the computational domain. The L-BFGS method is used as the optimization method for the inversion. We select a source and receiver pair located in the middle of the computational domain and plot the synthetic and observed waveforms in Fig. 3b. The synthetic and observed waveforms exhibit a significant phase difference, which could potentially lead to cycle-skipping.

Fig. 4 displays the inversion results obtained with L_2 distance, OT, RUOT, and USD, respectively. The convergence curves are depicted in Fig. 5, where the data misfit is the objective function of the inversion, and the model misfit represents the error between the reconstructed model and the true model. The result obtained by the L_2 misfit is incorrect and is trapped in a local minimum because of cycle-skipping. The other results are close to the true model. The associated data misfits are reduced by more than 90% within ten iterations. Table 2 presents the total computational cost of the inversions and the model error after 40 iterations. Since the RUOT inversion converges after 29 iterations, its execution time is much smaller than the others. In this example, the inversions using the

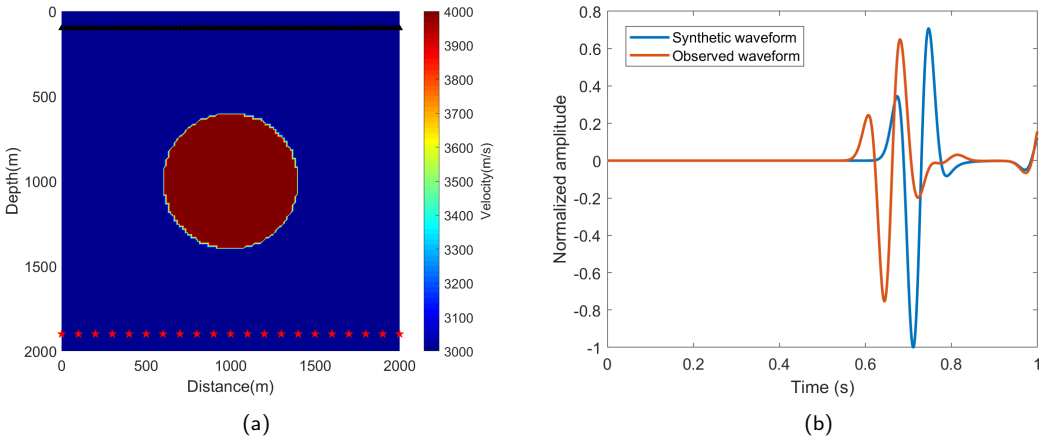


Figure 3: (a) True velocity for the Camembert model with a high-velocity anomaly. The triangles and stars indicate the receivers and sources, respectively. (b) The synthetic and observed waveforms.

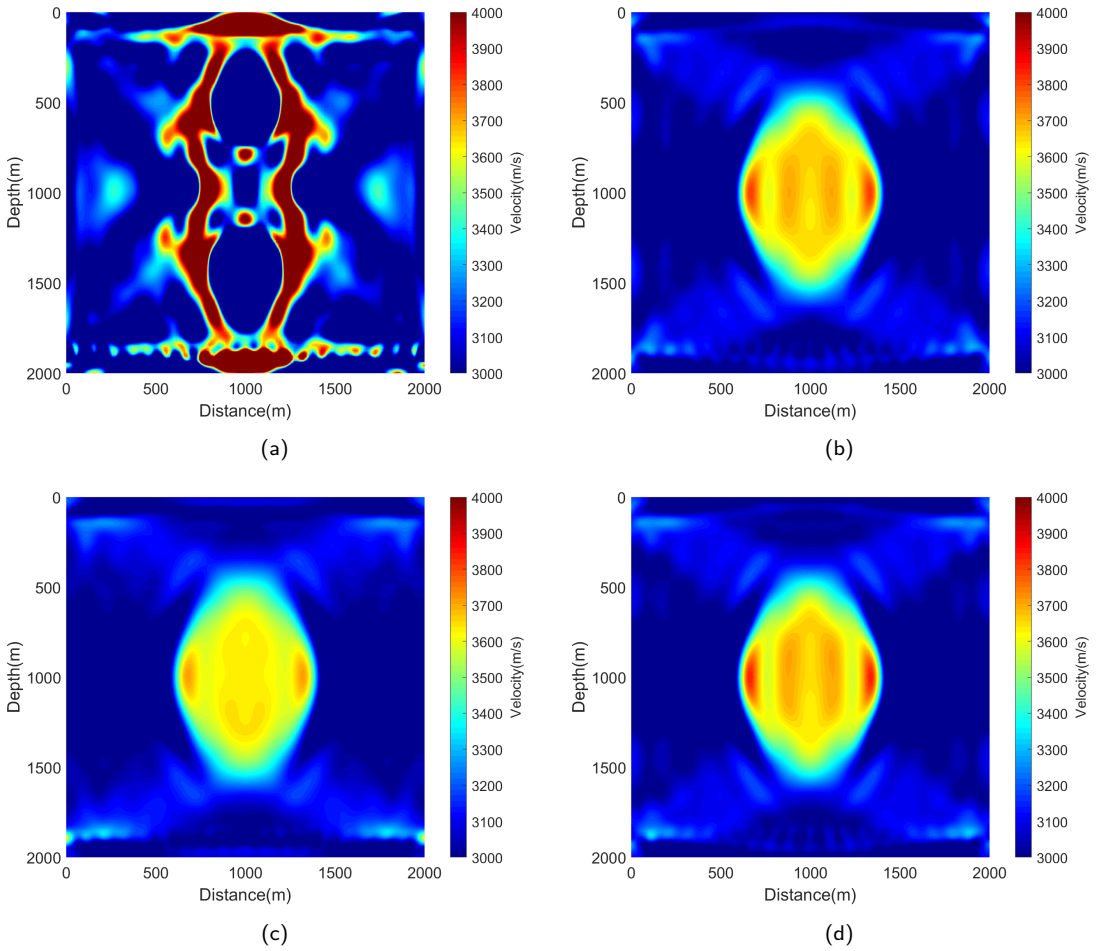


Figure 4: Inversion results of (a) L_2, (b) W_2^2, (c) $W_{\varepsilon,\lambda}$ and (d) $S_{\varepsilon,\lambda}$ for the Camembert model with a high-velocity anomaly.

Table 2: The total execution time and the model misfit after 40 iterations for the Camembert model with a high-velocity anomaly.

Method	L_2	OT	RUOT	USD
Time(min)	549	552	365	439
Model misfit	5.933	0.2736	0.3168	0.2755

OT and the UOT-based measures both yield good results, but the inversions employing the UOT-based methods perform better in computational efficiency.

Next, we modify the high-velocity anomaly to a low-velocity anomaly. The velocity inside the circle is 2.4km/s. Fig. 6a shows the true model. The initial model and other

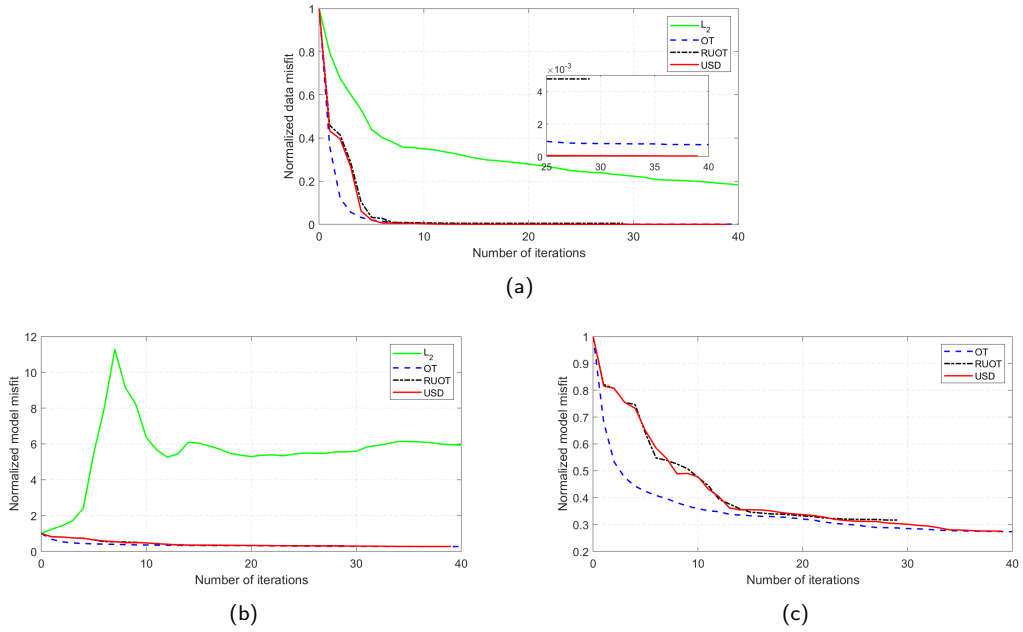


Figure 5: Convergence curves of the Camembert model with a high-velocity anomaly. (a) Data misfit, (b) model misfit and (c) model misfit without L_2.

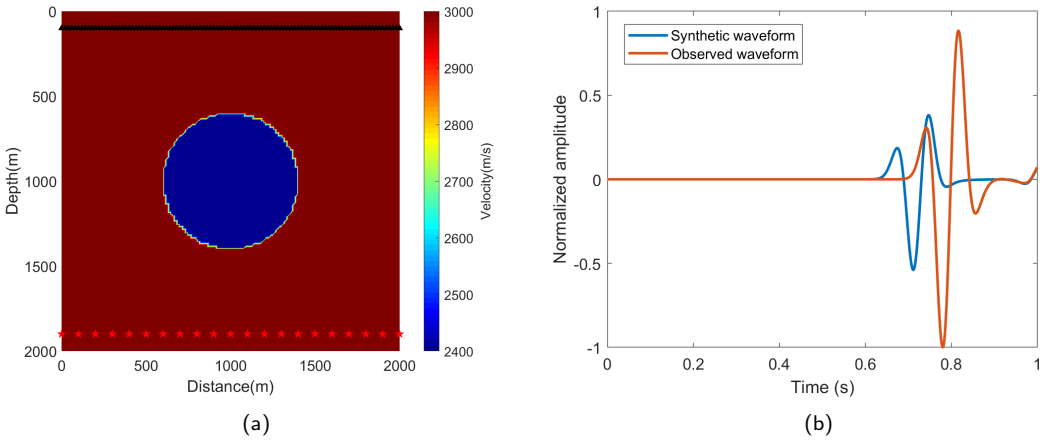


Figure 6: (a) True velocity for the Camembert model with a low-velocity anomaly. The triangles and stars indicate the receivers and sources, respectively. (b) The synthetic and observed waveforms.

parameters are set the same as before. The synthetic waveforms using the true and initial velocities are shown in Fig. 6b. The difference between the two waveforms is larger compared to Fig. 3b, and these two waveforms have a significant disparity in amplitude. The inversion results generated by conducting different methods are shown in Fig. 7.

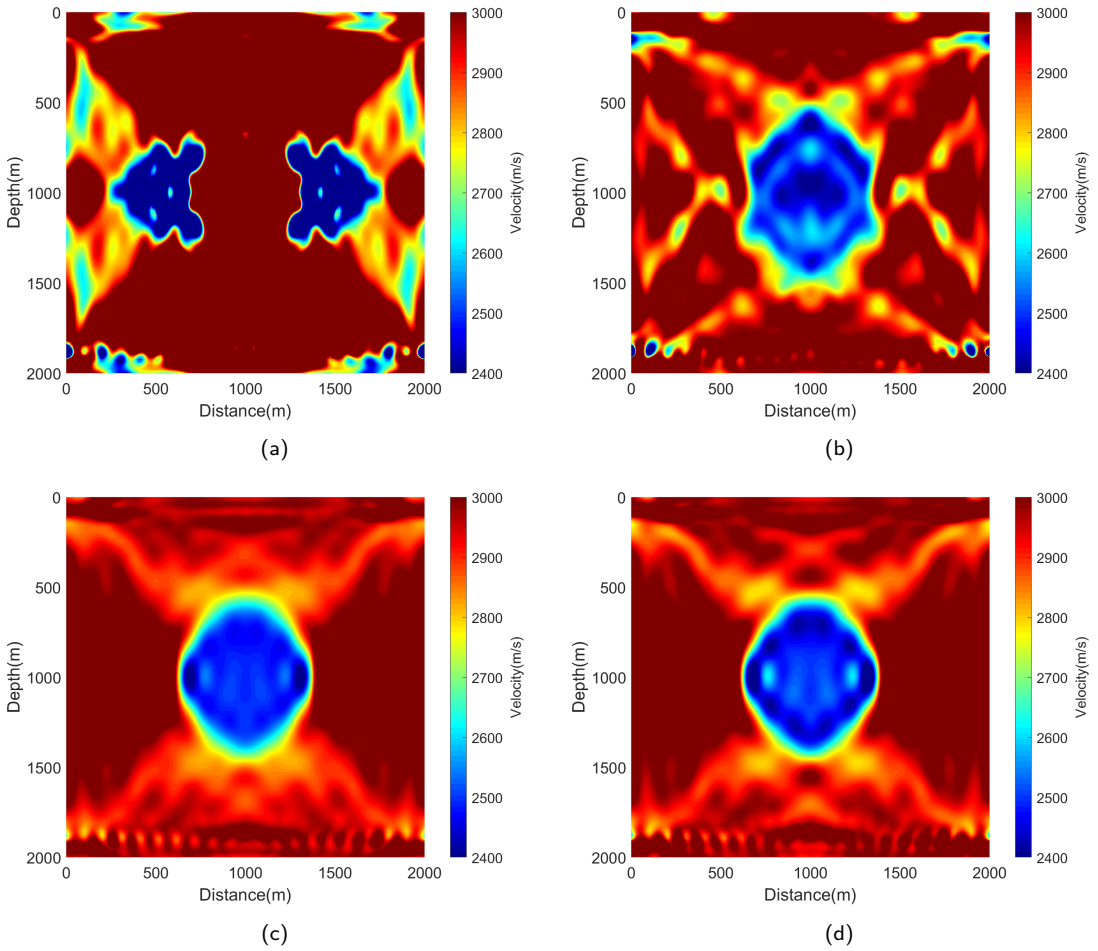


Figure 7: Inversion results of (a) L_2, (b) W_2^2, (c) $W_{\varepsilon,\lambda}$ and (d) $S_{\varepsilon,\lambda}$ for the Camembert model with a low-velocity anomaly.

Table 3: The total execution time and the model misfit after 50 iterations for the Camembert model with a low-velocity anomaly.

Method	L_2	OT	RUOT	USD
Time(min)	680	623	646	550
Model misfit	16.75	0.7812	0.2700	0.2737

Fig. 8 show the convergence curves. The result produced by the L_2 misfit is still inaccurate. However, the OT fails at this time, even though its data misfit decreases sufficiently. The total computational cost of the inversions and the model misfit after 50 iterations are listed in Table 3. The execution time for the USD inversion is faster than for the others.

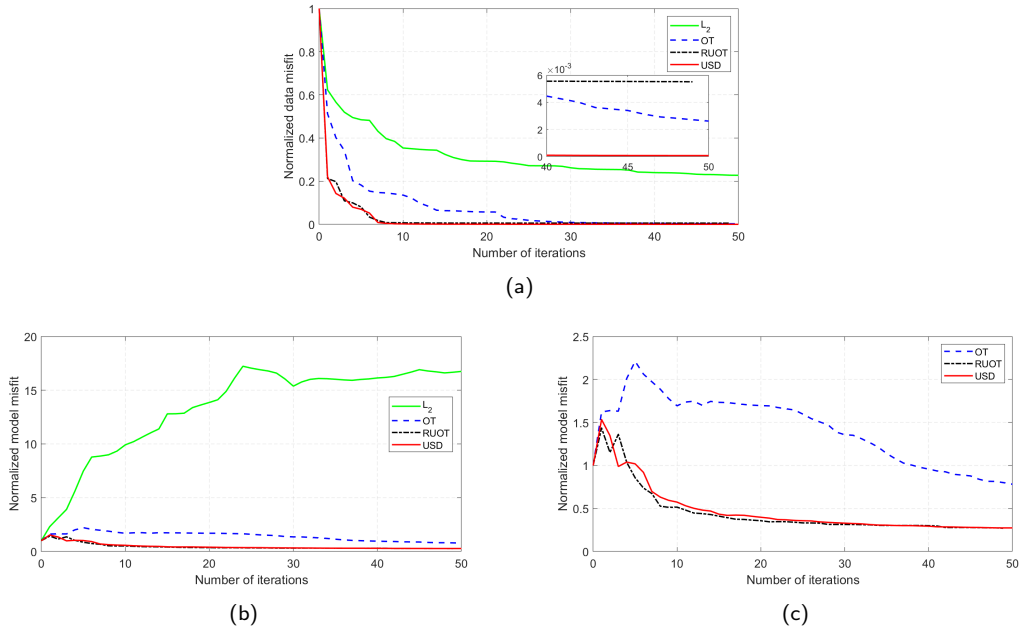


Figure 8: Convergence curves of the Camembert model with a low-velocity anomaly. (a) Data misfit, (b) model misfit and (c) model misfit without L_2.

The inversions based on the UOT-based distances display better performance than the inversions using the L_2 distance and the OT.

In these examples, the inversion mainly involves the direct wave, so the impact of phase information on inversion results is absent. Therefore, the inversions using the OT and the UOT-based metrics yield similar results when the anomal velocity is high. However, the mass conservation may result in the loss of the amplitude information of the wavefield, leading to the failure of the OT in the low-velocity anomaly model.

3.3 The crustal root model

Let us consider the crustal root model (Fig. 9a) [45] obtained by slightly modifying the S-wave velocity of the 1D IASP91 model [17]. A dipping and discontinuous Moho is constructed in this model. The computational domain is 100km$\times$75km. The model has three layers divided by the Conrad discontinuity at 20km depth and the Moho discontinuity whose location $(x,z(x))$ is a piecewise quadratic function:

$$z(x)=\begin{cases}35+\dfrac{1}{121}x^2\text{km}, & 0\text{km}\leqslant x\leqslant 55\text{km},\\ 35\text{km}, & 55\text{km}<x\leqslant 100\text{km}.\end{cases} \tag{3.1}$$

The velocities are 3.36km/s, 3.75km/s and 4.75km/s in the upper crust, lower crust, and mantle, respectively. The initial model shown in Fig. 9b is obtained by smoothing

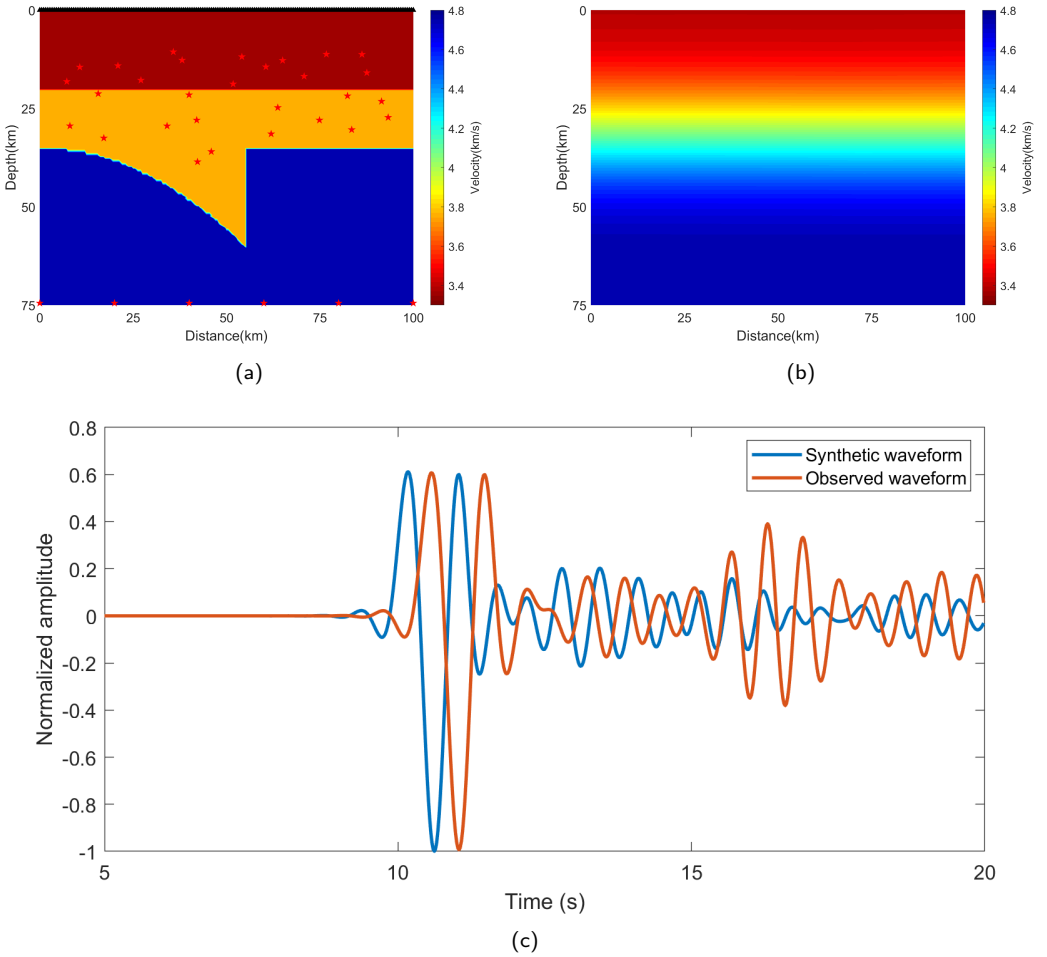


Figure 9: (a) True velocity and (b) initial velocity for the crustal root model. The triangles and stars indicate the receivers and sources, respectively. (c) The synthetic and observed waveforms.

the three-layer model without crustal root. In Fig. 9a, the triangles and stars indicate the receivers and sources, respectively. 201 equally spaced receivers are deployed on the surface. 29 sources are randomly distributed in the crust, and 6 sources are equally spaced at the bottom. The source time function of the sources is a Ricker wavelet with a peak frequency of 2 Hz. We set $\lambda = 20$ and $\varepsilon = 10^{-3}$ for the RUOT and the USD. The L-BFGS method is used as the optimization method for the inversion. We select a source and receiver pair located in the middle of the computational domain and plot a portion of the synthetic and observed waveforms, as shown in Fig. 9c. This example includes both direct waves and reflected waves, along with multiple phases. The location of the sources in our setup limits the proportion of direct waves in the inversion to a small percentage. Two discontinuities are mainly recovered by reflected waves.

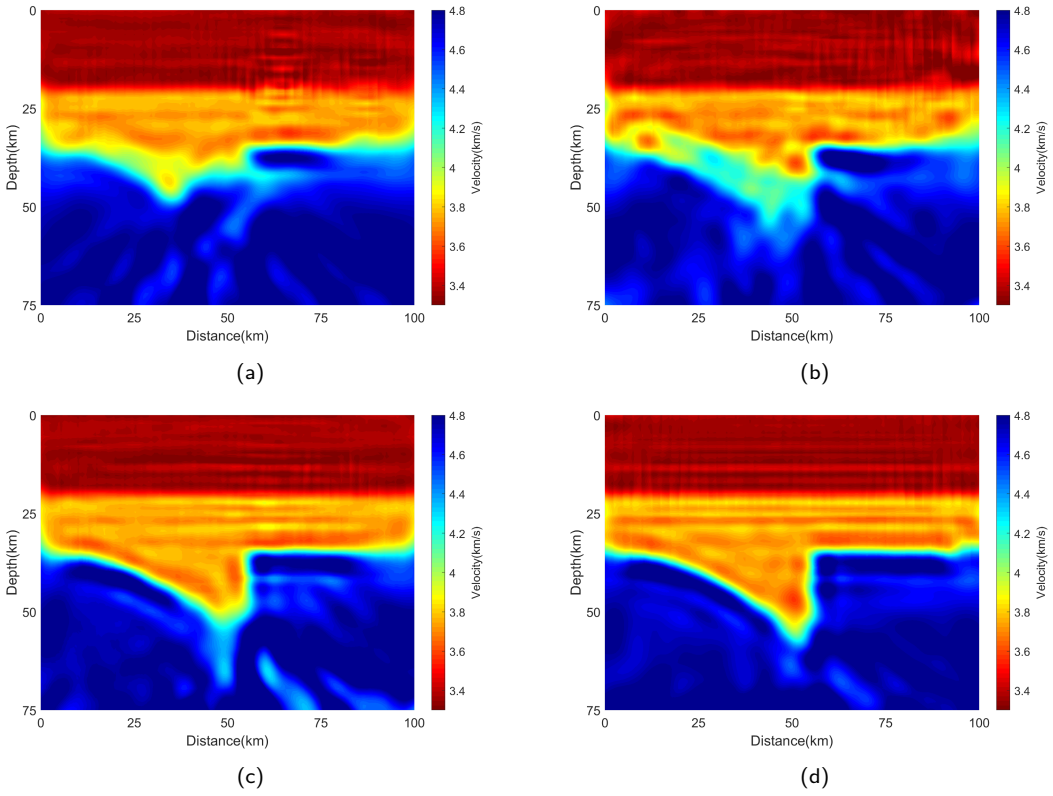


Figure 10: Inversion results of (a) L_2, (b) W_2^2, (c) $W_{\varepsilon,\lambda}$ and (d) $S_{\varepsilon,\lambda}$ for the crustal root model.

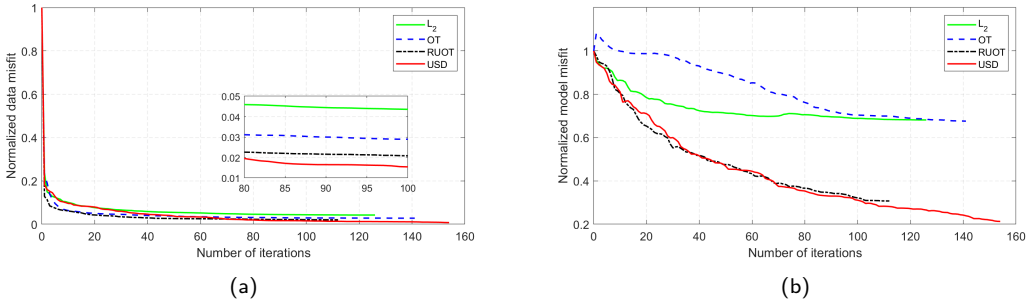


Figure 11: Convergence curves of the crustal root model. (a) Data misfit, (b) model misfit.

We present the inversion results obtained with L_2 distance, OT, RUOT, and USD in Fig. 10. Obviously, the inversions with L_2 distance and OT do not accomplish correct results, although their data misfits (Fig. 11a) reduce rapidly to a small value. Not only do they fail to capture the crustal root construct, but the Conrad interface is not clear either.

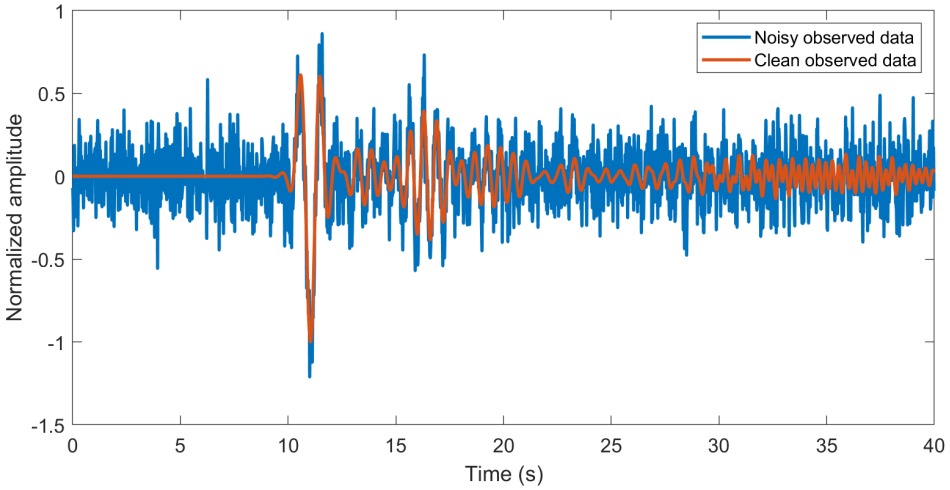


Figure 12: Noisy and clean observed data.

In contrast, the results based on the RUOT and the USD basically reconstruct the wave velocity. The USD inversion result exhibits a better resolution at the edge of the Moho surface than the RUOT inversion result, and its model misfit (Fig. 11b) thus decreases even further.

One of the ideal properties of the OT-based metric is its insensitivity to noise [12, 38]. We conduct the previous experiment again with a noisy reference by introducing normal random iid noise to the data from the true velocity (Fig. 12). The signal-to-noise ratio (SNR) is -3.29dB, indicating that the noise power significantly exceeds the signal power. Other settings remain consistent with the previous experiment. We perform the inversion without applying any filtering to the waveforms. After 65 iterations, the inversion results obtained with RUOT and USD are presented in Fig. 13. Despite the lower resolution compared to Fig. 10, these results still recover the main structure of the crustal root model, highlighting the insensitivity of UOT-based distances to noise.

4 Discussion

FWI with the L_2 distance is effective when the initial model is close to the true model. However, when the initial model has a large perturbation relative to the true model, the L_2 misfit may suffer from cycle-skipping. Unlike the point-to-point comparison of the traditional L_2-norm waveform difference, OT-based distances derive sensitivity kernels by comparing the mass of the waveforms, specifically their energy, which provides a more global perspective. Consequently, OT-based distances have greater potential to address cycle-skipping issues. The classical quadratic Wasserstein metric alleviates the cycle-skipping problem in some cases [12]. When the inversion involves a single seismic

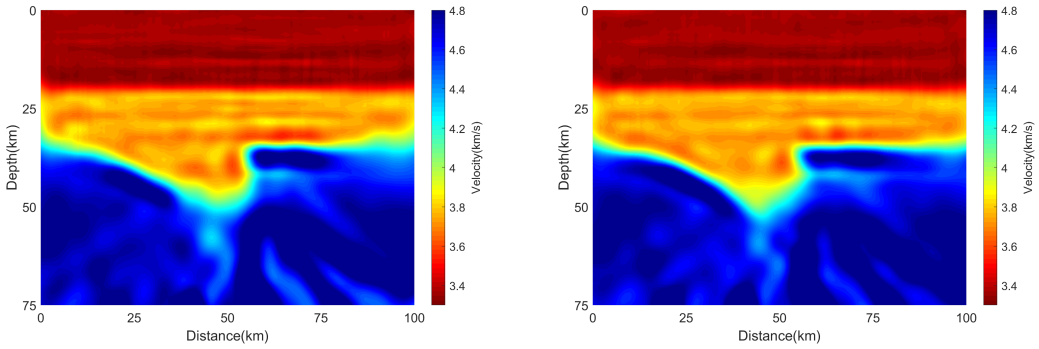
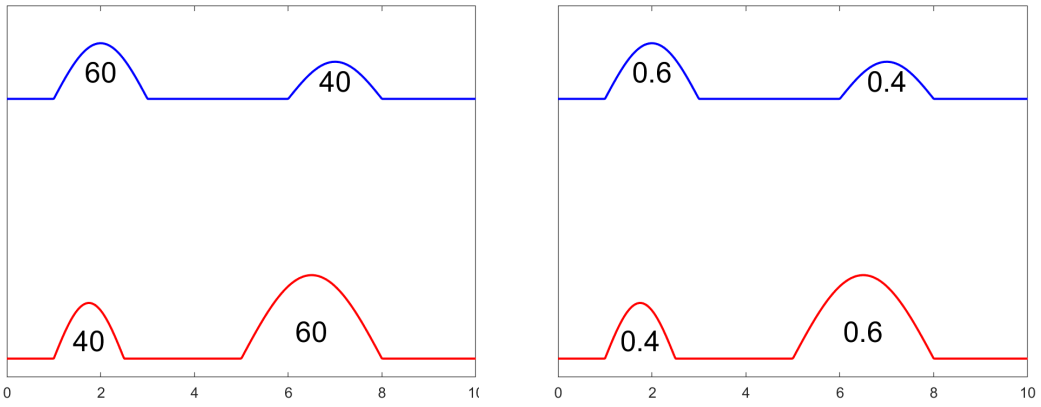


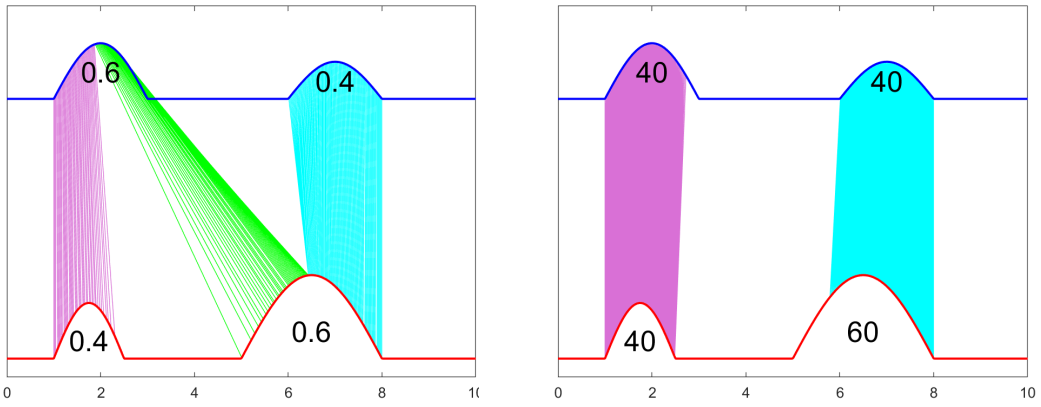
Figure 13: Inversion results of (a) $W_{\varepsilon,\lambda}$ and (b) $S_{\varepsilon,\lambda}$ for the crustal root model with the noisy data.

phase, the classical quadratic Wasserstein metric can overcome the defect of the L_2 distance and acquire reliable inversion results in most cases. When the seismic phases used in the inversion are complex, the classical quadratic Wasserstein metric may transport the mass between different seismic phases, leading to an unacceptable result as shown in Fig. 10b. Two UOT-based distances we introduced can overcome the flaws of the classical quadratic Wasserstein metric. Consider two mass distributions with a total mass of 100, as shown in Fig. 14a. The red line represents the original mass, and the blue line represents the target mass. Both mass distributions have two sections, which can be considered distinct phases. The mass of each part is shown in Fig. 14a. UOT can directly transport the mass between them, while OT should first rescale them to unit masses (Fig. 14b). Fig. 15 shows the two classes of mass transport. The mass transport of UOT exists only within the corresponding phases, whereas the illegal mass transport between different phases occurs in the mass transport of the balanced OT. In terms of UOT, the surplus mass will be retained if the mass exceeds the target mass, and the deficient mass will be disregarded if the mass is less than the target mass. Consequently, only 80% of the mass is transported in UOT. The additional cost resulting from the unequal mass is offset by the mass balancing terms. In contrast, the insufficient mass will be complemented by the transportation between different phases in OT. As a result of mass conservation, the balanced OT may lose the amplitude and phase information in FWI. In FWI, different seismic phases correspond to distinct waves or subsurface structures. The properties of the UOT facilitate comparison of corresponding seismic phases, enabling the accurate determination of the sensitivity kernels. In contrast, the balanced OT may permit illegal transport between different seismic phases, potentially resulting in incorrect sensitivity kernels. Therefore, in scenarios involving multiple seismic phases or significant attenuation, inversions utilizing the balanced OT are more likely to yield erroneous results. Selecting seismic phases carefully and applying the multiwindow method [9], which requires the operators to have sufficient experience, can avoid the aforementioned draw-



(a) (b)

Figure 14: (a) Mass distributions with a total mass of 100. (b) Normalized mass distributions. The red line is the origin mass and the blue line is the target mass. The number denotes the mass of each part.



(a) (b)

Figure 15: Illustration of the optimal transport maps between two mass distributions. (a) Balanced optimal transport. (b) Unbalanced optimal transport. The red line is the origin mass and the blue line is the target mass. The purple, green, and cyan lines represent the mass transport from the origin mass to the target mass. The number in the blue line denotes the mass being transported.

back of the balanced OT. Nevertheless, the UOT-based methods can achieve good results without tedious operations. Therefore, the UOT-based methods are more practical in FWI.

As previously discussed, the inversions employing UOT-based distances can provide more appropriate gradients, resulting in lower costs compared to the inversions using

the L_2 distance and the classical quadratic Wasserstein distance in line search for determining the step length. However, the calculation speed of the regularized UOT inversion will slow down because of the regularization when the data misfit is close to the minimum. Therefore, the FWI with the unbalanced Sinkhorn divergence demonstrates higher computational efficiency than the other three methods.

The UOT distances introduce two new parameters ε and λ, which have been empirically determined. We generally recommend that λ be set between 0.1 and 50. Small λ produces poor inversion results, while large λ leads to slow calculations. ε is always set to a small value to ensure the accuracy of the regularized UOT. On the other hand, if ε is too small, the computational burden would be intolerable. Pham et al. [33] mentioned the ideal value of ε to balance computational accuracy and efficiency, but it is not practical in our experiments. Li et al. [22] set ε as a small value multiplied by the maximal element of the cost matrix. As for the unbalanced Sinkhorn divergence, due to the aforementioned favorable properties, it offers a more flexible choice of ε.

5 Conclusion

In this study, we explore the FWI using two distinct UOT-based distances. We introduce a regularized UOT distance approach to address the limitations imposed by mass conservation. To compute the regularized UOT distance and its gradient, we employ an entropy regularization and scaling algorithm. Besides, we propose a truncation approximation method aiming at reducing computational costs of UOT-based distances. However, entropic penalization introduces a bias into the regularized UOT distance. To mitigate this bias, we first introduce the unbalanced Sinkhorn divergence. Numerical results demonstrate the superiority of UOT-based distances over both the traditional L_2 distance and the quadratic Wasserstein distance. Moreover, UOT-based distances circumvent the issue of cycle-skipping, even in scenarios where the classical quadratic Wasserstein distance proves ineffective. Therefore, the unbalanced optimal transport distances are more appropriate and dependable for the FWI problem. The unbalanced Sinkhorn divergence-based FWI is more accurate and cost-effective than the regularized unbalanced optimal transport-based FWI, while the regularized UOT distance is less expensive to calculate.

Notably, UOT-based distances can be easily extended to higher dimensions through a convolution approach [22], without significantly increasing computation burden. In this situation, the inter-receiver coherences are taken into account, and the misfit can be calculated source gather by source gather. Thus, theoretically, more accurate results are achievable. However, establishing the ground metric becomes challenging due to the simultaneous measurement of the cost in both time and space. One strategy involves normalizing the cost to achieve dimensionlessness [31]; however, this does not consistently generate favorable outcomes. This topic requires further investigation.

Acknowledgments

The authors appreciate the editor and anonymous reviewer for their constructive suggestions and insightful comments, which have significantly contributed to the improvement of the manuscript. This work is supported by the National Natural Science Foundation of China (Grant Nos. 42330801).

References

[1] K. Aki and W.H.K. Lee. Determination of three-dimensional velocity anomalies under a seismic array using first P arrival times from local earthquakes: 1. A homogeneous initial model. J. Geophys. Res., 81(23):4381-4399, 1976.
[2] J.-D. Benamou. Numerical resolution of an "unbalanced" mass transport problem. ESAIM: Math. Modell. Numer. Anal., 37(5):851-868, 2003.
[3] D. Borisov, R. Modrak, F. Gao, and J. Tromp. 3D elastic full-waveform inversion of surface waves in the presence of irregular topography using an envelope-based misfit function. Geophysics, 83(1):R1-R11, 2017.
[4] J. Chen, Y. Chen, H. Wu, and D. Yang. The quadratic Wasserstein metric for earthquake location. J. Comput. Phys., 373:188-209, 2018.
[5] M. Chen, F. Niu, Q. Liu, J. Tromp, and X. Zheng. Multiparameter adjoint tomography of the crust and upper mantle beneath East Asia: 1. Model construction and comparisons. J. Geophys. Res.: Solid Earth, 120(3):1762-1786, 2015.
[6] L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. Scaling algorithms for unbalanced optimal transport problems. Math. Comput., 87(314):2563-2609, 2018.
[7] L. Chizat, P. Roussillon, F. Léger, F.-X. Vialard, G. Peyré. Faster Wasserstein distance estimation with the Sinkhorn divergence. arXiv:2006.08172, 2020.
[8] X. Dong and D. Yang. Novel phase-sensitive full-waveform tomography for seismic imaging. Seismol. Res. Lett., 95(4):2328-2336, 2024.
[9] X. Dong, D. Yang, and F. Niu. Passive adjoint tomography of the crustal and upper mantle beneath eastern Tibet with a W2-norm misfit function. Geophys. Res. Lett., 46(22):12986-12995, 2019.
[10] T.A. El Moselhy and Y.M. Marzouk. Bayesian inference with optimal maps. J. Comput. Phys., 231(23):7815-7850, 2012.
[11] B. Engquist and B.D. Froese. Application of the Wasserstein metric to seismic signals. Commun. Math. Sci., 12(5):979-988, 2014.
[12] B. Engquist, B.D. Froese, and Y. Yang. Optimal transport for seismic full waveform inversion. Commun. Math. Sci., 14(8):2309-2330, 2016.
[13] B. Engquist and Y. Yang. Optimal transport based seismic inversion:beyond cycle skipping. Commun. Pure Appl. Math., 75(10):2201-2244, 2021.
[14] A. Fichtner. Full seismic waveform modelling and inversion. Springer-Verlag Berlin Heidelberg, 2011.
[15] O. Gauthier, J. Virieux, and A. Tarantola. Two-dimensional nonlinear inversion of seismic waveforms; numerical results. Geophysics, 51(7):1387-1403, 1986.
[16] L. Kantorovitch. On the translocation of masses. Manage. Sci., 5(1):1-4, 1958.
[17] B.L.N. Kennett and E.R. Engdahl. Traveltimes for global earthquake location and phase identification. Geophys. J. Int., 105(2):429-465, 1991.

[18] S. Kolouri, S. Park, M. Thorpe, D. Slepčev, and G.K. Rohde. Transport-based analysis, modeling, and learning from signal and data distributions. arXiv:1609.04767, 2016a.
[19] S. Kolouri, A.B. Tosun, J.A. Ozolek, and G.K. Rohde. A continuous linear optimal transport approach for pattern analysis in image datasets. Pattern Recognit, 51:453-462, 2016b.
[20] D. Komatitsch and J. Tromp. Introduction to the spectral element method for three-dimensional seismic wave propagation. Geophys. J. Int., 139(3):806-822, 1999.
[21] P. Lailly. The seismic inverse problem as a sequence of before stack migrations. Conference on Inverse Scattering, Theory and Application, pp.206-220, 1983.
[22] D. Li, M.P. Lamoureux, and W. Liao. Application of an unbalanced optimal transport distance and a mixed L1/Wasserstein distance to full waveform inversion. Geophys. J. Int., 230(2):1338-1357, 2022.
[23] P. Li, Q. Wang, and L. Zhang. A novel Earth mover's distance methodology for image matching with Gaussian mixture models. 2013 IEEE International Conference on Computer Vision, pp.1689-1696, 2013.
[24] Q. Liu and J. Tromp. Finite-frequency kernels based on adjoint methods. Bull. Seismol. Soc. Am., 96(6):2383-2397, 2006.
[25] J. Luo and R.S. Wu. Seismic envelope inversion: reduction of local minima and noise resistance. Geophys. Prospect., 63(3):597-614, 2015.
[26] S. Luo and P. Sava. A deconvolution-based objective function for wave-equation inversion. SEG Technical Program Expanded Abstracts, pp.2788-2792, 2011.
[27] Y. Luo and G.T. Schuster. Wave-equation traveltime inversion. Geophysics, 56(5):645-653, 1991.
[28] L. Métivier, A. Allain, R. Brossier, Q. Mérigot, E. Oudet, and J. Virieux. Optimal transport for mitigating cycle skipping in full-waveform inversion: A graph-space transform approach. Geophysics, 83(5):R515-R540, 2018.
[29] L. Métivier, R. Brossier, F. Kpadonou, J. Messud, and A. Pladys. A review of the use of optimal transport distances for high resolution seismic imaging based on the full waveform. arXiv:.08514, 2022.
[30] L. Métivier, R. Brossier, Q. Mérigot, and E. Oudet. A graph space optimal transport distance as a generalization of L^p distances: application to a seismic imaging inverse problem. Inverse Probl., 35(8), 2019.
[31] L. Métivier, R. Brossier, Q. Mérigot, E. Oudet, and J. Virieux. Measuring the misfit between seismograms using an optimal transport distance: application to full waveform inversion. Geophys. J. Int., 205(1):345-377, 2016.
[32] D. Peter, D. Komatitsch, Y. Luo, R. Martin, N. Le Goff, E. Casarotti, P. Le Loher, F. Magnoni, Q. Liu, C. Blitz, T. Nissen-Meyer, P. Basini, and J. Tromp. Forward and adjoint simulations of seismic wave propagation on fully unstructured hexahedral meshes. Geophys. J. Int., 186(2):721-739, 2011.
[33] K. Pham, K. Le, N. Ho, T. Pham, and H. Bui. On unbalanced optimal transport: An analysis of Sinkhorn algorithm. Proceedings of the 37th International Conference on Machine Learning. PMLR, pp.7673-7682, 2020.
[34] R.E. Plessix. A review of the adjoint-state method for computing the gradient of a functional with geophysical applications. Geophys. J. Int., 167(2):495-503, 2006.
[35] L. Qiu, J. Ramos-Martínez, A. Valenciano, Y. Yang, and B. Engquist. Full-waveform inversion with an exponentially encoded optimal-transport norm. SEG Technical Program Expanded Abstracts, pp.1286-1290, 2017.
[36] L. Qiu. Analysis of seismic inversion with optimal transportation and softplus encoding.

Inverse Probl., 37(9), 2021.
[37] J. Rabin and G. Peyré. Wasserstein regularization of imaging problem. 2011 18th IEEE International Conference on Image Processing, pp.1541-1544, 2011.
[38] T. Séjourné, J. Feydy, F.-X. Vialard, A. Trouvé, and G. Peyré. Sinkhorn divergences for unbalanced optimal transport. arXiv:1910.12958, 2019.
[39] J. Solomon, R. Rustamov, L. Guibas, and A. Butscher. Wasserstein propagation for semi-supervised learning. Proceedings of the 31st International Conference on Machine Learning. PMLR, pp.306-314, 2014.
[40] C. Tape, Q. Liu, A. Maggi, and J. Tromp. Adjoint tomography of the southern California Crust. Sci., 325(5943):988-992, 2009.
[41] A. Tarantola. Inversion of seismic-reflection data in the acoustic approximation. Geophysics, 49(8):1259-1266, 1984.
[42] J. Tromp, C. Tape, and Q. Liu. Seismic tomography, adjoint methods, time reversal and banana-doughnut kernels. Geophys. J. Int., 160(1):195-216, 2005.
[43] C. Villani. Topics in optimal transportation. Graduate Studies in Mathematics, AMS, 2003.
[44] J. Virieux and S. Operto. An overview of full-waveform inversion in exploration geophysics. Geophysics, 74(6):WCC1-WCC26, 2009.
[45] J. Wang, D. Yang, H. Jing, and H. Wu. Full waveform inversion based on the ensemble Kalman filter method using uniform sampling without replacement. Sci. Bull., 64(5):321-330, 2019.
[46] M. Warner and L. Guasch. Adaptive waveform inversion: Theory. Geophysics, 81(6):R429-R445, 2016.
[47] D. Yang, X. Dong, J. Huang, Z. Fang, X. Huang, S. Liu, M. Liu, and W. Meng. High-resolution full waveform seismic imaging: Progresses, challenges, and prospects. Sci. China Earth Sci., 68(2):315-342, 2025.
[48] Y.N. Yang, B. Engquist, J.Z. Sun, and B.F. Hamfeldt. Application of optimal transport and the quadratic Wasserstein metric to full-waveform inversion. Geophysics, 83(1):R43-R62, 2018.
[49] P. Yong, W. Liao, J. Huang, Z. Li, and Y. Lin. Misfit function for full waveform inversion based on the Wasserstein metric with dynamic formulation. J. Comput. Phys., 399, 2019.
[50] Y.O. Yuan, F.J. Simons, and E. Bozdağ. Multiscale adjoint waveform tomography for surface and body waves. Geophysics, 80(5):R281-R302, 2015.
[51] H. Zhu, E. Bozdağ, and J. Tromp. Seismic structure of the European upper mantle based on adjoint tomography. Geophys. J. Int., 201(1):18-52, 2015.